BIOTECH - GO

Blended learning in Bioinformatics - the SMEs instrument for Biotech innovations

"Information is Knowledge and Today's Economy is Knowledge economy"

The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Funded by the Erasmus+ Programme of the European Union

A few Notes

Circle Economy advocates a new economic approach. It requires building of programs and tools to help accelerate the scalable adoption of the circular economy across businesses, governments and communities. This has imposed development and implementation of relevant European policies and tools to answer this challenge. Thus, "European Area of Skills and Qualifications" intends to further strengthen the links between business, education/training, mobility and the labor market. In this respect Europe's economic development is becoming increasingly dependent on SMEs. To answer the needs related to the transparency and recognition of skills and qualifications of SMEs personnel became a crucial importance. Furthermore, these companies (SMEs) lack many of the support networks that are taken for granted by larger companies. For example, each small Biotech company relies on Bioinformatics for its research, and effective bioinformatics tools are often key part of business strategy. Yet many SMEs have only a single member of staff responsible for this important aspect of their business. On this basis the engagement of staff in education and training in order to update and upgrade their skills within the continuous or life-long learning approach is a key issue. In order to achieve this, the small businesses need to engage relevant training providers or VET professionals.

Taking into account all above the main goal of BIOTECH-GO project is focused on the provision of innovation in skills improvement for VET professionals in the fields of Bioinformatics, thus assuring new ways of talent development for small and medium-sized enterprises (SMEs) employees. Project contributes to the advance of a European Area of Skills and Qualifications through creating specific VET tools in the subject area (EQF/NQF, ECVET). Knowledge, skills, responsibility & autonomy update of VET specialists working in the project subject area will further promote excellence, and will raise awareness of the fundamental concepts underlying bioinformatics in different biotech companies, such as:

- **contribution to the advancement of biology research in Biotech SMEs through bioinformatics tools application;**
- **provision of advanced bioinformatics training to SMEs personnel at all levels, from technicians to independent investigators;**
- **helping for dissemination of cutting-edge technologies to industry;**
- **coordination of biological data provision across Europe.**

Biology, biological databases, and high-throughput data sources

Basic level

Ventsislava Petrova
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Kliment Petrov
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Contents

Biology in the Computer Age

Bioinformatics is the science combining utilization of computer and biological data. It's the instrument we can use to understand biological processes and to answer of numerous others questions. Entirely, bioinformatics is a subset of the bigger field of computational science, the use of quantitative scientific strategies in modelling biological systems. The field of bioinformatics depends vigorously on work by specialists with statistical methods and pattern recognition. Scientists come to bioinformatics from many fields, including arithmetic, software engineering, and semantics. Unfortunately, biology is a study of the particular and in addition the general. Bioinformatics is full of pitfalls for the individuals who search for examples and make expectations without an entire comprehension of where biological data originates from and what it implies. By giving calculations, databases, UIs, and measurable devices, bioinformatics makes it conceivable to do things like compare DNA sequences and generate results that are potentially significant. Possibly critical" is maybe the most essential expression. “These new approaches additionally give the chance to overinterpret information and assign meaning where none truly exists”. We can't exaggerate the significance of understanding the restrictions of these tools. In any case, once you gain that understanding and turn into smart user of bioinformatics strategies, the speed at which your research advances can be genuinely astonishing.

Bioinformatics deals with any type of data that is of interest to biologists

- DNA and protein sequences
- Gene expression (microarray)
- Articles from the literature and databases of citations
- Images
- Raw data collected from any type of field or laboratory experiment
- Software

How Informatics Change Biology?

Biological genetic and functional data are stored as DNA, RNA, and proteins, which are all linear chains composed of smaller molecules. These macromolecules are composed from a defined alphabet of well-studied chemicals: DNA is comprised of four deoxyribonucleotides (adenine, thymine, cytosine, and guanine), RNA is made up from the four ribonucleotides (adenine, uracil, cytosine, and guanine), and proteins are built using the 20 amino acids. Since these macromolecules are straight chains of characterized parts, they can be represented as sequences of symbols. These sequences can then be compared to find similarities that suggest the molecules are related by form or function. Sequences examination is conceivably the most valuable computational tool to emerge for molecular biologists. The World Wide Web has made it possible for a single public database of genome sequence data to give benefits through a uniform interface to an overall group of users. With an ordinarily utilized PC program called fsBLAST, molecular biologists can compare an uncharacterized DNA with the all openly available DNA sequence collections.

Bioinformatics and Databases Building

A lot of what we currently consider as a major aspect of bioinformatics— sequence comparison, sequence database searching, sequence analysis —is more complicated than simply outlining and setting public databases. Bioinformaticians (or computational scientists) go beyond simply downloading, managing, and introducing information, drawing motivation from a wide variety of quantitative fields, including statistics, physics, material science, software engineering. Figure 1 indicates how quantitative science intersects with biology at each level, from investigation of sequence information and macromolecules structure, to metabolic modelling, to quantitative study of populations and ecology.

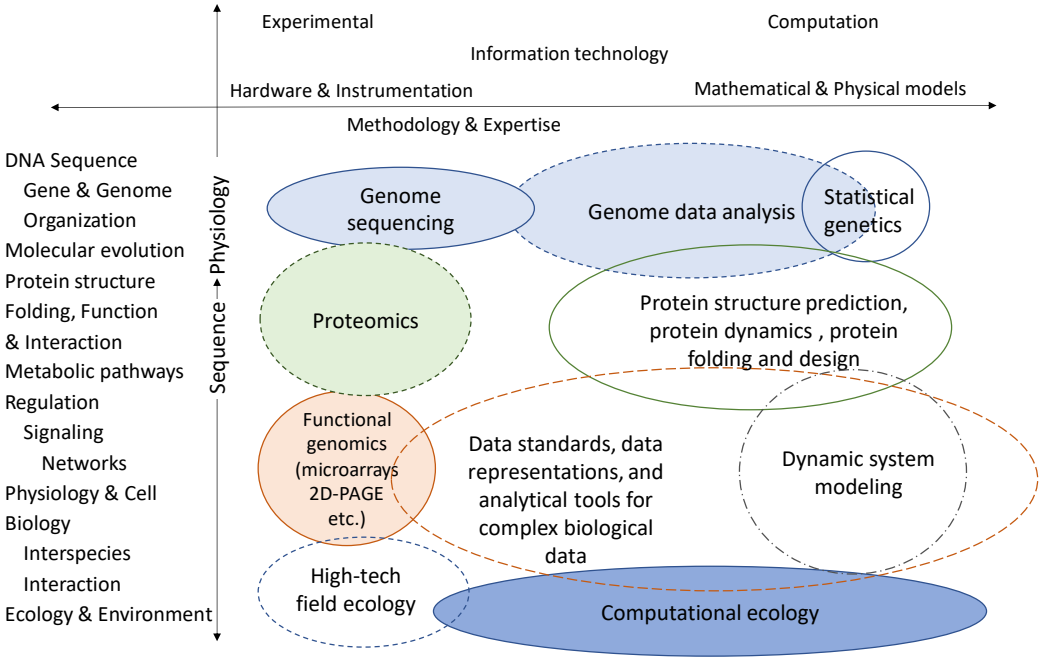


Figure 1. How technology intersects with biology

Bioinformatics is above all else a part of the biological sciences. The principle objective of bioinformatics isn't building up the most sophisticated algorithms or the most hidden analysis; the

objective is discovering how living organism function. Like the molecular biology science strategies that extraordinarily extended what researcher were fit for examining, bioinformatics is an approach and not an end in itself. Bioinformaticians are the tool- developers, and it's important that they comprehend natural issues and computational arrangements so as to create valuable instruments. Research in bioinformatics and computational science can incorporate abstraction of the properties of a biological system into a mathematical or physical model, to execution of new calculations for information investigation, to the improvement of databases and web tools to assess them.

Informatics and Biologists

The science of informatics is focused on the representation, organization, manipulation, distribution, maintenance, and use of data, especially in computerized frame. The functional part of bioinformatics is the representation, storage, and distribution of data. Smart outline of information configurations and databases, formation of instruments to search in those databases, and advancement of UIs that unite diverse apparatuses to enable the user to make complex inquiries about the information are generally parts of the improvement of bioinformatics foundation.

Creating analytical tools to find information in information is the second, and more logical, part of bioinformatics. There are many levels at which we utilize biological data, regardless of whether we are comparing sequences to build up a theory about the function of a newfound gene, examining known 3D protein structures to discover patterns that can help foresee how the protein folds, or displaying how proteins and metabolites in a cell cooperate to make the cell function. A ultimate objective of analytical bioinformaticians is to create prescient techniques that enable researchers to display the function and phenotype of a living organisms based only on its genome sequence.

Bioinformatician Skills?

There's an extensive variety of points that are helpful in case you're interested in bioinformatics, and it's not possible to learn them all. However, the following "core requirements" for bioinformaticians could be underlined:

- ✓ Have a genuinely profound background in some part of molecular biology, like: biochemistry, molecular biology, molecular biophysics, or even molecular modelling.
- ✓ Completely comprehend the “central dogma” of molecular biology. Understanding how and why DNA sequence is transcribed into RNA and then translated into protein.
- ✓ Have significant experience with at least one or two major molecular biology software packages, either for sequence analysis or molecular modelling. The experience of learning one of these softwares makes it substantially much esier to figure out how to utilize other available programmes.
- ✓ Be open to work in command-line computing environment.
- ✓ Have experience with programming in a computer language, for example, C/C++, as well as in a scripting language, for example, Perl or Python.

Biologists and Computers

Computers are powerful devices for study any system that can be described in a mathematical way. As our comprehension of biological processes has developed and extended, it isn't amazing, at that point, that computational biology and bioinformatics, have advanced from the intersection of traditional biology, mathematics, and computer science.

The expanding automation of experimental molecular biology and the use of increasing data in the biological sciences have prompt a major change in the way biological research is performed.

Notwithstanding narrative research — finding and studying in detail a single gene at a time — we are presently classifying all the information that is accessible, making complete maps to which we to can later return and mark the points of interest. This is occurring in the domains of sequence and structure, and has started to be the way to deal with different sorts of information also. The trend is toward storage of row biological information in numerous public databases with open access. Rather than doing preparatory research in the lab, investigators are going to the databases initially to save time and assets.

Web Information Use

While you can rapidly locate a single protein structure file or DNA sequence file by filling in a web form and looking through a public database, it's reasonable that in the end you will want to work with more than one bit of information. You may gathering and archiving your own particular information; as well as you might need to make newly discovered information accessible to a broader research community. To do these things effectively, you have to store information on your own PC. In the event that you need to process your data utilizing a computer program, you have to structure your information. Understanding the contrast amongst organized and unstructured information and outlining an information arrange that suits your data storage and access needs is the way to making your information valuable and accessible.

There are numerous approaches to sort out information. While most biological data is stored in flat file databases, this sort of database becomes inefficient when the quantity of data being stored becomes extremely large. More information regarding differences between flat file and relational databases, introduce the best public -domain tools for managing databases, and show you how to use them to store and access your data you could find in GM2 (Advance level).

Understanding Sequence Alignment Data

It's difficult to comprehend your data, or make a point, without visualization tools. The extraction of cross sections or subsets of complex multivariate data is regularly required to understand biological information. Once you've stored data in an open, flexible format, the next stage is to extract what is essential to you and visualize it. You have to make a histogram of your information or show a molecular structure in three dimensions and watch it move in real time using a specific visualization instruments.

Predicting Protein Structure from Sequence

There are a few questions that Bioinformatics can't answer, and this is one of them. Indeed, it's one of the greatest open research inquiries in computational science. What is conceivable is to give the instruments to discover data about such issues and different authors who are working on them. Bioinformatics, similar to some other science, doesn't generally give fast and simple responses to all issues.

Questions That Bioinformatics Can Answer

The questions that drive bioinformatics development are similar that people have at in applied biology for the last couple of hundred years. How might we cure disease? How might we prevent infection? How might we produce enough food to sustain all of mankind? Organizations working in the field of drugs development, agricultural chemicals, hybrid plants, plastics and other petroleum derivatives, and biological approaches to environmental remediation, among others, are creating bioinformatics divisions and looking to bioinformatics to give new targets and to help replace scarce natural resources.

The presence of genome projects infers our goal to utilize the information they create. The important objectives of modern molecular biology are to read the entire genomes of living organisms, to identify each gene, to match every gene with the protein it encodes, and to determine the structure and function of each protein. Detailed knowledge of gene sequence, protein structure and function, and gene expression patterns is expected to enable us to see how life functions at the most noteworthy conceivable resolution. In this way the ability to manipulate living organisms will be performed with exactness and precision.

Computational Approaches to Biological Questions

There is a standard range of approaches that are applied in bioinformatics. Currently, the greater part of the important methods depends on one key principle: that sequence and structural *homology* (or similarity) between molecules can be utilized to define basic and functional similarity. Here, an outline for the standard computer tools accessible to researcher is given; in GM2 how specific software packages implement these strategies is examined and how a researcher should utilize them.

Molecular Biology's Central Dogma

The central dogma of molecular biology states that:

- ✓ DNA is a template to replicate itself,
- ✓ DNA is transcribed into RNA, and
- ✓ RNA is translated into protein.

In brief, genomic DNA contains all the necessary information about functioning of a define living organism. Without DNA, organisms wouldn't be able to replicate themselves. The raw "one-dimensional" sequence of DNA, however, doesn't actually do anything biochemically; it's only store information, a blueprint that is read by the cell's protein synthesizing machinery. DNA sequences are the punch cards; cells are the computers.

Replication of DNA

The specific structure of DNA molecules assures its special properties. These properties allow the information stored in DNA to be preserved and transferred from one cell to another, and thus from parents to their offspring.

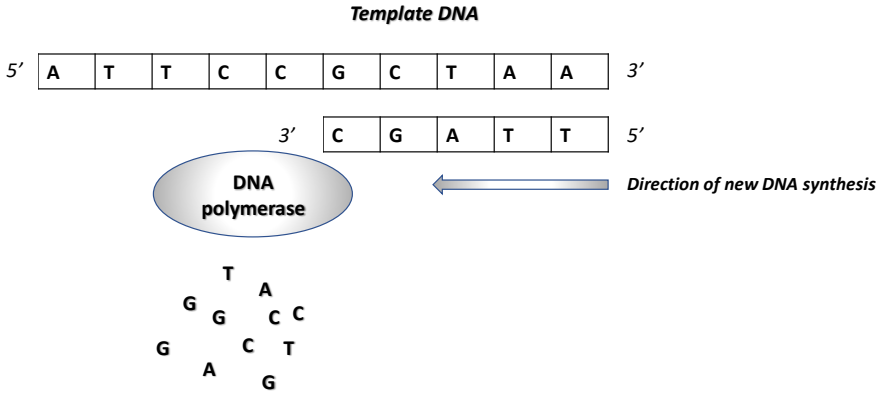


Figure 2. Schematic replication of DNA helix

Genomes and Genes

The genome comprises individual genes. There are three classes of genes: *protein-coding* genes, *RNA-specifying* genes are untranscribed *genes*.

Transcription of DNA

DNA act as a blueprint for a synthesis of ribonucleic acid (RNA).

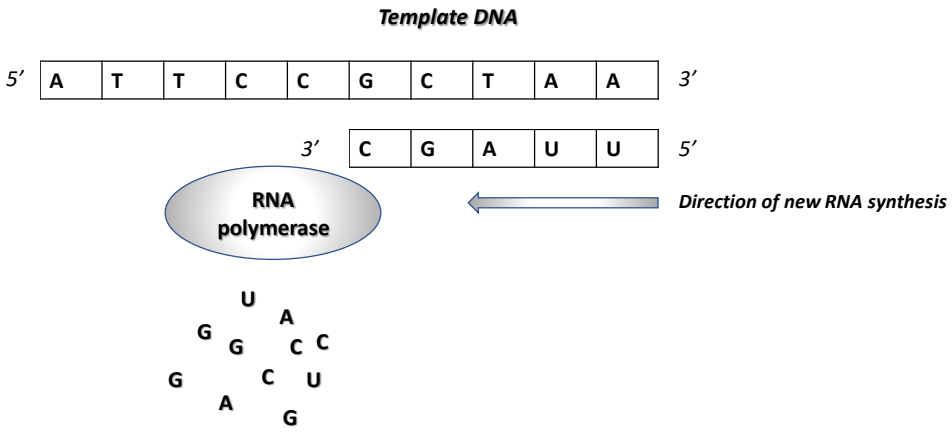
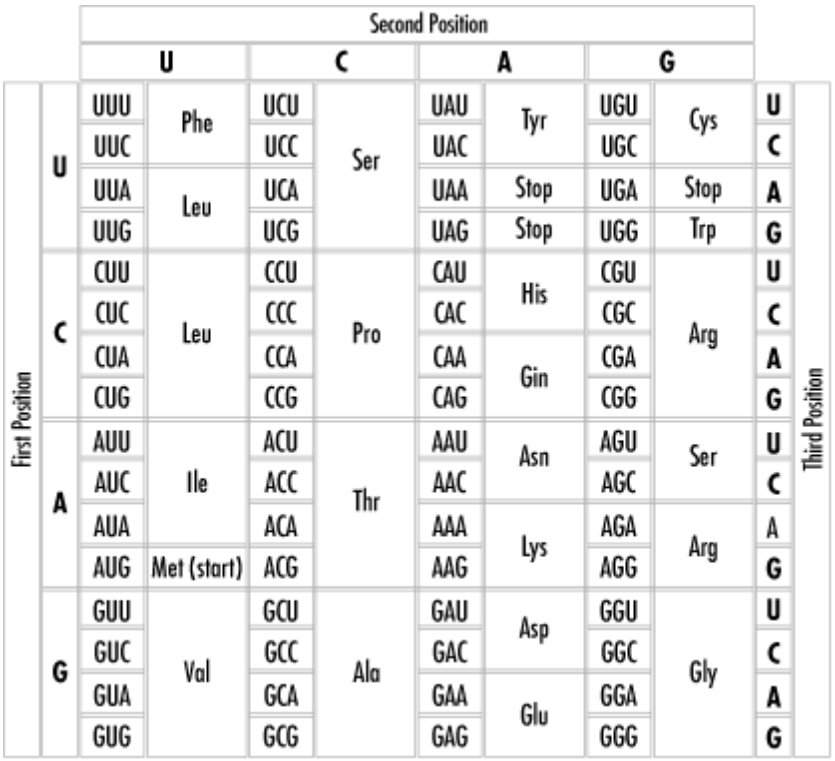


Figure 3. Schematic transcription of DNA into RNA

Translation of mRNA

Translation of mRNA into protein is the final key step in putting the information in the genome to work in the cell.



First Position	Second Position: U		Second Position: C		Second Position: A		Second Position: G		Third Position
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gin	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met (start)	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 4. The genetic code

Molecular Evolution

Errors in replication and transcription of DNA are relatively common. If these errors occur in in dividing cells, they can be passed to its offspring. Modifications in the DNA sequence can have harmful effect, they can also have beneficial, or they can be neutral. If a mutation doesn't kill the organism before it reproduces, the mutation can become fixed in the population over many generations. The slow accumulation of such mutations is the background of the *evolution*. Thus, knowing the DNA sequences provide us with more precise understanding of evolution. Knowing the molecular mechanism of evolution as a gradual process of accumulating DNA sequence mutations is the reason for creating theories based on DNA and protein sequence comparison.

Biological Models

One of the most important exercises in biology and bioinformatics is modeling. A *model* is an abstract way of describing a complicated system. Turning something as complex (and confusing) as a chromosome, or the cycle of cell division, into a simplified representation that captures all the features you are trying to study can be extremely difficult. A model helps us see the larger picture. One feature of a good model is that it makes systems that are otherwise difficult to study easier to analyze using quantitative approaches. Bioinformatics tools rely on our ability to extract relevant parameters from a biological system (be it a single molecule or something as complicated as a cell), describe them quantitatively, and then develop computational methods that use those parameters to compute the properties of a system or predict its behavior.

Accessing 3D Molecules Through a 1D Representation

In reality, DNA and proteins are complicated 3D molecules, composed of thousands or even millions of atoms bonded together. However, DNA and proteins are both *polymers*, chains of repeating *monomers*. Not too long after the chemical natures of DNA and proteins were understood, researchers recognized that it was convenient to represent them by strings of single letters. Instead of representing each nucleic acid in a DNA sequence as a detailed chemical entity, they could be represented simply as A, T, C, and G. Thus, a short piece of DNA that contains thousands of individual atoms can be represented by a sequence of few hundred letters.

Not only does this abstraction save storage space and provide a convenient form for sharing sequence information, it represents the nature of a molecule uniquely and correctly and ignores levels of detail (such as atomic structure of DNA and many proteins) that are experimentally inaccessible. Many computational biology methods exploit this 1D abstraction of 3D biological macromolecules.

The abstraction of nucleic acid and protein sequences into 1D strings has been one of the most fruitful modeling strategies in computational molecular biology, and analysis of character strings is a longstanding area of research in computer science. One of the elementary questions you can ask about strings is, "Do they match?" There are well-established algorithms in computer science for finding exact and inexact matches in pairs of strings. These algorithms are applied to find pairwise matches between biological sequences and to search sequence databases using a sequence query.

In addition to matching individual sequences, string-based methods from computer science have been successfully applied to a number of other problems in molecular biology. For example, algorithms for reconstructing a string from a set of shorter substrings can assemble DNA sequences from overlapping sequence fragments. Techniques for recognizing repeated patterns in single sequences or conserved patterns across multiple sequences allow researchers to identify signatures associated with biological structures or functions. Finally, multiple sequence-alignment techniques allow the simultaneous comparison of several molecules that can infer evolutionary relationships between sequences.

This simplifying abstraction of DNA and protein sequence seems to ignore a lot of biology. The cellular context in which biomolecules exist is completely ignored, as are their interactions with other molecules and their molecular structure. And yet it has been shown over and over that matches between biological sequences can be biologically meaningful.

Abstractions for Modeling Protein Structure

There is more to biology than sequences. Proteins and nucleic acids also have complex 3D structures that provide clues to their functions in the living organism. Structure analysis can be performed on static structures, or movements and interactions in the molecules can be studied with molecular simulation methods.

Standard molecular simulation approaches model proteins as a collection of point masses (atoms) connected by bonds. The bond between two atoms has a standard length, derived from experimental chemistry, and an associated applied force that constrains the bond at that length. The angle between three adjacent atoms has a standard value and an applied force that constrains the bond angle around that value. The same is true of the dihedral angle described by four adjacent atoms. In a molecular dynamics simulation, energy is added to the molecular system by simulated "heating." Following standard Newtonian laws, the atoms in the molecule move. The energy added to the system provides an opposing force that moves atoms in the molecule out of their standard conformations. The actions and reactions of hundreds of atoms in a molecular system can be simulated using this abstraction.

In any case, the computational requests for molecular simulations are huge, and there is some weakness both in the force field - the accumulation of standard forces that model the molecule — and

in the displaying of nonbonded interactions - interactions between nonadjacent atoms. In this way, it has not demonstrated conceivable to anticipate protein structure utilizing the all-atom modeling approach.

A few researchers have recently moderate success in predicting protein topology for small proteins utilizing a moderate level of abstraction — more than linear sequence, but less than an all atom model. For this situation, the protein is dealt with as a progression of globules (speaking to the individual amino acids) on a string (speaking to the backbone). Globules may have distinctive characters to represent the distinctions in the amino acids sidechains. They might be positively or negatively charged, polar or nonpolar, small or large. There are rules overseeing which globules will attract each other. Polar groups cluster with other polar groups, and nonpolar with nonpolar. There are also rules concerning the the string; essentially that it can't go through itself throughout the course of simulation. Modeling the protein folding itself is directed through sequential or simultaneous perturbations of the position of each globule.

Mathematical Modeling of Biochemical Systems

Using theoretical models in biology goes far beyond the single molecule level. For years, ecologists have been using mathematical models to help them understand the dynamics of changes in interdependent populations. What effect does a decrease in the population of a predator species have on the population of its prey? What effect do changes in the environment have on population? The answers to those questions are theoretically predictable, given an appropriate mathematical model and a knowledge of the sizes of populations and their standard rates of change due to various factors.

In molecular biology, a similar approach, called *metabolic control analysis*, is applied to biochemical reactions that involve many molecules and chemical species. While cells contain hundreds or thousands of interacting proteins, small molecules, and ions, it's possible to create a model that describes and predicts a small corner of that complicated metabolism. For instance, if you are interested in the biological processes that maintain different concentrations of hydrogen ions on either side of the mitochondrial inner membrane in eukaryotic cells, it's probably not necessary for your model to include the distant group of metabolic pathways that are closely involved in biosynthesis of the heme structure.

Metabolic models depict a biochemical process in respect to the concentrations of chemical substances engaged with a pathway, and the reactions and fluxes that influence those concentrations. Reactions and fluxes can be identified by differential equations; they are basically rates of change in concentration.

What makes metabolic modeling intriguing is the possibility of displaying many reactions at the same time to perceive what impact they have on the concentration of specific chemical compound. Utilizing a properly built metabolic model, you can test diverse presumptions about cell conditions and fine-tune the model to simulate experimental trials. That, in turn, can propose testable speculations to drive further research.

Bioinformatics Approaches

Molecular biology research is a fast-growing area. The amount and type of data that can be gathered is exploding, and the trend of storing this data in public databases is spilling over from genome sequence to all sorts of other biological datatypes. The information landscape for biologists is changing so rapidly that often more of the provided information is somewhat behind the times.

Yet, since the inception of the Human Genome Project, a core set of computational approaches has emerged for dealing with the types of data that are currently shared in public databases—DNA, protein sequence, and protein structure. Although databases containing results from new high-throughput molecular biology methods have not yet grown to the extent the sequence databases have, standard methods for analyzing these data have begun to emerge.

The following list gives an overview of the key computational methods:

Using public databases and data formats

The first key skill for biologists is to learn to use online search tools to find information. Literature searching is no longer a matter of looking up references in a printed index. You can find links to most of the scientific publications you need online. There are central databases that collect reference information, so you can search dozens of journals at once. You can even set up "agents" that notify you when new articles are published in an area of interest. Searching the public molecular-biology databases requires the same skills as searching for literature references: you need to know how to construct a query statement that will pluck the particular needle you're looking for out of the database haystack.

Sequence alignment and sequence searching

Having the capacity to analyze pairs of DNA or protein sequences and extract partial matches has made it conceivable to utilize a biological sequence as a database query. Sequence-based searching is another key expertise for biologists; a little investigation of the biological databases toward the start of a scientific project often saves a lot of valuable time in the lab. Recognizing homologous sequences gives a basis to phylogenetic examination and sequence pattern recognition. Sequence-based searching should be possible online through web platforms, so it requires no extraordinary computer skills, yet to judge the quality of your search results or you have to understand how the sequence-alignment method functions and how to go beyond different kinds of further investigations.

Gene prediction

Gene prediction is just one of a bunch of techniques for recognition of meaningful signals in uncharacterized DNA sequences. Up to this point, most sequences deposit in GenBank were already characterized at the time of deposition. That is, somebody had officially gone in and, utilizing molecular biology, genetic, or biochemical approaches, made sense of what the gene did. Nonetheless, now that the genome projects are going all out, a lot of DNA sequence out there that isn't characterized.

Programming for forecast of open reading frames, genes, exon splice sites, promoter binding sites, repeat sequences, and tRNA genes enables researchers to make sense out of this unmapped DNA.

Multiple sequence alignment

Multiple sequence-alignment techniques assemble pairwise sequence alignment for some related sequences into a image of sequence homology among all individuals from a gene family. Multiple sequence alignments help in visual distinguishing of sites in a DNA or protein sequence that might be functionally important. Such sites are normally conserved; the same amino acid is present at that site in each one of a group of related sequences. Multiple sequence alignments can also be quantitatively examined to obtain data about certain gene family. This technique is a basic advance in phylogenetic investigation of a group of related sequences, and they additionally provide the basis for identifying sequence patterns that describe specific protein families.

Phylogenetic analysis

Phylogenetic analysis endeavors to depict the evolutionary relatedness of a group of sequences. A traditional phylogenetic tree or cladogram groups species into a diagram presenting their relative evolutionary similarity / divergence. Branching of the tree that occur uttermost from the root isolate

individual species; branching that that occur close to the root assembly species into kingdoms, phyla, classes, families, genera, et cetera.

The information in a molecular sequence alignment can be used to compute a phylogenetic tree for a particular family of gene sequences. The branching in phylogenetic trees represent evolutionary distance based on sequence similarity scores or on information-theoretic modeling of the number of mutational steps required to change one sequence into the other. Phylogenetic analyses of protein sequence families talk not about the evolution of the entire organism but about evolutionary change in specific coding regions, although our ability to create broader evolutionary models based on molecular information will expand as the genome projects provide more data to work with.

Extraction of patterns and profiles from sequence data

A *motif* is a sequence of amino acids that defines a substructure in a protein that can be connected to function or to structural stability. In a group of evolutionarily related gene sequences, motifs appear as conserved sites. Sites in a gene sequence tend to be *conserved*—to remain the same in all or most representatives of a sequence family—when there is selection pressure against copies of the gene that have mutations at that site. Nonessential parts of the gene sequence will diverge from each other in the course of evolution, so the conserved motif regions show up as a signal in a sea of mutational noise. Sequence profiles are statistical descriptions of these motif signals; profiles can help identify distantly related proteins by picking out a motif signal even in a sequence that has diverged radically from other members of the same family.

Protein sequence analysis

The amino-acid content of a protein sequence can be used as the basis for many analyses, from computing the isoelectric point and molecular weight of the protein and the characteristic peptide mass fingerprints that will form when it's digested with a particular protease, to predicting secondary structure features and post-translational modification sites.

Protein structure prediction

It's a lot harder to determine the structure of a protein experimentally than it is to obtain DNA sequence data. One very active area of bioinformatics and computational biology research is the development of methods for predicting protein structure from protein sequence. Methods such as secondary structure prediction and threading can help determine how a protein might fold, classifying it with other proteins that have similar topology, but they don't provide a detailed structural model. The most effective and practical method for protein structure prediction is *homology modeling*—using a known structure as a template to model a structure with a similar sequence. In the absence of homology, there is no way to predict a complete 3D structure for a protein.

Protein structure property analysis

Protein structures have numerous quantifiable properties that are important to crystallographers and structural biologists. Protein structure validation devices are utilized by crystallographers to measure how well a structure model fits in with auxiliary standards extricated from existing structures or chemical model compounds. These instruments may also examine the "fitness" of each amino acid in a structure model for its environment, hailing such peculiarities as hidden charges with no countercharge or large patches of hydrophobic amino acids found on a protein surface. These tools are valuable for assessing both experimental and hypothetical structure models.

Another class of methods can figure inner geometry and physicochemical properties of proteins. These instruments generally are used to create models of the protein's catalytic mechanism or other chemical features. Probably the most fascinating properties of protein structures are the locations of deeply concave surface clefts and internal cavities, both of which may point to the area of a cofactor binding site or active site. Different tools register hydrogen-bonding patterns or investigate intramolecular interactions. An especially intriguing properties are the electrostatic potential field encompassing the protein and other electrostatically controlled parameters, for example, individual amino acid pKa, protein solvation energies, and binding constants.

Protein structure alignment and comparison

Notwithstanding when two gene sequences aren't obviously homologous, the structures of the proteins they encode can be similar. New instruments for computing structural similarity are making it conceivable to recognize distant homologies by comparing structures, even without much sequence similarity. These tools also are helpful for comparing developed homology models with the known protein structures they are based on.

Biochemical simulation

Biochemical simulation utilizes the instruments of dynamical systems modeling to mimic the chemical reactions involved in metabolism. Simulations can reach out from individual metabolic pathways to transmembrane transport process and even properties of entire cells or tissues. Biochemical and cell simulations generally depended on the capacity of the researcher to describe a system mathematically, building up an arrangement of differential conditions that represent the different reactions and fluxes occurring in the system. In any case, new software tools can develop the mathematical framework of a simulation automatically from a description given interactively by the user. This make mathematical modeling accessible to any biologist who knows enough about a system to describe it according to the conventions of dynamical systems modeling.

Whole genome analysis

As more and more genomes are sequenced completely, the analysis of raw genome data has become a more important task. There are a number of perspectives from which one can look at genome data: for example, it can be treated as a long linear sequence, but it's often more useful to integrate DNA sequence information with existing genetic and physical map data. This allows you to navigate a very large genome and find what you want. The National Center for Biotechnology Information (NCBI) and other organizations are making a concerted effort to provide useful web interfaces to genome data, so that users can start from a high-level map and navigate to the location of a specific gene sequence.

Genome navigation is far from the only issue in genomic sequence analysis, however. Annotation frameworks, which integrate genome sequence with results of gene finding analysis and sequence homology information, are becoming more common, and the challenge of making and analyzing complete pairwise comparisons between genomes is beginning to be addressed.

Primer design

Many molecular biology protocols require the design of oligonucleotide primers. Proper primer design is critical for the success of polymerase chain reaction (PCR), oligo hybridization, DNA sequencing, and microarray experiments. Primers must hybridize with the target DNA to provide a clear answer to the question being asked, but, they must also have appropriate physicochemical properties;

they must not self-hybridize or dimerize; and they should not have multiple targets within the sequence under investigation. There are several web-based services that allow users to submit a DNA sequence and automatically detect appropriate primers, or to compute the properties of a desired primer DNA sequence.

DNA microarray analysis

DNA microarray analysis is a relatively new molecular biology method that expands on classic probe hybridization methods to provide access to thousands of genes at once. Microarray experiments are amenable to computational analysis because of the uniform, standardized nature of their results—a grid of equally sized spots, each identifiable with a particular DNA sequence. Computational tools are required to analyze larger microarrays because the resulting images are so visually complex that comparison by hand is no longer feasible.

The main tasks in microarray analysis as it's currently done are an image analysis step, in which individual spots on the array image are identified and signal intensity is quantitated, and a clustering step, in which spots with similar signal intensities are identified. Computational support is also required for the chip -design phase of a microarray experiment to identify appropriate oligonucleotide probe sequences for a particular set of genes and to maintain a record of the identity of each spot in a grid that may contain thousands of individual experiments.

Proteomics analysis

Before they're at any point crystallized and biochemically characterized, proteins are frequently analysid utilizing a combination of gel electrophoresis, partial sequencing, and mass spectroscopy. 2D gel electrophoresis can separate a mixture of thousands of proteins into particular segments; the individual spots of material can be blotted or even cut from the gel and examined. Simple computational instruments can give some data to help in the process of analyzing the protein mixtures. It's easier to calculate the molecular weight and pI from a protein sequence; by utilizing these values, sets of putative candidate identities can be identified for each spot on a gel. It's also conceivable to compute, from a protein sequence, the *peptide fingerprint* that is made when that protein is broken down into fragments by enzymes with specific protein cleavage sites. Mass spectrometry investigations of protein fragments can be compared with processed peptide fingerprints to further limit the search.

The Public Biological Databases

The nomenclature problem in biology at the molecular level is immense. Genes are commonly known by unsystematic names. These may come from developmental biology studies in model systems, so that some genes have names like *flightless*, *shaker*, and *antennapedia* due to the developmental effects they cause in a particular animal. Other names are chosen by cellular biologists and represent the function of genes at a cellular level, like *homeobox*. Still other names are chosen by biochemists and structural biologists and refer to a protein that was probably isolated and studied before the gene was ever found.

Though proteins are direct products of genes, they are not always referred to by the same names or codes as the genes that encode them. This kind of confusing nomenclature generally means that only a scientist who works with a particular gene, gene product, or the biochemical process that it's a part of can immediately recognize what the common name of the gene refers to. The biochemistry of a single organism is a more complex set of information than the taxonomy of living species was at the time of Linnaeus, so it isn't to be expected that a clear and comprehensive system of nomenclature will be arrived at easily. There are many things to be known about a given gene: its source organism, its chromosomal

location, and the location of the activator sequences and identities of the regulatory proteins that turn it on and off. Genes also can be categorized by when during the organism's development they are turned on, and in which tissues expression occurs. They can be categorized by the function of their product, whether it's a structural protein, an enzyme, or a functional RNA. They can be categorized by the identity of the metabolic pathway that their product is part of, and by the substrate it modifies or the product it produces. They can be categorized by the structural architecture of their protein products. Clearly this is a wealth of information to be condensed into a reasonable nomenclature. Figure 5 shows a portion of the information that may be associated with a single gene.

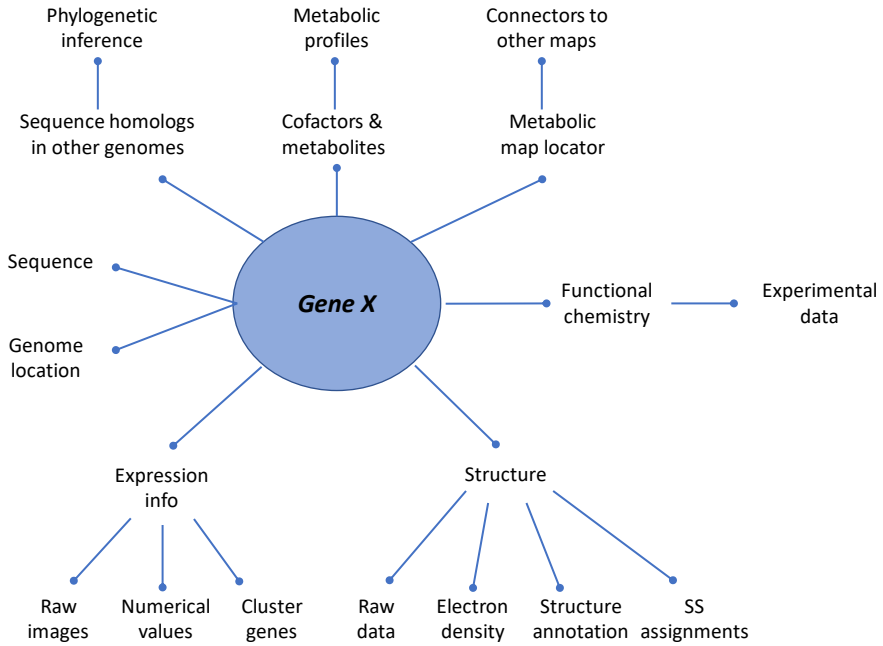


Figure 5. Information associated with a single gene

The issue for maintainers of biological databases turns out to be mostly one of annotation; that is, putting adequate data into the database that there is no doubt of what the gene is, regardless of whether it has a cryptic common name, and making the best possible links between that data and the gene sequence and serial number. Correct annotation of genomic data is a dynamic research area itself, as scientists attempt to discover approaches to exchange data crosswise over genomes without spreading error. Storage of macromolecular information in electronic databases has offered ascend to a method for working around the issue of classification. The solution has been to give each new entry into the database a serial number and afterward to store it in a relational database that knows the correct linkages between that serial number, any number of names for the gene or gene product it encodes, and all manner of other information about the gene. This technique is the the one currently in use in the major biological databases.

The questions databases resolve are essentially the same questions that arise in developing a nomenclature. However, by using relational databases and complex querying strategies, they (perhaps somewhat unfortunately) avoid the issue of finding a concise way for scientists to communicate the identities of genes on a nondigital level.

Data Annotation and Data Formats

The representation and distribution of biological data is still an open problem in bioinformatics. The nucleotide sequences of DNA and RNA and the amino acid sequences of proteins reduce neatly to

character strings in which a single letter represents a single nucleotide or amino acid. The remaining challenges in representing sequence data are verification of the correctness of the data, thorough annotation of data, and handling of data that comes in ever-larger chunks, such as the sequences of chromosomes and whole genomes.

The standard reduced representation of the 3D structure of biomolecule consists of the Cartesian coordinates of the atoms in the molecule. This aspect of representing the molecule is straightforward. On the other hand, there are a host of complex issues for structure databases that are not completely resolved. Annotation is still an issue for structural data, although the biology community has attempted to form a consensus as to what annotation of a structure is currently required. In the last 15 years, different researchers have developed their own styles and formats for reporting biological data. Biological sequence and structure databases have developed in parallel in the United States and in Europe. The use of proprietary software for data analysis has contributed a number of proprietary data formats to the mix. While there are many specialized databases, we focus here on the fields in which an effort is being made to maintain a comprehensive database of an entire class of data.

3D Molecular Structure Data

Though DNA sequence, protein sequence, and protein structure are in some sense just different ways of representing the same gene product, these datatypes currently are maintained as separate database projects and in unconnected data formats. This is mainly because sequence and structure determination methods have separate histories of development.

The first public molecular biology database, set up about 10 years before the public DNA sequence databases, was the Protein Data Bank (PDB). It represents the central repository for x-ray crystal structures of protein molecules. While the first finish protein structure was presented in the 1950s, there were not a noteworthy number of protein structures accessible until the late 1970s. Computers had not created to the point where graphical representation of protein coordinate structure information was possible, at least at useful speeds. However, in 1971, the PDB was set up at the Brookhaven National Laboratory, to store protein structure information in a computer-based archive. A data format created, which owed a lot of its style to the prerequisites of early computer technology. All through the 1980s, the PDB grew. From 15 sets of entries in 1973, it augments to 69 entries in 1976. The number of coordinate sets deposited each year remained under 100 until 1988, at which time there were still fewer than 400 PDB entries.

In the vicinity of 1988 and 1992, the PDB hit the the turning point in its exponential growth curve. By January 1994, there were 2,143 entries in the PDB; and at the moment the PDB has more than 14,000 entries. Administration of the PDB has been exchanged to a consortium of entry mark, called the Research Collaboratory for Structural Bioinformatics, and and a new format for recording of crystallographic data, the Macromolecular Crystallographic Information File (mmCIF), is being introduced in to replace the antiquated PDB format. Journals that publish crystallographic results require submission to the PDB as a condition of publication, which means that nearly all protein structure data obtained by academic researchers becomes available in the PDB.

A typical issue for information driven investigations of protein structure is the excess and absence of thoroughness of the PDB. There are numerous proteins for which various crystal structures have been submitted to the database. Choosing subsets of the PDB information with which to work is in this manner a critical step in any statistical investigation of protein structure. Numerous statistical studies of protein structure depend on sets of protein chains that have close to 25% of their sequence in common; if this paradigm is utilized, there are still just around 1,000 unique protein folds represented in the PDB. As the amount of biological sequence data available has grown, the PDB now falls a long ways behind the gene-sequence databases.

DNA, RNA, and Protein Sequence Data

Sequence databases generally specialize in one type of sequence data: DNA, RNA, or protein. There are major sequence data collections and deposition sites in Europe, Japan, and the United States, and there are independent groups that mirror all the data collected in the major public databases, often offering some software that adds value to the data.

In 1970, Ray Wu sequenced the first segment of DNA; twelve bases that occurred as a single strand at the end of a circular DNA that was opened utilizing a cleaving enzyme. In any case, DNA sequencing demonstrated considerably more troublesome than protein sequencing, on the grounds that there is no chemical process that selectively cleaves the first nucleotide from a nucleic acid chain. At the point when Robert Holley announced the sequencing of a 76-nucleotide RNA molecule from yeas, it was following seven years of work. After Holley's sequence was published, different groups refined the protocols for sequencing, even succeeding in sequence effectively a 3,200-base bacteriophage genome. Genuine advance with DNA sequencing came after 1975, with the chemical cleavage method created by Allan Maxam and Walter Gilbert, and with Frederick Sanger's chain terminator procedure.

The first DNA sequence database, established in 1979, was the Gene Sequence Database (GSDB) at Los Alamos National Lab. While GSDB has since been supplanted by the worldwide collaboration that is the modern GenBank, up-to-date gene sequence information is still available from GSDB through the National Center for Genome Resources.

The European Molecular Biology Laboratory, the DNA Database of Japan, and the National Institutes of Health cooperate to make all freely accessible sequence data through GenBank. NCBI has built up a standard relational database format for sequence information presentation and storage, known as the ASN.1 format. While this format guarantees to locate the right sequences of the right kind in GenBank simpler, there are also various services tions giving access to nonredundant versions of the database. The DNA sequence database developed gradually through its first decade. In 1992, GenBank contained just 78,000 DNA sequences — a little more than 100 million pairs of DNA. In 1995, the Human Genome Project, and advances in sequencing innovation, kicked GenBank's growth into high gear. GenBank currently doubles in size every 6 to 8 months, and its rate of increase is constantly growing.

Genomic Data

In addition to the Human Genome Project, there are now separate genome project databases for a large number of model organisms. The sequence content of the genome project databases is represented in GenBank, but the genome project sites also provide everything from genome maps to supplementary resources for researchers working on that organism. As of October 2000, NCBI's Entrez Genome database contained the partial or complete genomes of over 900 species. Many of these are viruses. The remainder include bacteria; archaea; yeast; commonly studied plant model systems such as *A. thaliana*, rice, and maize; animal model systems such as *C. elegans*, fruit flies, mice, rats, and puffer fish; as well as organelle genomes. NCBI's web-based software tools for accessing these databases are constantly evolving and becoming more sophisticated.

Biochemical Pathway Data

The most vital biological activities don't occur by the action of single molecule, however as the orchestrated activities of multiple molecules. Since the mid twentieth century, biochemists have analyzed these functional ensembles of enzymes and their substrates. A couple of research groups have started work at intelligently arranging and storing these pathways in databases. Key example of pathway database is KEGG. The Kyoto Encyclopedia of Genes and Genomes (KEGG) stores comparative

information about sequence, structure, and genetic linkage databases. This database is queryable through web interfaces and are curated by a combination of automation and human expertise. In addition to these whole genome "parts catalogs," other, more specialized databases that focus on specific pathways (such as intercellular signaling or degradation of chemical compounds by microbes) have been developed.

Gene Expression Data

DNA microarrays (or *gene chips*) are miniaturized laboratories for the study of gene expression. Each chip contains a deliberately designed array of probe molecules that can bind specific pieces of DNA or mRNA. Labeling the DNA or RNA with fluorescent molecules allows the level of expression of any gene in a cellular preparation to be measured quantitatively. Microarrays also have other applications in molecular biology, but their use in studying gene expression has opened up a new way of measuring genome functions.

Since the advancement of DNA microarray technology in the late 1990s, it has turned out that the increase in available gene expression data will eventually parallel the growth of the sequence and structure databases. Raw microarray information has been started to be made accessible to the general audience in particular databases, and the building up of a central data repository for such data is done (Gene Expression Omnibus).

Since a significant number of the early microarray experiments were performed at Stanford, their genome resources site has connections to raw information and databases that can be queried utilizing gene names or functional descriptions. Furthermore, the European Bioinformatics Institute has been instrumental in setting up of standards for deposition of microarray data in databases. Several databases additionally exist for the deposition of 2D gel electrophoresis results, including SWISS-2DPAGE and HSC-2DPAGE. 2D-PAGE is an innovation that permits quantitative investigation of protein concentrations in the cell, for many proteins at the same time. The combination of these two systems is an intense tool for understanding how genomes function.

Table 1 summarizes sources on the Web for some of the most important databases we've discussed in this section.

Table 1. Major Biological Data and Information Sources

Subject	Source	Link
Biomedical literature	PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Nucleic acid sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide
	SRS at EMBL/EBI	http://srs.ebi.ac.uk
Genome sequence	Entrez Genome	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
	TIGR databases	http://www.tigr.org/tdb/
Protein sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein
	SWISS-PROT at ExPASy	http://www.expasy.ch/spro/
	PIR	http://www-nbrf.georgetown.edu
Protein structure	Protein Data Bank	http://www.rcsb.org/pdb/
	Entrez Structure DB Protein and peptide mass spectroscopy PROWL	http://prowl.rockefeller.edu
Post-translational modifications	RESID	http://www-nbrf.georgetown.edu/pirwww/search/textresid.html
Biochemical and biophysical information	ENZYME	http://www.expasy.ch/enzyme/
	BIND	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Structure
Biochemical pathways	PathDB	http://www.ncgr.org/software/pathdb/
	KEGG	http://www.genome.ad.jp/kegg/
	WIT	http://wit.mcs.anl.gov/WIT2/
Microarray	Gene Expression Links	http://industry.ebi.ac.uk/~alan/MicroArray/
2D-PAGE	SWISS-2DPAGE	http://www.expasy.ch/ch2d/ch2d-top.html
Web resources	The EBI Biocatalog	http://www.ebi.ac.uk/biocat/
	IUBio Archive	http://iubio.bio.indiana.edu

References

1. Baxevanis A.D., Ouellette B. F. F. (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition, John Wiley & Son, New York
2. Elloumi M., Zomaya A. Y. (2011) Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley a& Son, New York
3. Liu L., Agren R., Bordel S., Nielsen J. (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Letters 584: 2556–2564.
4. Milne C.B., Kim P.J., Eddy J.A., Price N.D. (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnol J. 4(12):1653-70
5. Pevzner P., Shamir R. (2011) Bioinformatics for Biologists, 1st Edition, Cambrage University Press
6. Ramsden J. (2015) Bioinformatics: An Introduction, Springer-Verlag, London
7. Singh G. B. (2015) Fundamentals of Bioinformatics and Computational Biology, Springer International Publishing, Switzerland

Alignments and phylogenetic trees

Basic level

Ventsislava Petrova
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Kliment Petrov
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Contents

Bioinformatics tools

There are several tools that study protein and DNA sequences, the most abundant type of biological data available electronically. The importance of sequence databases is from crucial importance to biological investigations and the pairwise sequence comparison is the most essential technique in bioinformatics. It allows you to search sequence-based datasets, to build evolutionary trees, to recognize specific features of protein families, to create homology models. But it's also the key for the development of larger projects,

such as analyzing whole genomes, exploring the sequence determinants of protein structure, connecting expression data to genomic information, etc.

The following types of analysis can be performed by using sequence data:

- Single sequence analysis and sequence characterization
- Pairwise alignment and DNA / protein sequence searching
- Multiple sequence alignment
- Sequence motif discovery in multiple alignments
- Phylogenetic analysis

Pairwise sequence comparison is the main tool of connecting biological function with genome and of transferring known information from one genome to another. The techniques for analysis of biological sequences is the most significant approaches for sequence data assessment. There are numerous freely accessible software tools for performing pairwise sequence comparison. Some of them are summarized in Table 1.

Table 1. Sequence Analysis Tools and Techniques

What you do	**Why you do it**	**What you use to do it**
Gene finding	Identify possible coding regions in genomic DNA sequences	GENSCAN, GeneWise, PROCRUSTES, GRAIL
DNA feature detection	Locate splice sites, promoters, and sequences involved in regulation of gene expression	CBS Prediction Server
DNA translation and reverse translation	Convert a DNA sequence into protein sequence or vice versa	"Protein machine" server at EBI
Pairwise sequence alignment (local)	Locate short regions of homology in a pair of longer sequences	BLAST, FASTA
Pairwise sequence alignment (global)	Find the best full-length alignment between two sequences	ALIGN
Sequence database search by pairwise comparison	Find sequence matches that aren't recognized by a keyword search; find only matches that actually have some sequence homology	BLAST, FASTA, SSEARCH

Mechanisms of Molecular Evolution

The discovery of DNA as the molecular basis of heredity and evolution made it possible to understand the process of evolution in a whole new way. It is known that often mutations occur in different parts of an organism's DNA: in the middle of genes that code for proteins or functional RNA molecules, in the middle of regulatory sequences that govern whether a gene to be expressed or not, or in the "middle of nowhere", in the regions between gene sequences. Mutations can have important effects on the organism's phenotype or they can have no apparent consequence. Over time mutations that are beneficial or at least not harmful to a species can become fixed in the population.

By comparative study of DNA sequences or of whole genomes, it's possible to develop quantitative methods for understanding when and how mutational events occurred, as well as how and why they were preserved to survive in existing species and populations. Genomics and bioinformatics have made it possible to study the evolutionary record and make statements about the phylogenetic relationship of one species to another. Changes in the identity of the residue (nucleotide or amino acid) at a given position in the sequence are scored using standard substitution scores (for example, a positive score for a match and a negative score for a mismatch) or substitution matrices. Insertions and deletions are scored with penalties for gap opening and gap extension.

Genefinders and DNA Features Detection

Once a large piece of DNA has been mapped and sequenced, the next important task is to understand its function. Analysis of single DNA for sequence features is a rapidly growing research area in bioinformatics. There are two reasons that genefinding and feature detection represent difficult problems. First, there are a huge number of protein-DNA interactions, many of which have not yet been experimentally characterized, and some of which differ from organism to organism. Current promoter detection algorithms yield about 20-40 false positives for each real promoter identified. Some proteins bind to specific sequences; others are more flexible and recognize different attachment sites. To complicate matters further, a protein can bind in one part of a chromosome but affect completely different region hundreds or thousands of base pairs away.

Genefinders are programs that try to identify all the open reading frames in unannotated DNA. They use a variety of approaches to locate genes, but the most successful combine content-based and pattern-recognition approaches. Content-based tools for gene prediction take advantage of the fact that the distribution of nucleotides in genes is different than in non-genes. Pattern-recognition methods look for characteristic sequences associated with genes (start and stop codons, promoters, splice sites) to deduce the presence and structure of a gene. In fact, the current generation of genefinders combine both methods with additional knowledge, such as gene structure or sequences of other, known genes.

Some genefinders are accessible only though web interfaces: the sequence that needs to be examined for genes is submitted to the program, it is processed, and the corresponding result is returned. On one hand, this eliminates the need for installation and maintenance of the specific software on your system, and it provides a relatively uniform interface for the different programs. On the other, if you plan to rely on the results of a genefinder, you should take the time to understand underlying algorithm, find out if the model is specific for a given species or family, and, in the case of content-based models, know which sequences they are.

Some frequently used programs in gene finding include Oak Ridge National Labs' GRAIL, GENSCAN, PROCRUSTES, and GeneWise. GRAIL combines evidence from a variety of signal and

content information using a neural network. GENSCAN combines information about content statistics with a probabilistic model of gene structure. PROCRUSTES and GeneWise find open reading frames by translating the DNA sequence and comparing the resulting protein sequence with known protein sequences. PROCRUSTES compares potential ORFs with close homologs, while GeneWise compares the gene against a single sequence or a model of an entire protein family.

Feature Detection

In addition to their role in genefinder systems, feature-detection algorithms can be used on their own to find patterns in DNA sequences. Frequently, these tools help interpret newly sequenced DNA or choose targets for designing PCR primers or microarray oligomers. Some starting places for tools like these include the Center for Biological Sequence Analysis at the Technical University of Denmark, the CodeHop server at the Fred Hutchinson Cancer Research Center, and the Tools collection at the European Bioinformatics Institute. In addition to these special-purpose tools, another popular approach is to use motif discovery programs that automatically find common patterns in sequences.

DNA Translation

Before a protein can be synthesized, its sequence must be translated from the DNA into protein sequence. However, any DNA sequence can be translated in six possible ways. The sequence can be translated backward and forward. Because each amino acid in a protein is specified by three bases in the DNA sequence, there are three possible translations of any DNA sequence in each direction: one beginning with the very first character in the sequence, one beginning with the second character, and one beginning with the third character.

Figure 1 shows "back-translation" of a protein sequence (shown on the top line) into DNA, using the bacterial and plant plastid genetic code. However, note that nature has grouped the codons "sensibly": alanine (A) is always specified by a "G-C-X" codon, arginine (R) is specified either by a "C-G-X" codon or an "A-G-pyrimidine" codon, etc. This reduces the number of potential sequences that have to be checked if you (for example) try to write a program to compare a protein sequence to a DNA sequence database.

The more computationally efficient solution to this problem is simply to translate the DNA sequence database in all six reading frames.

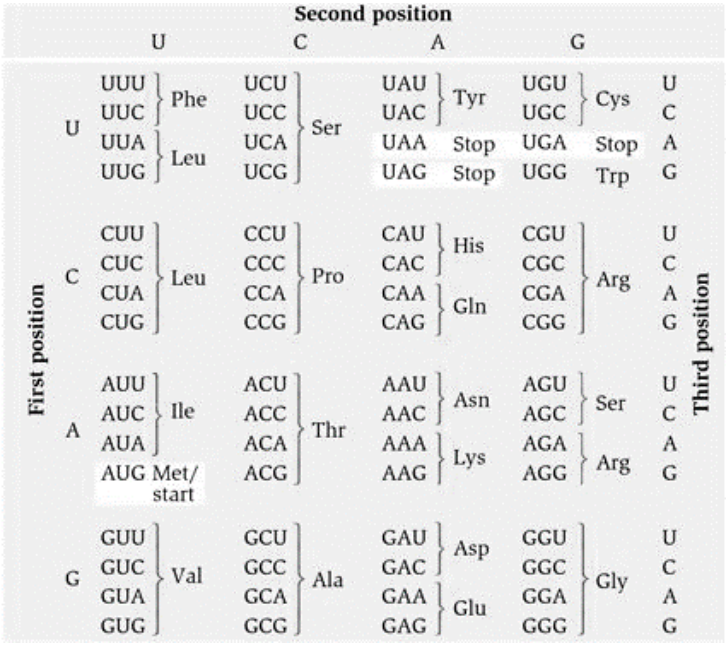


Figure 1. Back-translation from a protein sequence

There are no markers in the DNA sequence to indicate where one codon ends and the next one begins. Consequently, unless the location of the start codon is known ahead of time, a double-stranded DNA sequence can be interpreted in any of six ways: an open reading frame can start at nucleotide *i*, at *i*+1, or at *i*+2 on either of both DNA strand. To interpret this uncertainty, when a protein is compared with a set of DNA sequences, the DNA sequences are translated into all six possible amino acid sequences, and the protein query sequence is compared with these resulting conceptual translations. This exhaustive translation is called a "six-frame translation" and is illustrated in Figure 2.

```
  M  S  K  L  G  Q  E  K  N  E  V  N  Y  S  D  V  R  E  D  R     F1
   C  R  N  W  D  K  K  K  M  K  *  I  T  L  M  *  E  R  I  E    F2
    V  E  I  G  T  R  K  K  *  S  K  L  L  *  C  K  R  G  *  S   F3
1 ATGTCGAAATTGGGACAAGAAAAAAATGAAGTAAATTACTCTGATGTAAGAGAGGATAGA 60
  ----:----|----:----|----:----|----:----|----:----|----:----|
1 TACAGCTTTAACCCTGTTCTTTTTTTACTTCATTTAATGAGACTACATTCTCTCCTATCT 60
  X  D  F  N  P  C  S  F  F  S  T  F  *  E  S  T  L  S  S  L    F6
  X  T  S  I  P  V  L  F  F  H  L  L  N  S  Q  H  L  L  P  Y    F5
    H  R  F  Q  S  L  F  F  I  F  Y  I  V  R  I  Y  S  L  I  S  F4


  V  V  T  N  S  T  G  N  P  I  N  E  P  F  V  T  Q  R  I  G     F1
   L  *  Q  T  P  L  V  I  Q  S  M  N  H  L  S  P  N  V  L  G    F2
    C  D  K  L  H  W  *  S  N  Q  *  T  I  C  H  P  T  Y  W  G   F3
61 GTTGTGACAAACTCCACTGGTAATCCAATCAATGAACCATTTGTCACCCAACGTATTGGG 120
  ----:----|----:----|----:----|----:----|----:----|----:----|
61 CAACACTGTTTGAGGTGACCATTAGGTTAGTTACTTGGTAAACAGTGGGTTGCATAACCC 120
   T  T  V  F  E  V  P  L  G  I  L  S  G  N  T  V  W  R  I  P    F6
  L  Q  S  L  S  W  Q  Y  D  L  *  H  V  M  Q  *  G  V  Y  Q     F5
    N  H  C  V  G  S  T  I  W  D  I  F  W  K  D  G  L  T  N  P   F4
```

Figure 2. A DNA sequence and its translation in three of six possible reading frames

Because of the large number of codon possibilities for some amino acids, back-translation of a protein into DNA sequence can result in an extremely large number of possible sequences. However, codon usage statistics for different species are available and can be used to suggest the most likely backtranslation out of the range of possibilities. However, if you need to produce a six-frame translation of a single DNA sequence or translate a protein back into a set of possible DNA sequences, and you don't want to script it yourself, the Protein Machine server at the European Bioinformatics Institute (EBI) will do it for you.

Pairwise Sequence Comparison

Comparison of protein and DNA sequences is one of the fundamentals of bioinformatics. The ability to perform rapid automated comparisons of sequences facilitates assignment of function to a new sequence, prediction and construction of model protein structures, design and analysis of gene expression experiments. As biological sequence data has accumulated, it has become apparent that nature is conservative. A new biochemistry isn't created for each new species, and new functionality isn't created by the sudden appearance of whole new genes. Instead, incremental modifications give rise to genetic diversity and novel function. Thus, detection of similarity between sequences allows transferring of information about one sequence to other similar sequences with reasonable, though not always total, confidence.

Before making a comparative conclusion about one nucleic acid or protein sequence, a sequence alignment is required. The basic concept of selecting an optimal sequence alignment is simple. The two sequences are matched up in an arbitrary way. The quality of the match is scored. Then one sequence is moved with respect to the other and the match is scored again, until the best-scoring alignment is found.

What sounds simple in principle isn't at all simple in practice. So, using an automated method for finding the optimal alignment is the most suitable approach. Next question is how should alignments be scored? A scoring scheme can be as simple as +1 for a match and -1 for a mismatch. But, should gaps be allowed to open in the sequences to facilitate better matches elsewhere? If gaps are allowed, how should they be scored? What is the best algorithm for finding the optimal alignment of two sequences? And when an alignment is produced, is it necessarily significant? Can an alignment of similar quality be produced for two random sequences?

Figure 3 shows examples of three kinds of alignment. In each alignment, the sequences being compared are displayed, one above the other, such that matching residues are aligned. Similarities are indicated with plus (+). Information about the alignment is presented at the top, including percent identity (the number of identical matches divided by the length of the alignment) and score. Finally, gaps in one sequence relative to another are represented by dashes (-) for each position in that sequence occupied by a gap.

```
Score = 27.7 bits (60), Expect = 2.1,   Method: Composition-based stats.
 Identities = 15/72 (20%), Positives = 34/72 (47%), Gaps = 15/72 (20%)

Query: 9   SEFDSAIAQDKLVVVDFYATWCGPCKMIAPMIEKFSEQYPQA-DFYKLDVDELGDVAQKN 67
           SE++S + +DKL ++D              + +    ++P + DF ++D+  +    QK
Sbjct: 273 SEYNSIVHEDKLYIID--------------VSQSVQPEHPMSLDFLRMDIKNVNSYFQKL 318

Query: 68  EVSAMPTLLLFK 79
            +   P  ++F+
Sbjct: 319 GIDIFPERVIFQ 330


Score =  176 bits (447), Expect = 5e-59,   Method: Compositional matrix
adjust. Identities = 86/102 (84%), Positives = 92/102 (90%)

Query: 1   MVTQFKTASEFDSAIAQDKLVVVDFYATWCGPCKMIAPMIEKFSEQYPQADFYKLDVDEL 60
           MV Q  + SEFDSAIA DKLVVVDF+ATWCGPCKMIAPMIEKF+ +Y  ADFYKLDVDEL
Sbjct: 1   MVKQITSVSEFDSAIAVDKLVVVDFFATWCGPCKMIAPMIEKFAAEYSTADFYKLDVDEL 60

Query: 61  GDVAQKNEVSAMPTLLLFKNGKEVAKVVGANPAAIKQAIAAN 102
            +VAQKNEVSAMPTL+LFKNGKEVAKVVGANPAAIKQAIA N
Sbjct: 61  PEVAQKNEVSAMPTLVLFKNGKEVAKVVGANPAAIKQAIANN 102


Score = 69.7 bits (169), Expect = 2e-15,   Method: Compositional matrix
adjust. Identities = 29/96 (30%), Positives = 60/96 (62%), Gaps = 2/96 (2%)

Query: 6   KTASEFDSAIAQDKLVVVDFYATWCGPCKMIAPMIEKFSEQYP--QADFYKLDVDELGDV 63
           K+  +F+  ++ +K +V +F A WCGPC+ I P+++ F ++    + D  ++D+D  G++
Sbjct: 9   KSQQDFELYLSNNKYLVANFTAQWCGPCQQIKPVVDNFYQETEGQKFDIVRVDLDSQGEL 68

Query: 64  AQKNEVSAMPTLLLFKNGKEVAKVVGANPAAIKQAI 99
           A K  ++A+PT +  +   EV ++ GA+ +A+  A+
Sbjct: 69  ASKYAITAVPTFIFLEGKNEVNRIRGADTSALLTAL 104
```

Figure 3. Three alignments: random, high scoring, and low scoring but meaningful

The first alignment is a random alignment, a comparison between two unrelated sequences. Notice that, in addition to the few identities and conservative mutations between the two, large gaps have been opened in both sequences to achieve this alignment. Second alignment is a high-scoring one: it shows a comparison of two closely related proteins. Compare that alignment with the third, a comparison of two distantly related proteins. It shows that fewer identical residues are shared by the sequences in the low-scoring alignment than in the high-scoring one. Still, there are several similarities or conservative changes.

In describing sequence comparisons, several different terms are frequently used. Sequence identity, sequence similarity, and sequence homology are the most important. *Sequence similarity* is meaningful only when possible substitutions are scored according to the probability with which they occur. In protein sequences, amino acids of similar chemical properties are found to substitute for each other much more readily than dissimilar amino acids. *Sequence homology* is a more general term that indicates evolutionary relatedness among sequences. It is common to speak of a percentage of sequence homology when comparing two sequences, although that percentage may include a mixture of identical and similar sites. Finally, sequence homology refers to the evolutionary relatedness between sequences. Two sequences are said to be homologous if they are both derived from a common ancestral sequence. The terms similarity and homology are often used interchangeably to describe sequences, but, however, they mean different things. Similarity refers to the presence of identical and similar sites in the two sequences, while homology reflects a sharing of a common ancestor.

Scoring Matrices

The most important information when evaluating a sequence alignment is whether it is random, or meaningful. If the alignment is meaningful, the question is how meaningful it is. This is assessed by constructing a scoring matrix. A scoring matrix is a table of values that describe the probability of a residue (amino acid or base) pair occurring in an alignment. The values in a scoring matrix are logarithms of ratios of two probabilities. One is the probability of random occurrence of an amino acid in a sequence alignment. This value is simply the product of the independent frequencies of occurrence of each of the amino acids. The other is the probability of meaningful occurrence of a pair of residues in a sequence alignment. These probabilities are derived from samples of actual sequence alignments that are known to be valid.

Figure 4 shows an example of a BLOSUM62 substitution matrix for amino acids.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-2	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	0	-3	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	-2	-2	-2	-2	0	1	0	-2	-3
H	-2	-3	1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	0	-2	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-2	1	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	0	-3	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-2	3	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	-2	-1	-1
N	-2	-3	1	0	-3	0	-1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-1	7	-1	-2	-1	1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	0	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	-1	-1	1	0	-2	1	0	-2	0	-2	-1	0	1	0	-1	1	4	-2	-3	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	-2	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-3	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

Figure 4. The BLOSUM62 substitution matrix for amino acids

Substitution matrices for amino acids are complicated because they reflect the chemical nature and frequency of occurrence of the amino acids. For example, in the BLOSUM matrix, glutamic acid (E) has a positive score for substitution with aspartic acid (D) and also with glutamine (Q). Both these substitutions are chemically conservative. Aspartic acid has a sidechain that is chemically similar to glutamic acid, though one methyl group shorter. On the other hand, glutamine is similar in size and chemistry to glutamic acid, but it is neutral while glutamic acid is negatively charged. Substitution scores for glutamic acid with residues such as isoleucine (I) and leucine (L) are negative

Substitution matrices for bases in DNA or RNA sequence are very simple. In most cases, it is reasonable to assume that A:T and G:C occur in roughly equal proportions. Commonly used substitution

matrices include the BLOSUM and PAM matrices. When using BLAST, you need to select a scoring matrix. Most automated servers select a default matrix for you, and if you're just doing a quick sequence search, it's fine to accept the default.

BLOSUM matrices are derived from the Blocks database. The numerical value (e.g., 62) associated with a BLOSUM matrix represents the cutoff value for the clustering step. A value of 62 indicates that sequences were put into the same cluster if they were more than 62% identical. By allowing more diverse sequences to be included in each cluster, lower cutoff values represent longer evolutionary time scales, so matrices with low cutoff values are appropriate for seeking more distant relationships. BLOSUM62 is the standard matrix for ungapped alignments, while BLOSUM50 is more commonly used when generating alignments with gaps.

Point accepted mutation (PAM) matrices are scaled according to a model of evolutionary distance from alignments of closely related sequences. The most commonly used PAM matrix is PAM250. However, comparison of results using PAM and BLOSUM matrices suggest that BLOSUM matrices are better at detecting biologically significant similarities.

Gap Penalties

DNA sequences change not only by point mutation, but by insertion and deletion of residues as well. Consequently, it is often necessary to introduce gaps into one or both of the sequences being aligned to produce a meaningful alignment between them. Most algorithms use a gap penalty for the introduction of a gap in the alignment. Most sequence alignment models use affine gap penalties, in which the rate of opening a gap in a sequence is different from the rate of extending a gap that has already been started. Of these two penalties—-the gap opening penalty and the gap extension penalty—-the gap opening penalties tend to be much higher than the associated extension penalty. Scores of -11 for gap opening and -1 for gap extension are commonly used in conjunction with the BLOSUM 62 matrix.

Global Alignment

One possibility is to align two sequences along their whole length. This algorithm is called the Needleman-Wunsch algorithm. In this case, an optimal alignment is built up from high-scoring alignments of subsequences, stepping through the matrix from top left to bottom right. Only the best-scoring path can be traced through the matrix, resulting in an optimal alignment.

Local Alignment

The most commonly used sequence alignment tools rely on a strategy called local alignment. The global alignment strategy assumes that the two sequences to be aligned are known and are to be aligned over their full length. However, often a sequence is searched against a sequence database with unknown sequences, or a short query sequence is used to match with a very long DNA sequence. For example, in protein or gene sequences that do have some evolutionary relatedness, but which have diverged significantly from each other, short homologous segments may be all the evidence of sequence homology that remains. The algorithm that performs local alignment of two sequences is known as the Smith-Waterman algorithm. A local alignment isn't required to extend from beginning to end of the two sequences being aligned. If the

cumulative score up to some point in the sequence is negative, the alignment can be abandoned and a new alignment started. The alignment can also end anywhere in the matrix.

Tools for local alignment

One of the most frequently reported implementations of the Smith-Waterman algorithm for database searching is the program SSEARCH, which is part of the FASTA distribution. LALIGN, also part of the FASTA package, is an implementation of the Smith-Waterman algorithm for aligning two sequences.

Sequence Queries Against Biological Databases

A common application of sequence alignment is searching a database for sequences that are similar to a query sequence. In these searches, an alignment of a sequence hundreds or thousands of residues long is matched against a database of at least tens of thousands of comparably sized sequences.

Local Alignment-Based Searching Using BLAST

By far, the most popular tool for searching sequence databases is a program called BLAST (Basic Local Alignment Search Tool). It performs pairwise comparisons of sequences, seeking regions of local similarity, rather than optimal global alignments between whole sequences. BLAST can perform hundreds or even thousands of sequence comparisons in a matter of minutes. And in less than a few hours, a query sequence can be compared to an entire database to find all similar sequences.

The BLAST algorithm

Local sequence alignment searching using a standard Smith-Waterman algorithm is a fairly slow process. The BLAST algorithm, which speeds up local sequence alignment, has three basic steps. First, it creates a list of all short sequences (called *WORDS*) that score above a threshold value when aligned with the query sequence. Next, the sequence database is searched for occurrences of these words. Because the word length is so short (3 residues for proteins, 11 residues for nucleic acids), it's possible to search a precomputed table of all words and their positions in the sequences for improved speed. These matching words are then extended into ungapped local alignments between the query sequence and the sequence from the database. Extensions are continued until the score of the alignment drops below a threshold. The top-scoring alignments in a sequence, or maximal-scoring segment pairs (MSPs), are combined where possible into local alignments. The new additions to the BLAST software package also search for gapped alignments.

NCBI BLAST and WU-BLAST

There are two implementations of the BLAST algorithm: NCBI BLAST and WU-BLAST. Both can be used as web services and as downloadable software packages. NCBI BLAST is available from the National Center for Biotechnology Information (NCBI), while WU-BLAST is developed and maintained at Washington University. NCBI BLAST is the more commonly used of the two. The most recent versions of this program have focused on the development of methods for comparing multiple-sequence profiles. WU-BLAST, on the other hand, has developed a different system for handling gaps as well as a number of features that are useful for searching genome sequences.

Different BLAST programs

The four main executable programs in the BLAST distribution are:

[blastall]

Performs BLAST searches using one of five BLAST programs: *blastp, blastn, blastx, tblastn,* or *tblastx*

[blastpgp]

Performs searches in PSI-BLAST or PHI-BLAST mode

[bl2seq]

Performs a local alignment of two sequences

[formatdb]

Converts a FASTA-format flat file sequence database into a BLAST database

blastall encompasses all the major options for ungapped and gapped BLAST searches. A full list of its command-line arguments can be displayed with the command *blastall* - :

[-p]

Program name. Its options include:

blastp

Protein sequence (PS) query versus PS database

blastn

Nucleic acid sequence (NS) query versus NS database

blastx

NS query translated in all six reading frames versus PS database

tblastn

PS query versus NS database dynamically translated in all six reading frames

tblastx

Translated NS query versus translated NS database—computationally intensive

blastpgp allows you to use two new BLAST modes: PHI-BLAST (Pattern Hit Initiated BLAST) and PSI-BLAST (Position Specific Iterative BLAST). PHI-BLAST uses protein motifs, such as those found in PROSITE and other motif databases, to increase the likelihood of finding biologically significant matches. PSI-BLAST uses an iterative alignment procedure to develop position-specific scoring matrices, which increases its capability to detect weak pattern matches.

bl2seq allows the comparison of two known sequences using the *blastp* or *blastn* programs. Most of the command-line options for *bl2seq* are similar to those for *blastall*.

Evaluating BLAST results

A BLAST search provides three related pieces of information that allow you to interpret its results: raw scores, bit scores, and E-values.

The *raw score* for a local sequence alignment is the sum of the scores of the maximal-scoring segment pairs (MSPs) that make up the alignment. *Bit score*s are raw scores that have been converted from the log base of the scoring matrix that creates the alignment to log base 2. *E-values* provide information about the likelihood that a given sequence alignment is significant. An alignment's E-value indicates the number of alignments one expects to find with a score greater than or equal to the observed alignment's score in a search against a random database. Thus, a large E-value (5 or 10) indicates that the alignment probably has occurred by chance, and that the target sequence has been aligned to an unrelated sequence in the database. E-values of 0.1 or 0.05 are typically used as cutoffs in sequence database searches. Using a larger E-value cutoff in a database search allows more distant matches to be found, but it also results in a higher rate of spurious alignments. Of the three, E values are the values most often reported in the literature.

There is a limit beyond which sequence similarity becomes uninformative about the relatedness of the sequences being compared. This limit is encountered below approximately 25% sequence similarity for protein sequences. In the case of protein sequences with low sequence similarity that are still believed to be related, structural analysis techniques may provide evidence for such a relationship. Where structure is unknown, sequences with low similarity are categorized as unrelated, but that may mean only that the evolutionary distance between sequences is so great that a relationship can't be detected.

Local Alignment Using FASTA

Another method for local sequence alignment is the FASTA algorithm. FASTA precedes BLAST and like BLAST, it is available both as a service over the Web and as a downloadable set of programs.

The FASTA algorithm

FASTA first searches for short sequences (called ktups) that occur in both the query sequence and the sequence database. Then, using the BLOSUM50 matrix, the algorithm scores the 10 ungapped alignments that contain the most identical ktups. These ungapped alignments are tested for their ability to be merged into a gapped alignment without reducing the score below a threshold. For those merged alignments that score over the threshold, an optimal local alignment of that region is then computed, and the score for that alignment (called the optimized score) is reported.

FASTA ktups are shorter than BLAST words, typically 1 or 2 for proteins, and 4 or 6 for nucleic acids. Lower ktup values result in slower but more sensitive searches, while higher ktup values yield faster searches with fewer false positives.

The FASTA programs

The FASTA distribution contains search programs that are analogous to the main BLAST modes, with the exception of PHI-BLAST and PSI-BLAST.

[fasta]

Compares a protein sequence against a protein database (or a DNA sequence against a DNA database) using the FASTA algorithm

[ssearch]

Compares a protein sequence against a protein database (or DNA sequence against a DNA database) using the Smith-Waterman algorithm

[fastx /fasty]

Compares a DNA sequence against a protein database, performing translations on the DNA sequence

[tfastx /tfasty]

Compares a protein sequence against a DNA database, performing translations on the DNA sequence database

[align]

Computes the global alignment between two DNA or protein sequences

[lalign]

Computes the local alignment between two DNA or protein sequences

Multifunctional Tools for Sequence Analysis

Several research groups and companies have assembled web-based interfaces to collections of sequence tools. The best of these have fully integrated tools, public databases, and the ability to save a record of user data and activities from one use to another. If you're searching for matches to just one or a few sequences and you want to search the standard public databases, these portals can save you a lot of time while providing most of the functionality and ease of use of a commercial sequence analysis package.

The Biology Workbench

The Biology Workbench resource is freely available to academic users and offers keyword and sequence-based searching of nearly 40 major sequence databases and over 25 whole genomes. Both BLAST and FASTA are implemented as search and alignment tools in the Workbench, along with several local and global alignment tools, tools for DNA sequence translation, protein sequence feature analysis, multiple sequence alignment, and phylogenetic tree drawing. Although its interface can be somewhat complicated, involving a lot of window scrolling and button clicking, the Biology Workbench is comprehensive, convenient, and accessible web-based toolkit. One of its main benefits is that many sequence file formats are accepted and can move easily from keyword-based database search, to sequence-based search, to multiple alignment, to phylogenetic analysis.

EMBOSS

EMBOSS is "The European Molecular Biology Open Software Suite". EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology user community. The software automatically copes with data in a variety of formats and even allows transparent

retrieval of sequence data from the web. Within EMBOSS you will find numerous applications covering areas such as:

- Sequence alignment,
- Rapid database searching with sequence patterns,
- Protein motif identification, including domain analysis,
- Nucleotide sequence pattern analysis---for example to identify CpG islands or repeats,
- Codon usage analysis for small genomes,
- Rapid identification of sequence patterns in large scale sequence sets,
- Presentation tools for publication, and much more.

References

1. Baxevanis A.D., Ouellette B. F. F. (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition, John Wiley & Son, New York
2. Elloumi M., Zomaya A. Y. (2011) Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley a& Son, New York
3. Liu L., Agren R., Bordel S., Nielsen J. (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Letters 584: 2556–2564.
4. Milne C.B., Kim P.J., Eddy J.A., Price N.D. (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnol J. 4(12):1653-70
5. Pevzner P., Shamir R. (2011) Bioinformatics for Biologists, 1st Edition, Cambrage University Press
6. Ramsden J. (2015) Bioinformatics: An Introduction, Springer-Verlag, London
7. Singh G. B. (2015) Fundamentals of Bioinformatics and Computational Biology, Springer International Publishing, Switzerland

Omics and system biology

Basic level

Ventsislava Petrova
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Kliment Petrov
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Contents

Tools for Genomics and Proteomics

The sequence alignment methods can be used to analyze a single sequence or structure and compare multiple sequences of single-gene length. These methods can help in understanding the function of a particular gene or the mechanism of a particular protein. However, it is also interesting to understand how gene functions manifest in the observable characteristics of an organism: its *phenotype*. In this respect, some datatypes and tools are available that allow studying the integrated function of all the genes in a genome.

Experimental strategies for analysing one gene or one protein are progressively replaced by parallel approaches in which many genes are examined simultaneously. Using bioinformatics algorithms information from multiple sources can be integrated to form a complete picture of genomic function and its expression, as well as to allow comparison between the genomes of different organisms. Figure 1 shows how genome information is transformed in phenotypic expression.

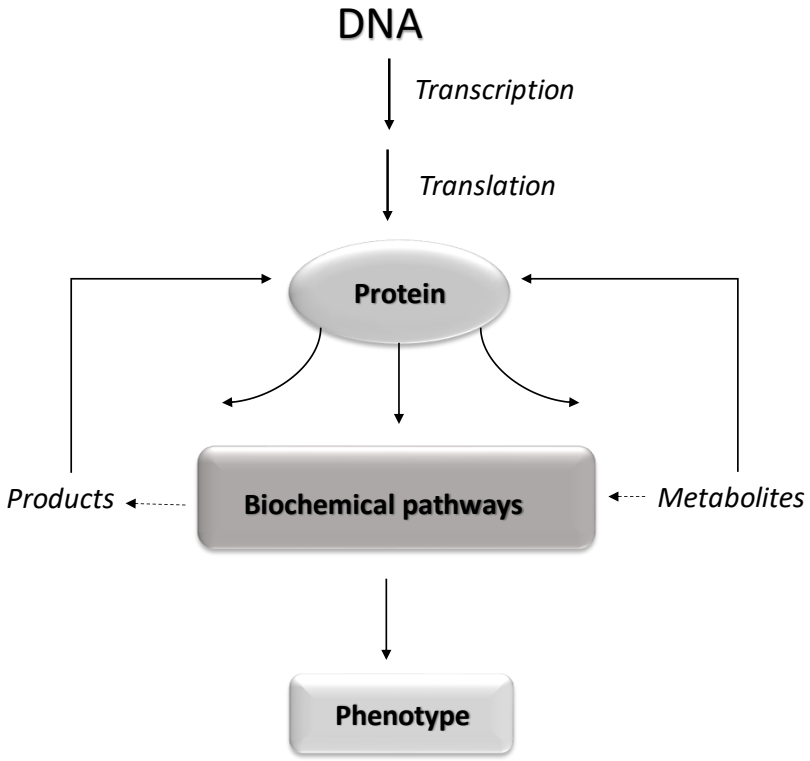


Figure 1. Transferring genome information to phenotype

For decades biologists have been collecting information from the molecular to the cellular level and beyond to see the functions of the genome as a whole. The process of automating and scaling up biochemical experimentation, and treating biochemical data as a public resource, is significantly facilitated by the use of bioinformatics.

The Human Genome Project has not only made gigabytes of biological sequence information available but it has begun to change the entire landscape of biological research by its example. Protein structure determination has not yet been automated at the same level as sequence determination, but several projects in structural genomics are launched, with the main goal to create a high-speed structure

determination approaches. The concept behind the DNA microarray experiment allows performance of comprehensive biochemical and molecular biology experiments.

One of the major tasks of bioinformatics is creating software systems for information management that can effectively annotate each part of a genome sequence with information about everything from its function, to the structure of its protein product (if it has one), to the rate at which the gene is expressed at different life stages of an organism. Another task of genome information management systems is to allow users to make intuitive, visual comparisons between large data sets. Many new data integration projects, from visual comparison of multiple genomes to visual integration of expression data with genome map data, are developed.

Sequencing Genes and Genomes

One of the first computational challenges in the process of sequencing a gene (or a genome) is the interpretation of the pattern of fragments on a sequencing gel.

Analysis of Raw Sequence Data: Basecalling

The process of assigning a sequence to raw data from DNA sequencing is called *basecalling*. If this step doesn't produce a correct DNA sequence, any subsequent analysis of the sequence is affected. All sequences deposited in public databases are affected by basecalling errors due to uncertainties in sequencer output or to equipment malfunctions. EST and genome survey sequences have the highest error rates (1/10 -1/100 errors per base), followed by finished sequences from small laboratories (1/100 - 1/1,000 per base) and finished sequences from large genome sequencing centers (1/10,000 -1/100,000 per base). Any sequence in GenBank is likely to have at least one error. Improving sequencing technology, and especially the signal detection and processing involved in DNA sequencing, is still the subject of active research.

There are two popular high-throughput methods for DNA sequencing. DNA sequencing relies on the ability to create a ladder of fragments of DNA at single base resolution and separate the DNA fragments by gel electrophoresis. Generally, the fragmented DNA is labeled with four different fluorescent labels, one for each base-specific fragmentation, and run a mixture of the four samples in one gel lane. Another commonly used sequencing method runs each sample in a separate, closely spaced lane. In both cases, the gel is scanned with a laser, which excites each fluorescent band on the gel in sequence. Each of these protocols has its advantages in different types of experiments, so both are in common use.

There are a variety of commercial and noncommercial tools for automated basecalling. Some of them are fully integrated with particular sequencing hardware and input datatypes. Most of them allow, and in fact require, curation by an expert user as sequence is determined.

The raw result of sequencing is a record of fluorescence intensities at each position in a sequencing gel. Figure 2 shows detector output from a modern sequencing experiment. The challenge for automated basecalling software is to translate the fluorescence peaks into four-letter DNA sequence code. As the separation of bands on a sequencing gel isn't perfect, the quality of the separation and the shape of the bands worsens over the length of the gel. Peaks broaden and intermix, and at some point (usually 400 -500 bases) the peaks become impossible to resolve. It is well-understood that systematic errors occurred, so computer algorithms are developed in a way to compensate them. The main goal of the basecalling software is to improve the accuracy of each sequence read, as well as to extend the range of sequencing runs, by providing means to deconvolute the more unclear fluorescence peaks at the end of the run.

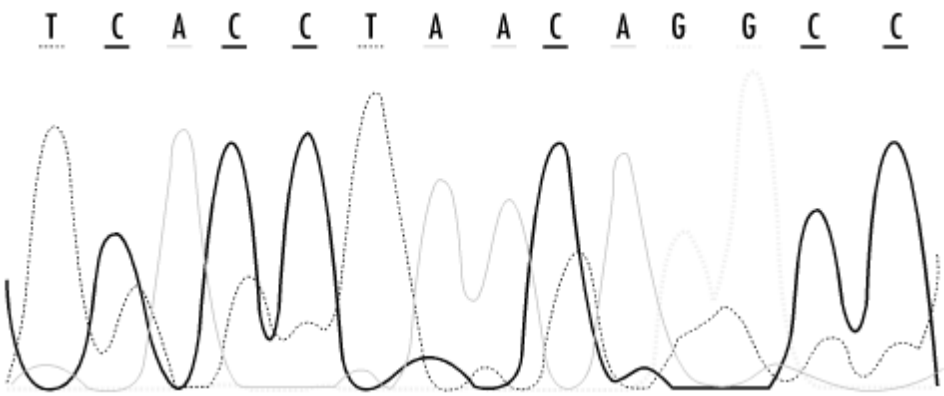


Figure 2. Detector output from a sequencing experiment

Modern sequencing technologies replace gels with microscopic capillary systems, but the core concepts of the process are the same as in gel-based sequencing: fragmentation of the DNA and separation of individual fragments by electrophoresis.

Sequencing an Entire Genome

Genome sequencing isn't simply a scaled -up version of a gene-sequencing run. The sequence length limit of something like 500 base pairs. And the length of a genome can range from tens of thousands to billions of base pairs. So, in order to sequence an entire genome, the genome has to be cleaved into fragments, and then the sequenced fragments need to be reassembled into a continuous sequence.

There are two popular strategies for sequencing genomes: the shotgun approach and the clone contig approach. Combinations of these strategies are often used to sequence larger genomes.

The shotgun approach

Shotgun DNA sequencing is an automated approach for DNA sequencing. Here, DNA is broken into random fragments of manageable length (around 2,000 KB). They are cloned into plasmids (called a *clone library*). If a sufficiently large amount of genomic DNA is fragmented, the set of clones spans every base pair of the genome many times. The end of each cloned DNA fragment is then sequenced, or in some cases, both ends are sequenced. Although only 400 -500 bases at the end(s) of the fragment are sequenced, if enough clones are randomly selected from the library, the amount of sequenced DNA still encompass every base pair of the genome several times. The final step in shotgun sequencing is sequence assembly. Usually, assembly of sequences results in multiple *contigs*—clearly assembled lengths of sequence that don't overlap each other. The final steps in sequencing a complete genome by shotgun sequencing are either to find clones that can fill in the missing regions, or to use PCR or other techniques to amplify DNA sequence from the gaps.

The clone contig approach

The *clone contig* approach relies on shotgun sequencing as well, but on a smaller scale. Instead of starting by breaking down the entire genome into random fragments, the clone contig approach starts by breaking it down into restriction fragments, which can then be cloned into artificial chromosome vectors

and amplified. Each of the cloned restriction fragments can be sequenced and assembled by a standard shotgun approach. When the genome is cleaved into restriction fragments, it is only partially degraded. The amount of restriction enzyme applied to the DNA sample is sufficient to cut at only approximately 50% of the available restriction sites in the sample. This means that some fragments will span a particular restriction site, while other fragments will be cut at that particular site and will span other restriction sites. So, the clone library that is made up of these restriction fragments will contain overlapping fragments. The process of assembly starts with so called chromosome walking. Finding a specific clone, then finding the next clone that overlaps it, and then the next, etc. Usually, a probe hybridization technique or PCR are used to help identify the restriction fragment that has been inserted into each clone.

Genomes can be mapped at various levels of detail. Genetic linkage maps could be created which assign the genes that give rise to particular traits to specific loci on the chromosome. Thus, they provide a set of ordered markers, sometimes very detailed depending on the organism, which can help researchers understand genome function (and provide a framework for assembling a full genome map). Also, physical maps can be built in several ways: by digesting the DNA with restriction enzymes that cut at particular sites, by developing ordered clone libraries, and by fluorescence microscopy of single, restriction enzyme-cleaved DNA molecules fixed to a glass substrate. The key to each method is that, using a combination of labeled probes and known genetic markers (in restriction mapping) or by identifying overlapping regions (in library creation), the fragments of a genome can be ordered correctly into a highly specific map.

LIMS: Tracking mini sequences

Tracking the millions of unique DNA samples that may be isolated from the genome is one of the biggest information technology challenges. The systems that manage output from high-throughput sequencing are called Laboratory Information Management Systems (LIMS), and its development and maintenance make up the biggest share of bioinformatics work in industrial settings. Other high throughput technologies, such as microarrays and cheminformatics, also require complicated LIMS support.

Sequence Assembly

Basecalling is only the first step in putting together a complete genome sequence (Fig. 3). Once the short fragments of sequence are obtained, they must be assembled into a complete sequence that may be many thousands of times their length. The next step is sequence assembly.

DNA sequencing using a shotgun approach provides thousands or millions of mini sequences, each 400-500 fragments in length. The fragments are random and can partially or completely overlap each other. Because of these overlaps, every fragment in the set can be identified by sequence identity as adjacent to some number of other fragments. Each of those fragments overlaps yet another set of fragments, and so on. Finally, all the fragments need to be optimally join together into one continuous sequence. However, the repetitive sequences can complicate the assembly process. Some fragments will be uncloneable, and the sequencing process will fail, leaving gaps in the DNA sequence that complicate automated assembly. If there isn't sufficient information at some point in the sequence for assembly to continue, the sequence contig that is being created comes to an end, and a new contig starts.

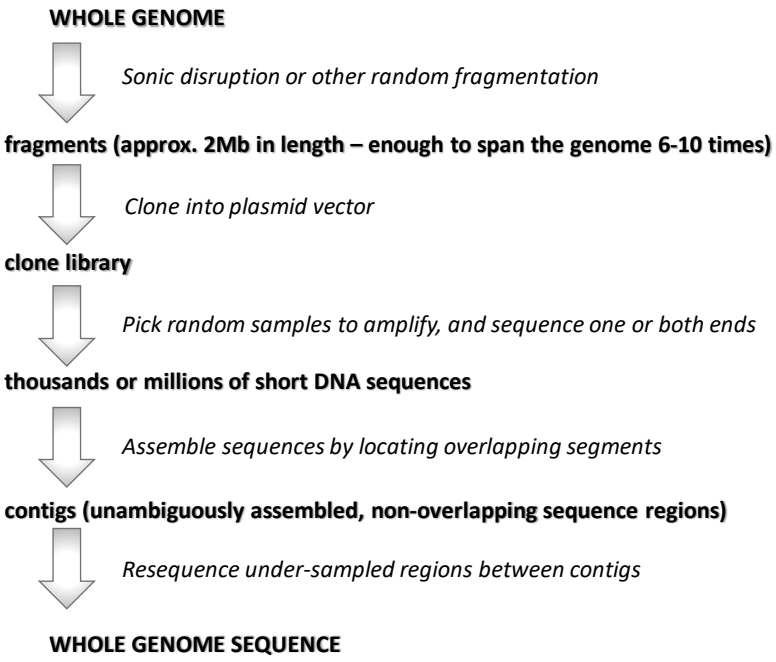


Figure 3. The shotgun DNA sequencing approach

Accessing Genome Information on the Web

Partial or complete DNA sequences from hundreds of genomes are available in GenBank. Putting those sequence records together into an intelligible representation of genome structure isn't so easy. There are several efforts underway to integrate DNA sequence with higher-level maps of genomes in a user-friendly format. So far, these efforts are focused on the human genome and genomes of important plant and animal model systems.

NCBI Genome Resources

NCBI offers access to a wide selection of web-based genome analysis tools from the Genomic Biology section of its main web site. Their interfaces are user-friendly, and NCBI supplies plenty of documentation explaining how to use the provided tools and databases.

Some of the available genomic tools are:

Genome Information

Genome project information is available from the *Entrez Genomes* page at NCBI. Database listings are available for the full database or for related groups of organisms such as microorganisms, archaea, bacteria, eukaryotes, and viruses. Each entry in the database is linked to a taxonomy browser entry or a home page with further links to available information about the organism. If a genome map of the organism

is available, a "See the Genome" link shows up on the organism's home page. From the home page, you can also download genome sequences and references.

Map Viewer

Depending on the genome, you can access links to overview maps showing known protein-coding regions, listings of coding regions for protein and RNA, and other information. Map Viewer distinguishes between four levels of information: the organism's home page, the graphical view of the genome, the detailed map for each chromosome (aligned to a master map from which the user can select where to zoom in), and the sequence view, which graphically displays annotations for regions of the genome sequence.

ORF Finder

The Open Reading Frame (ORF) Finder is a tool for locating open reading frames in a DNA sequence. ORF finders translate the sequence using standard or user-specified genetic code. In noncoding DNA, stop codons are frequently found. Information from the ORF finder can provide hints about the precise reading frame for a DNA sequence and about where coding regions start and stop. For many genomes found in the Entrez Genomes database, ORF Finder is available as an integrated tool from the map view of the genome.

HomoloGene

HomoloGene is an automated system for constructing putative homology groups from the complete gene sets of a wide range of eukaryotic species. The ortholog pairs are identified either by curation of literature reports or calculation of similarity. The HomoloGene database can be searched using gene symbols, gene names, GenBank accession numbers, and other features.

Clusters of Orthologous Groups (COG)

COG is a database of orthologous protein groups. The database was developed by comparing protein sequences across 97 genomes. The entries in COG represent genome functions that are conserved throughout much of evolutionary history. The COG database can be searched by functional category, phylogenetic pattern, and a number of other properties.

Genome Annotation

Genome annotation in practice is hyperlinking of content between multiple databases—sequence, structure, and functional genomics fully linked together in a queryable system. It is a difficult process because there are a huge number of different pieces of information attached to every gene in a genome and it generally relies on relational databases to integrate genome sequence information with other data.

Genome Comparison

Pairwise or multiple comparison of genomes is the tool that can be used in many different studies, such as answering of basic questions of evolutionary biology (genetic polymorphisms) or very specific clinical questions (variations in phenotype).

Comparing of whole genomes, rather than just comparing genes one at a time, can help in defining the regions of similarity within uncharacterized or even supposedly redundant DNA. Genome comparison will also aid in genomic annotation. Prototype genome comparisons allows justifying the sequencing of additional genomes and it is useful both at the map level and directly at the sequence level.

PipMaker

PipMaker is a tool that computes alignments of similar regions in two DNA sequences. This is useful in identifying large-scale patterns of similarity in longer sequences. The process of using PipMaker is relatively simple. Starting with two FASTA-format sequence files, you first generate a set of instructions for masking sequence repeats (using the RepeatMasker server). This reduces the number of uninformative hits in the sequence comparison. The resulting information, plus a simple file containing a numerical list of known gene positions, is submitted to the PipMaker web server at Penn State University and the results are emailed to you.

MUMmer

Another program for ultra-fast alignment of large-scale DNA and protein sequences is MUMmer. Its first application was a detailed comparison of genomes of two strains of *M. tuberculosis*. MUMmer can compare sequences millions of base pairs in length and produce colorful visualizations of regions of similarity. MUMmer is based on a computer algorithm called a *suffix tree*, which essentially makes it easy for the system to rapidly handle a large number of pairwise comparisons. MUMmer can also align incomplete genomes; it can easily handle the 100s or 1000s of contigs from a shotgun sequencing project and will align them to another set of contigs or a genome using the NUCmer program included with the system. If the species are too divergent for a DNA sequence alignment to detect similarity, then the PROmer program can generate alignments based upon the six-frame translations of both input sequences.

Functional Genomics

Launching of high-speed sequencing methods has changed the way we study the DNA sequences that code for proteins. It is now becoming possible to view the whole DNA sequence of a chromosome as a single entity and to examine how the parts of it work together to produce the complexity of the organism as a whole.

The functions of the genome break down loosely into a few obvious categories: metabolism, regulation, signaling, and construction. Metabolic pathways convert chemical energy derived from environmental sources into useful work in the cell. Regulatory pathways are biochemical mechanisms that control what genomic DNA does: when it is expressed or not. Genomic regulation involves not only expressed genes but structural and sequence signals in the DNA where regulatory proteins may bind. Signaling pathways control the fluxes of chemicals from one compartment in a cell to another. Many regulatory systems for the control of DNA transcription have been studied. Mapping these metabolic, regulatory, and signaling systems to the genome sequence is the goal of the field of functional genomics.

Sequence-Based Approaches for Analyzing Gene Expression

In addition to genome sequence, GenBank contains many other kinds of DNA sequence. Expressed sequence tag (EST) data for an organism can be an extremely useful starting point for analysis of gene expression. ESTs are partial sequences of cDNA clones of cellular mRNA. mRNA levels respond to changes in the cell or its environment; mRNA levels are tissue dependent, and they change during the life cycle of the organism as well. Quantitation of mRNA or cDNA provides a good measure of what a genome is doing under particular conditions.

NCBI offers a database called dbEST that provides access to several thousand libraries of ESTs. Quite a large number of these are human EST libraries, but there are libraries from dozens of other organisms as well.

DNA Microarrays

Recently, new technology has made it possible for researchers to rapidly explore expression patterns of entire genomes. A *microarray* (or gene chip) is a small glass which surface is covered with 20,000 or more precisely placed spots each containing a different DNA oligomer. cDNA can also be affixed to the slide to function as probes. Other media, such as thin membranes, can be used in place of slides. The key to the experiment is that each piece of DNA is immobilized and any reaction that results in a change in microarray signal can be precisely assigned to a specific DNA sequence.

Microarrays are conceptually no different from traditional hybridization experiments such as Southern Blots or Northern Blots. In traditional blotting, the protein sample is immobilized; in microarray experiments, the probe is immobilized, and the amount of information that can be collected in one experiment is vastly larger. Figure 4 shows just a portion of a microarray scan.

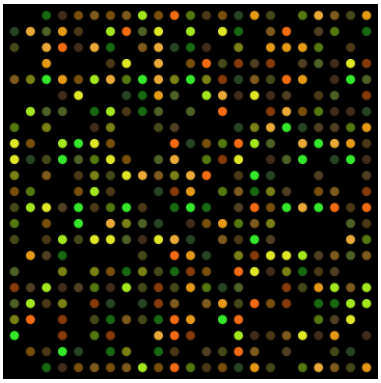


Figure 4. A microarray scan

Microarray technology is now routinely used for DNA sequencing experiments; for instance, in testing for the presence of polymorphisms. Another development is the use of microarrays for gene expression analysis. When a gene is expressed, an mRNA transcript is produced. If DNA oligomers complementary to the genes of interest are placed on the microarray, mRNA or cDNA can be hybridized to the chip, providing a rapid assay as to whether or not those genes are being expressed. Experiments like these for example have been performed in yeast to test differences in whole-genome expression patterns in response to changes in ambient sugar concentration. Microarray experiments can provide information about the behavior of every one of an organism's genes in response to environmental changes.

Bioinformatics Challenges in Microarray Design and Analysis

Bioinformatics plays multiple roles in microarray experiments. In fact, it is difficult to consider of microarrays as useful without the involvement of computers and databases. From the design of chips for specific purposes, to the quantitation of signals, to the extraction of groups of genes with linked expression

profiles, microarray analysis is a process that is difficult, if not impossible, to do without the use of specific bioinformatics software.

In the public domain, several projects for linking expression data with associated sequences and annotations are ongoing. The biggest microarray database is the EMBL-EBI's ArrayExpress. The National Human Genome Research Institute (NHGRI) is currently offering a demonstration version of an array data management system that includes both analysis tools and relational database support.

Planning array experiments

A key element in microarray experiments is chip design. Chip design is a process that can take months. In order for microarray results to be clear and unambiguous, each DNA probe in the array must be sufficiently unique that only one specific target gene can hybridize with it. Otherwise, the amount of signal detected at each spot will be quantitatively incorrect.

Analyzing scanned microarray images

Once the array experiment is complete, you'll find yourself in possession of a lot of very large TIFF files containing scanned images of your arrays. The standard for public-domain microarray analysis tools are the packages developed at Stanford. One package, ScanAlyze, is the image analysis tool, well regarded and widely used. It supports TIFF files as well as the Stanford SCN format.

Numerous others softwares exist for microarray data analysis, such as:

GenomeStudio Software enables you to visualize and analyze microarray data generated on Illumina platforms. The software package is composed of discrete application modules that enable you to obtain a comprehensive view of the genome, gene expression, and gene regulation.

TM4 Microarray Software Suite is an open-source tools for microarray data management and reporting, image analysis, normalization and pipeline control, and data mining and visualization.

MAIA is a software package for automatic processing of the one- and two-color images produced in cDNA, CGH or protein microarray technologies.

AIM (Automatic Image Processing system for Microarray) provides a method for uncalibrated microarray gridding and quantitative image analysis. AIM is a fast suffix array construction algorithm that performs very well even for worst-case strings. This system operates independently as well as command-line tools.

Koadarray, a fully automatic array image analysis software which can process single or multiple array images entirely unattended.

Clustering expression profiles

The most popular strategy for analysis of microarray data is the clustering of expression profiles. An *expression profile* can be visualized as a plot that describes the change in expression at one spot on a microarray grid over the course of the experiment. The course of the experiment changes with the context, anything from changes in the concentration of nutrients in the medium in which cells are being grown prior to having their DNA hybridized to the array, to cell cycle stages.

Different clustering methods, such as hierarchical clustering or SOMs (self-organizing maps) may work better in different situations, but the general aim of each of these methods is the same. If two genes change expression levels in the same way in response to a change in environment, it can be assumed that those genes are related. They may share something as simple as a promoter, or more likely, they are

controlled by the same complex regulatory pathway. Automated clustering of expression profiles looks for similar features but doesn't necessarily point to causes for those changes.

Proteomics

Proteomics refers to techniques that simultaneously study the entire protein complement of a cell. While protein purification and separation methods are constantly improving, and the time-to completion of protein structures determined by NMR and x-ray crystallography is decreasing, there is as yet no single way to rapidly crystallize the entire protein complement of an organism and determine every structure. The technological advance in biochemistry that most requires informatics support is the immobilized-gradient 2D-PAGE process and the subsequent characterization of separated protein products by mass spectrometry.

Experimental Approaches in Proteomics

Knowing when and at what levels genes are being expressed is only the first step in understanding how the genome determines phenotype. While mRNA levels are correlated with protein concentration in the cell, proteins are subject to post-translational modifications that can't be detected with a hybridization experiment. Experimental tools for determining protein concentration and activity in the cell are the crucial next step in the process.

Another high-throughput technology that is emerging as a tool in functional genomics is 2D gel electrophoresis. Two-dimensional gel electrophoresis can be used to separate protein mixtures containing thousands of components. The first dimension of the experiment is separation of the components of a solution along a pH gradient (isoelectric focusing). The second dimension is separation of the components orthogonally by molecular weight. Separation in these two dimensions can resolve even a complicated mixture of components. Figure 5 shows an example of 2D-PAGE map from *E. coli*. The 2D-PAGE experiment separates proteins from a mixed sample so that individual proteins can be identified. Each spot on the map represents a different protein.

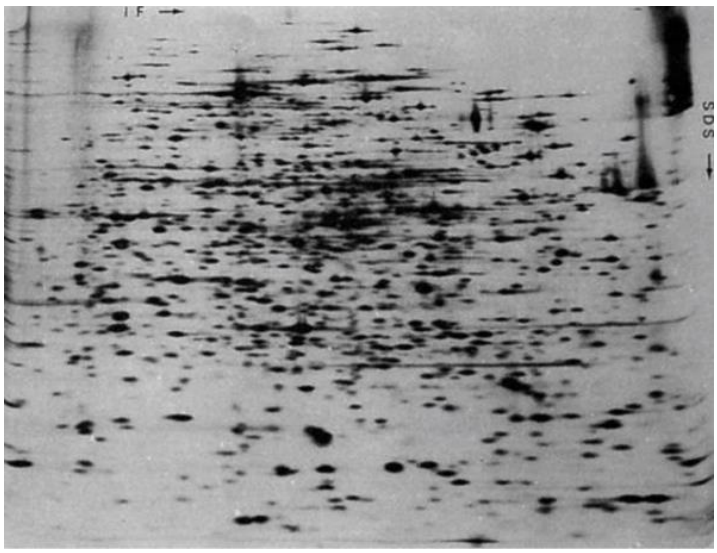


Figure 5. A 2D-PAGE map from *E. coli*

Using 2D gel electrophoresis allows very precise protein separations, resulting in standardized high-density data arrays. They can therefore be subjected to automated image analysis and quantitation and used for accurate comparative studies. The other advance that has put 2D gel technology at the forefront of modern molecular biology methods is the capacity to chemically analyze each spot on the gel using mass spectrometry. This allows the measurable biochemical phenomenon—the amount of protein found in a particular spot on the gel—to be directly connected to the sequence of the protein found at that spot.

Informatics Challenges in 2D-PAGE Analysis

The analysis pathway for 2D-PAGE gel images is essentially quite similar to that for microarrays. The first step is an image analysis, in which the positions of spots on the gel are identified and the boundaries between different spots are resolved. Molecular weight and isoelectric point (PI) for each protein in the gel can be estimated according to position.

Next, the spots are identified, and sequence information is used to make the connection between a particular spot and its gene sequence. In proteome analysis, the immobilized proteins can either be sequenced in situ or spots of protein can be physically removed from the gel, eluted, and analyzed using mass spectrometry methods such as electrospray ionization mass spectrometry (ESI-MS) or matrix-assisted laser desorption ionization mass spectrometry (MALDI).

Tools for Proteomics Analysis

Several public-domain programs for proteomics analysis are available on the Web. Most of these can be accessed through the excellent proteomics resource at Expert Protein Analysis System (ExPASy). ExPASy is the Swiss Institute of Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., resources) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.

Biochemical Pathway Databases

Gene and protein expression are only two steps in the translation of genetic code to phenotype. Once genes are expressed and translated into proteins, their products participate in complicated biochemical interactions called *pathways*, as shown in Figure 6. Each pathway may supply chemical precursors to many other pathways, meaning that each protein has relationships not only to the preceding and following biochemical steps in a single pathway, but possibly to steps in several pathways. The complicated branching of metabolic pathways are far more difficult to represent and search than the linear sequences of genes and genomes.

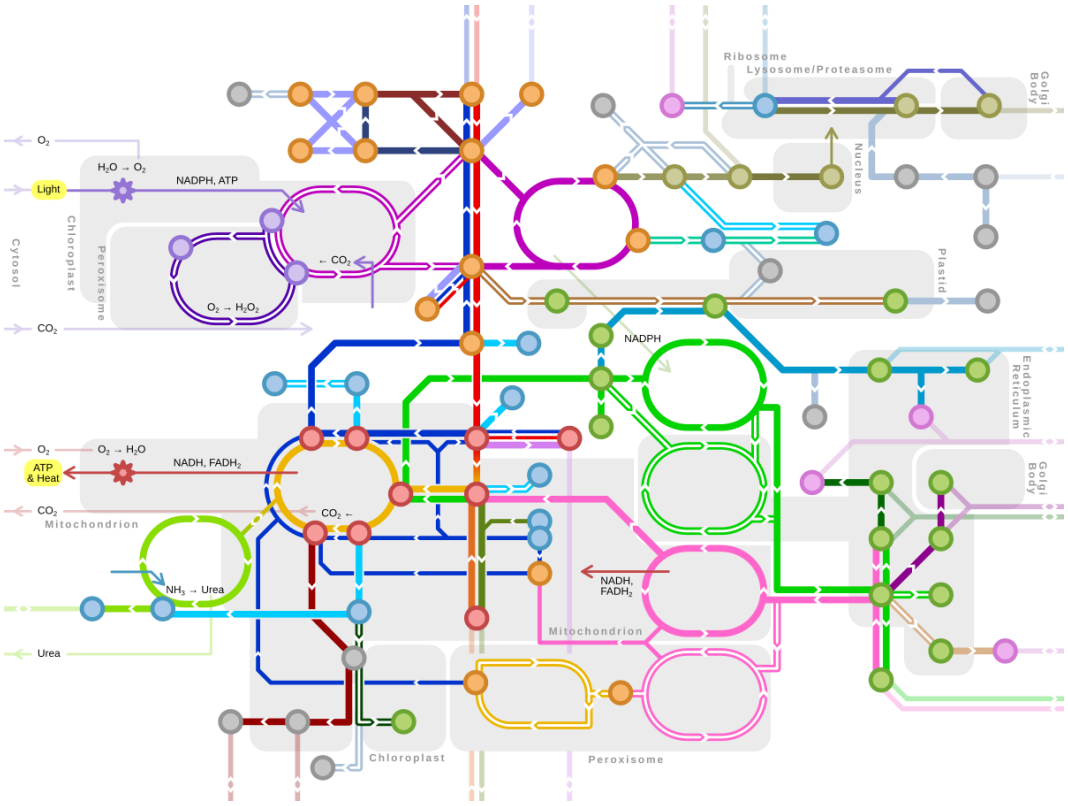


Figure 6. A complex metabolic pathway

Several web-based services offer access to metabolic pathway information.

KEGG

The best known metabolic pathway resources on the Web is the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG provides its metabolic overviews as map illustrations, rather than text-only, and can be easier to use for the visually-oriented user. KEGG also provides listings of EC numbers and their corresponding enzymes broken down by level, and many helpful links to sites describing enzyme and ligand nomenclature in detail. The LIGAND database, associated with KEGG, is a useful resource for identifying small molecules involved in biochemical pathways. KEGG is searchable by sequence homology, keyword, and chemical entity; you can also input the LIGAND ID codes of two small molecules and find all of the possible metabolic pathways connecting them.

PathDB

PathDB is another type of metabolic pathway database. While it contains roughly the same information as KEGG—identities of compounds and metabolic proteins, and information about the steps that connect these entities—it handles information in a far more flexible way than the other metabolic databases. Instead of limiting searches to arbitrary metabolic pathways and describing pathways with preconceived images, PathDB allows you to find any set of connected reactions that link point A to point B, or compound A to compound B. PathDB contains, in addition to the usual search tools, a pathway

visualization interface that allows you to review any selected pathway and display different representations of the pathway.

References

1. Baxevanis A.D., Ouellette B. F. F. (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition, John Wiley & Son, New York
2. Elloumi M., Zomaya A. Y. (2011) Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley a& Son, New York
3. Liu L., Agren R., Bordel S., Nielsen J. (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Letters 584: 2556–2564.
4. Milne C.B., Kim P.J., Eddy J.A., Price N.D. (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnol J. 4(12):1653-70
5. Pevzner P., Shamir R. (2011) Bioinformatics for Biologists, 1st Edition, Cambrage University Press
6. Ramsden J. (2015) Bioinformatics: An Introduction, Springer-Verlag, London
7. Singh G. B. (2015) Fundamentals of Bioinformatics and Computational Biology, Springer International Publishing, Switzerland

Biology, biological databases, and high-throughput data sources

Advanced level

Ventsislava Petrova
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Kliment Petrov
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Zlatyo Uzunov
JST Corporation Ltd.
Sofia, Bulgaria
https://jst.bg/

Contents

The Internet has completely changed the way scientists search for and exchange information. Data that once had to be communicated on paper is now digitized and distributed from centralized databases. Articles in journals are available online. And nearly every research group has a web page offering everything from reprints to software downloads to data to automated data-processing services.

Search Engines and Boolean Searching

AltaVista, Mozilla, Google, Internet explorer, Safari, and dozens of other search engines exist to help you find the billion or more pages that respond to your search. However, often scientists are looking for perhaps a couple of needles in a large haystack. Knowing how to structure a query to limit the majority of the junk that will come up in a search is very useful, both in web searching and in

keyword-based database searching. Understanding how to formulate boolean queries that limit your search space is a critical research skill.

Most web surfers approach searching randomly at best. But each search engine makes different default assumptions, so if you enter *protein structure* into Excite's query field, you are asking for an entirely different search than if you enter *protein structure* into Google's query field. In order to search effectively, you need to use boolean logic, which is an extremely simple way of stating how a group of things should be divided or combined into sets.

Search engines and public biological databases use some form of boolean logic. Boolean queries restrict the results that are returned from a database by joining a series of search terms with the operators AND, OR, and NOT. For example: joining two key terms with AND finds documents that contain only *key term1* and *key term2* ; using OR returns documents that contain either *key term1* or *key term2* (or both); and using NOT discovers documents that contain *key term1* but not *key term2*.

However, search engines differ in how they interpret a space. Some of them consider a space as OR, so when *protein structure* is typed, the search engine looks for protein or structure. As a result, a lot of advertisements for fad diets and protein supplements come up before to get to the scientific sites of interest. On the other hand, in Google space refers to AND, so the only references to be found are those that contain protein and structure.

Boolean queries are read from left to right, just like text. Parentheses can structure more complex boolean queries. For instance, if you look for documents that contain *key term1* and one of either *key term2* or *key term3*, but not *key term4*, your query would look like this: (*key term1* AND (*key term2* OR *key term3*)) NOT *key term4*.

Many search engines allow to use quotation marks to specify a phrase. In order to find only documents in which the key term *enzyme activity* appear together in sequence, searching for "enzyme activity" is one way to narrow the results.

There are many excellent web tutorials available on boolean searching. Try a search with the phrase *boolean searching* in Google, and see what comes up.

Finding Scientific Articles

An excellent resource for searching the scientific literature in the biological sciences is the free server sponsored by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. This server makes it possible for anyone with a web browser to search the Medline database. There are other literature databases of comparable quality available, but most of these are not free. Outside of refereed resources, however, anyone can publish information on the Web. Often research groups make papers available as technical reports on their web sites. These technical reports may never be peer reviewed or published outside the research group's home organization, and your only evidence to their quality is the reputation and expertise of the authors. This isn't to say that you shouldn't trust or seek out these sources. Many government organizations and academic research groups have reference material of near-textbook quality on their web sites. For example, the University of Washington Genome Center has an excellent tutorial on genome sequencing, and NCBI has a good practical tutorial on use of the BLAST sequence alignment program and its variants.

Using PubMed Effectively

PubMed is one of the most valuable web resources available to biologists. Over 4,000 journals are indexed in PubMed, including most of the well-regarded journals in cell and molecular biology, biochemistry, genetics, and related fields, as well as many clinical publications of interest to medical professionals. PubMed uses a keyword-based search strategy and allows the boolean operators AND, OR, and NOT in query statements. Users can specify which database fields to check for each search term by following the search term with a field name enclosed in square brackets. Additionally, users can search PubMed using Medical Subject Heading (MeSH) terms. MeSH is a library of standardized terms that may help locate manuscripts that use alternate terms to refer to the same concept. The MeSH browser allows users to enter a word or word fragment and find related keywords in the MeSH library. PubMed automatically finds MeSH terms related to query terms and uses them to enhance queries.

For example, we searched for "protein structure" in PubMed. The terms protein and structure are automatically joined with an AND unless otherwise specified. The resulting boolean query statement submitted to PubMed is actually:

("proteins"[MeSH Terms] OR "proteins"[All Fields] OR "protein"[All Fields]) AND ("Structure"[Journal] OR "structure"[All Fields])

The results of the search are shown in Figure 1.

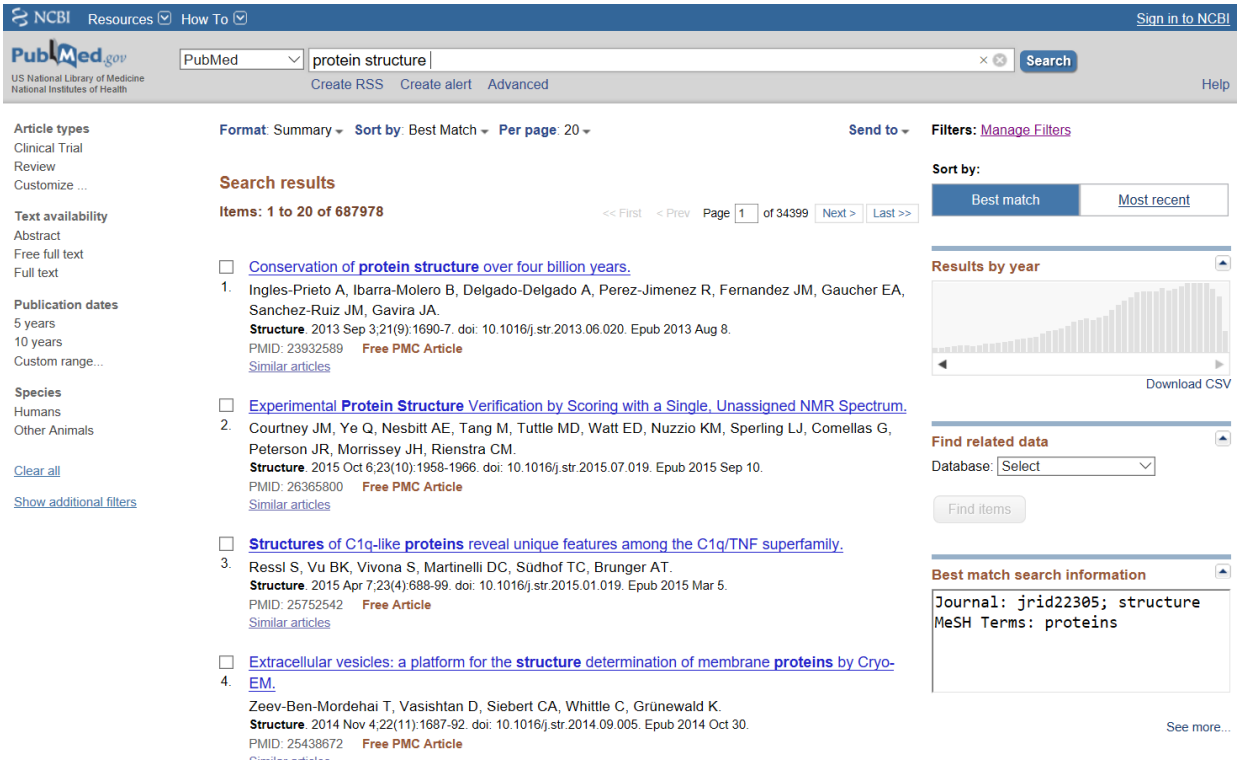


Figure 1. Results from a PubMed search

As you can see in Figure 2, PubMed also allows you to use a web interface to narrow your search.

The Advanced link immediately below the query box on the main PubMed page takes you to this web form.

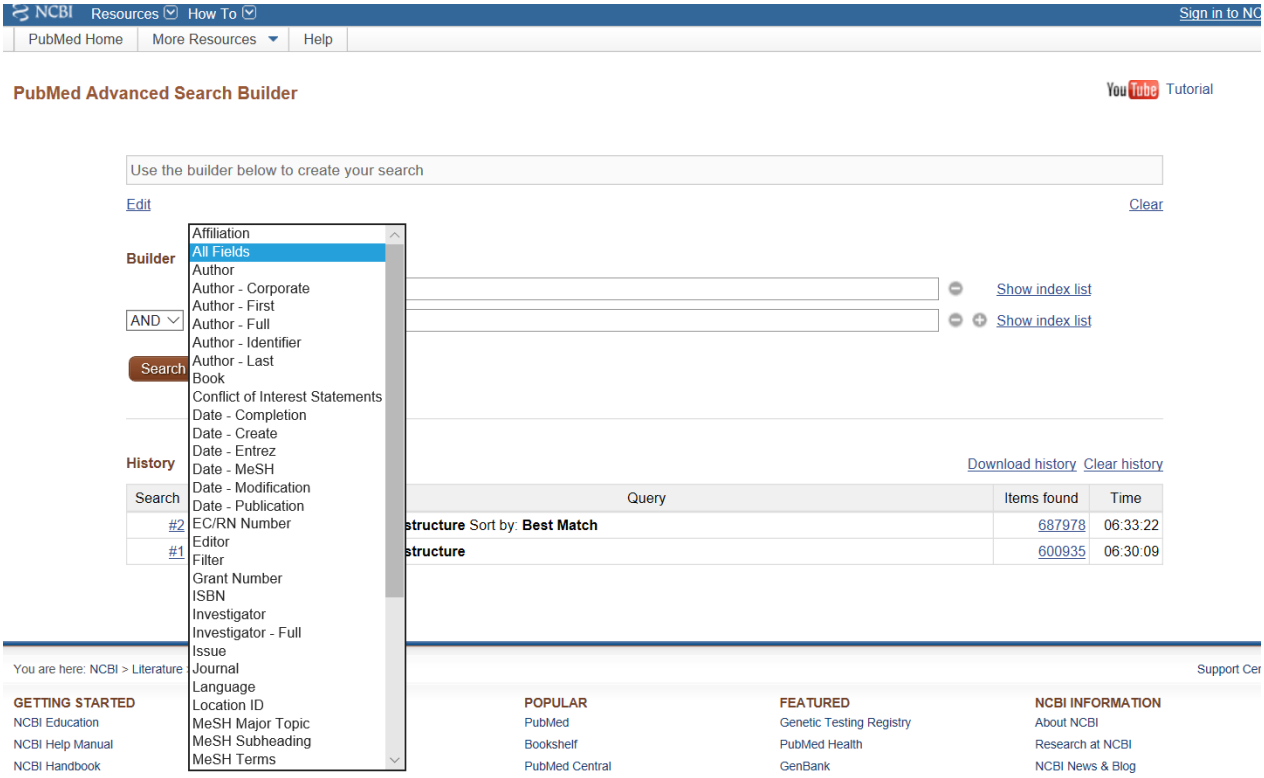


Figure 2. Narrowing a search strategy using the Advanced menu in PubMed

The Advanced form allows you to add specificity to your query. You can limit your search to particular fields in the PubMed database record, such as the Author Name or MeSH Major Topic. Searches can also be limited by language, content (e.g., searching for review articles or clinical trials only), and date.

The Public Biological Databases

The nomenclature problem in biology at the molecular level is immense. Genes are commonly known by unsystematic names. These may come from developmental biology studies in model systems, so that some genes have names like *flightless*, *shaker*, and *antennapedia* due to the developmental effects they cause in a particular animal. Other names are chosen by cellular biologists and represent the function of genes at a cellular level, like *homeobox*. Still other names are chosen by biochemists and structural biologists and refer to a protein that was probably isolated and studied before the gene was ever found.

Though proteins are direct products of genes, they are not always referred to by the same names or codes as the genes that encode them. This kind of confusing nomenclature generally means that only a scientist who works with a particular gene, gene product, or the biochemical process can immediately recognize what the common name of the gene refers to. The biochemistry of a single organism is a

more complex set of information than the taxonomy of living species was at the time of Linnaeus, so it isn't to be expected that a clear and comprehensive system of nomenclature will be arrived at easily. There are many things to be known about a given gene: its source organism, its chromosomal location, and the location of the activator sequences and identities of the proteins that down and up regulated it. Genes also can be categorized by when during the organism's development they are expressed, and in which tissues the expression occurs. They can be characterized by the function of their product, whether it's a structural protein, an enzyme, or a functional RNA. They can be determined by the metabolic pathway that their product is part of, by the substrate they modify or by the product they produce Moreover, they can be categorized by the structural characteristics of their protein products. Figure 3 shows some of the information that could be related with a single gene.

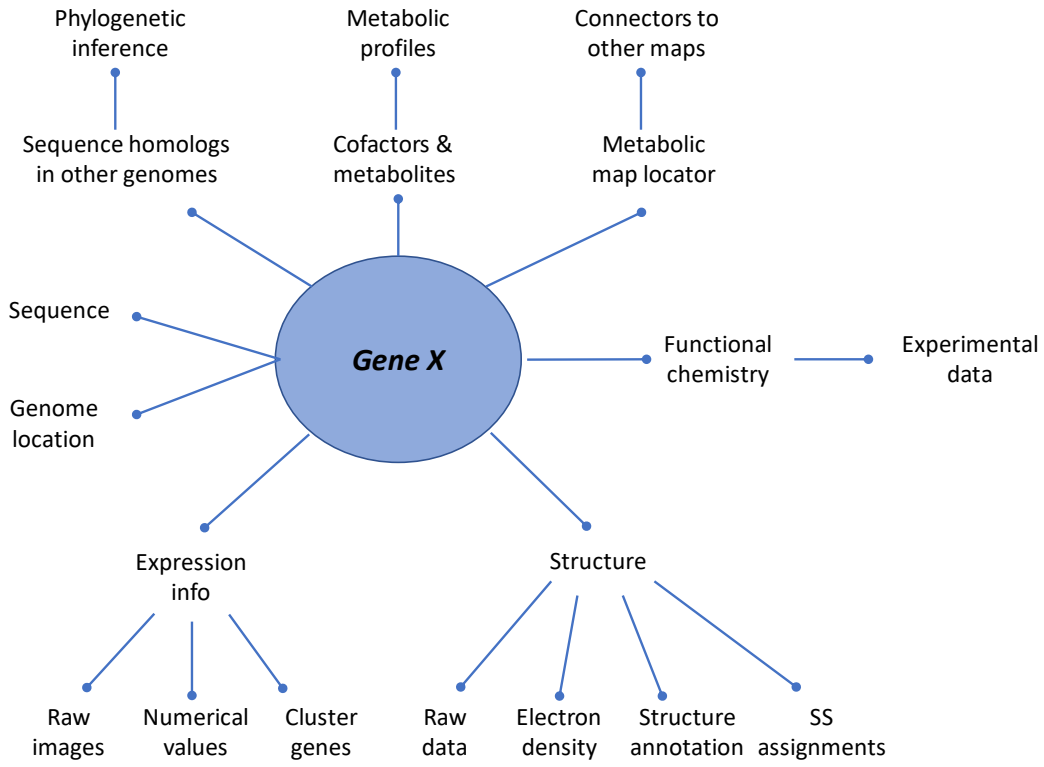


Figure 3. Part of the information associated with a single gene

The problem for maintainers of biological databases becomes mainly one of annotation. Correct annotation of genomic data may be achieved through putting the sufficient information into the database that there is no question of what the gene is, even if it does have a cryptic common name, and creating the proper links between that information and the gene sequence and serial number. Storage of macromolecular data in electronic databases has given rise to a way of working around the problem of nomenclature. The solution has been to give each new entry into the database a serial number and then to store it in a relational database that knows the proper linkages between that serial number, any number of names for the gene or gene product it represents, and all manner of other information about the gene. This strategy is the one currently in use in the major biological databases.

Data Annotation and Data Formats

The representation and distribution of biological data is still an open problem in bioinformatics. The nucleotide sequences of DNA and RNA and the amino acid sequences of proteins reduce neatly to character strings in which a single letter represents a single nucleotide or amino acid. The remaining challenges in representing sequence data are verification of the correctness of the data, thorough annotation of data, and handling of data that comes in ever-larger chunks, such as the sequences of chromosomes and whole genomes.

The standard reduced representation of the 3D structure of biomolecule consists of the Cartesian coordinates of the atoms in the molecule. This aspect of representing the molecule is straightforward. On the other hand, there are a host of complex issues for structure databases that are not completely resolved. Annotation is still an issue for structural data, although the biology community has attempted to form a consensus as to what annotation of a structure is currently required. In the last 15 years, different researchers have developed their own styles and formats for reporting biological data. Biological sequence and structure databases have developed in parallel in the United States and in Europe. The use of proprietary software for data analysis has contributed a number of proprietary data formats to the mix. While there are many specialized databases, we focus here on the fields in which an effort is being made to maintain a comprehensive database of an entire class of data.

3D Molecular Structure Data

Though DNA sequence, protein sequence, and protein structure are in some sense just different ways of representing the same gene product, these datatypes currently are maintained as separate database projects and in unconnected data formats. This is mainly because sequence and structure determination methods have separate histories of development.

The first public molecular biology database, set up about 10 years before the public DNA sequence databases, was the Protein Data Bank (PDB). It represents the central repository for x-ray crystal structures of protein molecules. While the first finish protein structure was presented in the 1950s, there were not a noteworthy number of protein structures accessible until the late 1970s. Computers had not created to the point where graphical representation of protein coordinate structure information was possible, at least at useful speeds. However, in 1971, the PDB was set up at the Brookhaven National Laboratory, to store protein structure information in a computer-based archive. A data format created, which owed a lot of its style to the prerequisites of early computer technology. All through the 1980s, the PDB grew. From 15 sets of entries in 1973, it augments to 69 entries in 1976. The number of coordinate sets deposited each year remained under 100 until 1988, at which time there were still fewer than 400 PDB entries.

In the vicinity of 1988 and 1992, the PDB hit the the turning point in its exponential growth curve. By January 1994, there were 2,143 entries in the PDB; and at the moment the PDB has more than 14,000 entries. Administration of the PDB has been exchanged to a consortium of entry mark, called the Research Collaboratory for Structural Bioinformatics, and and a new format for recording of crystallographic data, the Macromolecular Crystallographic Information File (mmCIF), is being introduced in to replace the antiquated PDB format. Journals that publish crystallographic results require submission to the PDB as a condition of publication, which means that nearly all protein structure data obtained by academic researchers becomes available in the PDB.

A typical issue for information driven investigations of protein structure is the excess and absence of thoroughness of the PDB. There are numerous proteins for which various crystal structures have been submitted to the database. Choosing subsets of the PDB information with which to work is in this manner a critical step in any statistical investigation of protein structure. Numerous statistical

studies of protein structure depend on sets of protein chains that have close to 25% of their sequence in common; if this paradigm is utilized, there are still just around 1,000 unique protein folds represented in the PDB. As the amount of biological sequence data available has grown, the PDB now falls a long ways behind the gene-sequence databases.

DNA, RNA, and Protein Sequence Data

Sequence databases generally specialize in one type of sequence data: DNA, RNA, or protein. There are major sequence data collections and deposition sites in Europe, Japan, and the United States, and there are independent groups that mirror all the data collected in the major public databases, often offering some software that adds value to the data.

In 1970, Ray Wu sequenced the first segment of DNA; twelve bases that occurred as a single strand at the end of a circular DNA that was opened utilizing a cleaving enzyme. In any case, DNA sequencing demonstrated considerably more troublesome than protein sequencing, on the grounds that there is no chemical process that selectively cleaves the first nucleotide from a nucleic acid chain. At the point when Robert Holley announced the sequencing of a 76-nucleotide RNA molecule from yeas, it was following seven years of work. After Holley's sequence was published, different groups refined the protocols for sequencing, even succeeding in sequence effectively a 3,200-base bacteriophage genome. Genuine advance with DNA sequencing came after 1975, with the chemical cleavage method created by Allan Maxam and Walter Gilbert, and with Frederick Sanger's chain terminator procedure.

The first DNA sequence database, established in 1979, was the Gene Sequence Database (GSDB) at Los Alamos National Lab. While GSDB has since been supplanted by the worldwide collaboration that is the modern GenBank, up-to-date gene sequence information is still available from GSDB through the National Center for Genome Resources.

The European Molecular Biology Laboratory, the DNA Database of Japan, and the National Institutes of Health cooperate to make all freely accessible sequence data through GenBank. NCBI has built up a standard relational database format for sequence information presentation and storage, known as the ASN.1 format. While this format guarantees to locate the right sequences of the right kind in GenBank simpler, there are also various services tions giving access to nonredundant versions of the database. The DNA sequence database developed gradually through its first decade. In 1992, GenBank contained just 78,000 DNA sequences — a little more than 100 million pairs of DNA. In 1995, the Human Genome Project, and advances in sequencing innovation, kicked GenBank's growth into high gear. GenBank currently doubles in size every 6 to 8 months, and its rate of increase is constantly growing.

Genomic Data

In addition to the Human Genome Project, there are now separate genome project databases for a large number of model organisms. The sequence content of the genome project databases is represented in GenBank, but the genome project sites also provide everything from genome maps to supplementary resources for researchers working on that organism. As of October 2000, NCBI's Entrez Genome database contained the partial or complete genomes of over 900 species. Many of these are viruses. The remainder include bacteria; archaea; yeast; commonly studied plant model systems such as A. thaliana, rice, and maize; animal model systems such as C. elegans, fruit flies, mice, rats, and puffer fish; as well as organelle genomes. NCBI's web-based software tools for accessing these databases are constantly evolving and becoming more sophisticated.

Biochemical Pathway Data

The most vital biological activities don't occur by the action of single molecule, however as the orchestrated activities of multiple molecules. Since the mid twentieth century, biochemists have analyzed these functional ensembles of enzymes and their substrates. A couple of research groups have started work at intelligently arranging and storing these pathways in databases. Key example of pathway database is KEGG. The Kyoto Encyclopedia of Genes and Genomes (KEGG) stores comparative information about sequence, structure, and genetic linkage databases. This database is queryable through web interfaces and are curated by a combination of automation and human expertise. In addition to these whole genome "parts catalogs," other, more specialized databases that focus on specific pathways (such as intercellular signaling or degradation of chemical compounds by microbes) have been developed.

Gene Expression Data

DNA microarrays (or *gene chips*) are miniaturized laboratories for the study of gene expression. Each chip contains a deliberately designed array of probe molecules that can bind specific pieces of DNA or mRNA. Labeling the DNA or RNA with fluorescent molecules allows the level of expression of any gene in a cellular preparation to be measured quantitatively. Microarrays also have other applications in molecular biology, but their use in studying gene expression has opened up a new way of measuring genome functions.

Since the advancement of DNA microarray technology in the late 1990s, it has turned out that the increase in available gene expression data will eventually parallel the growth of the sequence and structure databases. Raw microarray information has been started to be made accessible to the general audience in particular databases, and the building up of a central data repository for such data is done (Gene Expression Omnibus).

Since a significant number of the early microarray experiments were performed at Stanford, their genome resources site has connections to raw information and databases that can be queried utilizing gene names or functional descriptions. Furthermore, the European Bioinformatics Institute has been instrumental in setting up of standards for deposition of microarray data in databases. Several databases additionally exist for the deposition of 2D gel electrophoresis results, including SWISS-2DPAGE and HSC-2DPAGE. 2D-PAGE is an innovation that permits quantitative investigation of protein concentrations in the cell, for many proteins at the same time. The combination of these two systems is an intense tool for understanding how genomes function.

Table 1 summarizes sources on the Web for some of the most important databases we've discussed in this section.

Table 1. Major Biological Data and Information Sources

Subject	Source	Link
Biomedical literature	PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
Nucleic acid sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Nucleotide
	SRS at EMBL/EBI	http://srs.ebi.ac.uk
Genome sequence	Entrez Genome	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
	TIGR databases	http://www.tigr.org/tdb/
Protein sequence	GenBank	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Protein
	SWISS-PROT at ExPASy	http://www.expasy.ch/spro/
	PIR	http://www-nbrf.georgetown.edu
Protein structure	Protein Data Bank	http://www.rcsb.org/pdb/
	Entrez Structure DB Protein and peptide mass spectroscopy PROWL	http://prowl.rockefeller.edu
Post-translational modifications	RESID	http://www-nbrf.georgetown.edu/pirwww/search/textresid.html
Biochemical and biophysical information	ENZYME	http://www.expasy.ch/enzyme/
	BIND	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Structure
Biochemical pathways	PathDB	http://www.ncgr.org/software/pathdb/
	KEGG	http://www.genome.ad.jp/kegg/
	WIT	http://wit.mcs.anl.gov/WIT2/
Microarray	Gene Expression Links	http://industry.ebi.ac.uk/~alan/MicroArray/
2D-PAGE	SWISS-2DPAGE	http://www.expasy.ch/ch2d/ch2d-top.html
Web resources	The EBI Biocatalog	http://www.ebi.ac.uk/biocat/
	IUBio Archive	http://iubio.bio.indiana.edu

Searching Biological Databases

There are numerous biological databases, and many alternative web interfaces that provide access to the same sets of data. Which one to use depends on personal needs, but it's necessary to be

aware of what kind of data the central data repositories are, and how often the peripheral databases are synchronized with the central data sources.

The two most established databases are NCBI's GenBank, for DNA sequences; and the Protein Data Bank (PDB), for molecular structure data. Each database has its own deposition procedures. However, both NCBI and PDB have well developed, automated, web-based deposition systems that do not change often over time.

GenBank

NCBI, in cooperation with EMBL and other international organizations, provides the most complete collection of DNA sequence data in the world - the database, known as GenBank.

NCBI maintains sequence data from every organism, every source, every type of DNA—from mRNA to cDNA clones to expressed sequence tags (ESTs) to high-throughput genome sequencing data and information about sequence polymorphisms. Users of the NCBI database need to be aware of the differences between these datatypes so that they can search the data set that's most appropriate for the work they're doing. The main sequence types that you'll encounter in a full GenBank search include:

mRNA

Messenger RNA, the product of transcription of genomic DNA. mRNA may be edited by the cell to remove introns (in eukaryotes) or in other ways that result in differences from the transcribed genomic DNA. May be "partial" or "complete"; an mRNA may not cover the complete coding sequence of a gene.

cDNA

A DNA sequence artificially generated by reverse transcription of mRNA. cDNA represents the coding components of the genomic DNA region that produced the mRNA. May be "partial" or "complete."

Genomic DNA

A DNA sequence from genome sequencing that contains both coding and noncoding DNA sequences. May contain introns, repeat regions, and others. Genomic DNA is generally "complete"; it's a result of multiple sequencing experiments over a single stretch of a genome, and can generally be relied upon as a fairly good representation of the real DNA sequence of that region.

EST

Short cDNA sequences prepared from mRNA extracted from a cell under particular conditions or in specific developmental phases. ESTs are used for quick identification of genes and don't cover the entire coding sequence of a gene.

GSS

Genome survey sequence. Single-time sequenced part of DNA direct from the genome projects. Covers each region of sequence only once and may contain a relatively large percentage of sequencing errors. Genome survey sequence is included in a search only when search a very new hypothetical gene annotations in a genome project that is still in progress.

There are two ways to search GenBank. The first is to use a text-based query to search the annotations associated with each DNA sequence entry in the database. The second is to use a method called BLAST to compare a query DNA (or protein) sequence to a sequence database. Here's a sample

GenBank record. Each GenBank entry contains annotation—information about the gene's identity, the conditions under which it was characterized, etc.—in addition to sequence (Fig. 4).

Listeria monocytogenes Sod (sod) gene, sod-2 allele, partial cds

GenBank: AY533467.1

FASTA Graphics PopSet

Go to:

```
LOCUS       AY533467                405 bp    DNA     linear   BCT 26-JUL-2016
DEFINITION  Listeria monocytogenes Sod (sod) gene, sod-2 allele, partial cds.
ACCESSION   AY533467
VERSION     AY533467.1
KEYWORDS    .
SOURCE      Listeria monocytogenes
  ORGANISM  Listeria monocytogenes
            Bacteria; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (bases 1 to 405)
  AUTHORS   Jegot,G., Lanotte,P., Brun,S., Watt,S., Quentin,R. and
            Mereghetti,L.
  TITLE     Genetic diversity of Listeria monocytogenes housekeeping genes:
            evidence for three evolutionary lineages within the species
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 405)
  AUTHORS   Jegot,G., Lanotte,P., Brun,S., Watt,S., Quentin,R. and
            Mereghetti,L.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-JAN-2004) Faculte de Medecine de Tours, Departement
            de Microbiologie Medicale et Moleculaire - Unite de Bacteriologie,
            2 bis bd Tonnelle, Tours 37032, France
FEATURES             Location/Qualifiers
     source          1..405
                     /organism="Listeria monocytogenes"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:1639"
     gene            <1..>405
                     /gene="sod"
                     /allele="2"
     CDS             <1..>405
                     /gene="sod"
                     /allele="2"
                     /codon_start=3
                     /transl_table=11
                     /product="Sod"
                     /protein_id="AAS22321.1"
                     /translation="SAEELVTNLDSVPEDIRGAVRNHGGGHANHTLFWSILSPNGGGA
                     PTGNLKAAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNDGKLEIVSTANQDSPLS
                     DGKTPVLGLDVWEHAYYLKFQNRRPEYIDTFW"
ORIGIN
        1 aatctgcgga agaattagtt actaacctag atagcgttcc tgaagatatt cgcggcgctg
       61 tccgtaacca cggtggcggt catgctaacc atacattgtt ctggtctatt cttagcccaa
      121 atggtggcgg cgctccaact ggcaatttaa aagcagcaat cgaaagcgaa ttcggtactt
      181 ttgacgaatt taaagaaaaa ttcaatgcag cagctgcagc acgttttggt tctggttggg
      241 cttggctagt agttaatgat ggcaaattag aaatcgtttc tacagctaac caagattctc
      301 cattaagcga tggcaaaaca cccgttcttg gcttagatgt ttgggaacat gcttactacc
      361 ttaaattcca aaaccgtcgt cctgaatata tcgacacatt ttgga
//
```

Fig. 4. GeneBank record of Listeria monocytogenes superoxide dismutase gene

This sample GenBank record shows the types of fields that can be found in a record from the GenBank Nucleotide database. In the record could be found the relevant information for the identity of the protein product, the sequence of the protein product, and its starting and ending point within the

gene, to the authors who submitted the record and the journal references in which the experiment was described. The GenBank search interface is nearly identical to the PubMed search interface. The Advanced features for searching work the same way in the Protein, Nucleic Acid, and Genome databases as they do for PubMed, although the specific fields that can be searched and limits that can be set are more or less different.

Saving search results

Sequences can be downloaded from NCBI in several file formats: the simple FASTA format, which is readable by many sequence analysis programs but contains little information other than sequence; the GenBank flat file format, which is a legacy flat file format that was used at GenBank earlier in its history; and the modern ASN.1 (Abstract Syntax Notation One) format. ASN.1 is a generic data specification, designed to promote database interoperability, that is now used for storage and retrieval of all datatypes—sequences, genomes, structure, and literature—at NCBI. The NCBI Toolkit, a code library for developing molecular biology software, relies on the ASN.1 specification. NCBI, and increasingly, other organizations, rely on the NCBI Toolkit for software development.

The casual database user or depositor doesn't have to think too much about file formats, except if database files are to be exported and read by another piece of software. NCBI's forms-based interfaces convert user-entered data into the appropriate format for deposition, and the availability of GenBank files in FASTA format means that most sequence analysis software can handle sequence files you download from NCBI without complicated conversions.

When saving results of a GenBank search, the format in which to save them can be easily chosen. A particularly foolproof format in which to save your sequence files if you're going to process them with other software is the FASTA format. FASTA files have a simple format, a single comment line that begins with a > character, followed by single-character DNA sequence on as many lines as needed to hold the sequence, with no breaks. Of course, some information associated with the gene is lost when you save the data in FASTA format, but if the program can't read that extra data, it won't be useful to have it anyway.

Here's a sample of data in FASTA format:

> gene identifier and comments here

MATVQEIRNAQRADGPATVLAIGTATPAHSVNQADYPDYYFRITKSEHMTELKEKFKRMCDKSMIKKRYMYLTEEILKENPN
MCAYMAPSLDARQDIVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLIGLRPSVKRFMMYQQG
CFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPADTHLDSLVGQALFGDGAAAVIVGADPDTSVERPLYQLVSTSQTI
LPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLSEAFAPLGISDWNSIFWIAHPGGPAILDQVESKLGLKGEKLKATRQVL
SEYGNMSSACVLFILDEMRKKSVEEAKATTGEGLDWGVLFGFGPGLTVETVVLHSVPIKA

To save your files in FASTA format, simply use the pulldown menu at the top of the results page. When you first see it, it will say "Summary," but you can change it to FASTA, ASN.1, and other formats. Once you've chosen your format, you can click the Save button to save all your sequences into one big FASTA-format file. Figure 5 shows you how to change the file formats when doing a GenBank search.

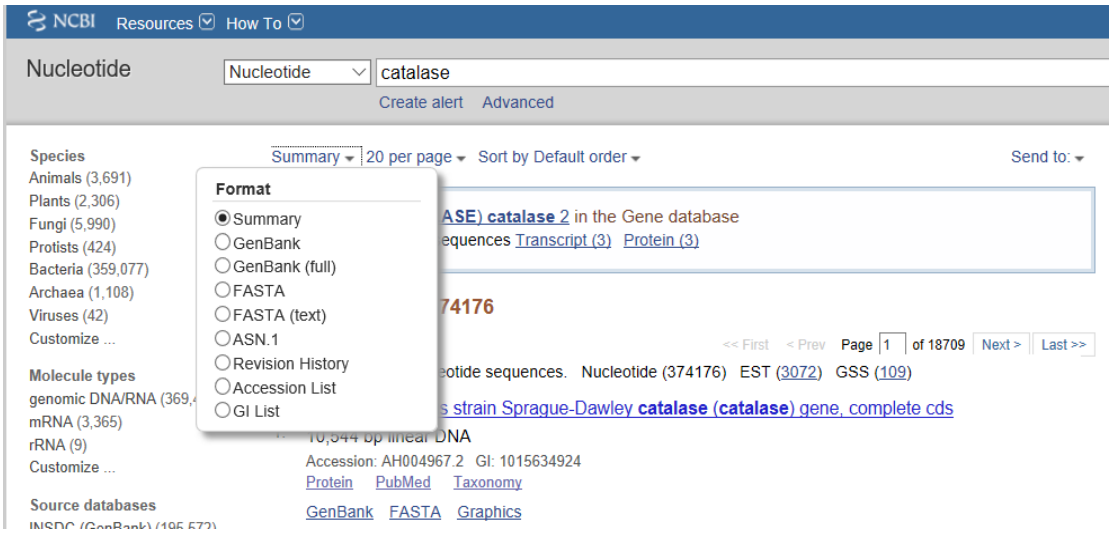


Figure 5. Selecting the file format to write out a GenBank search result

Saving large result sets

Modern bioinformatics studies increasingly deal with large amounts of sequence data. For example, gene finding programs are verified on hundreds or thousands of DNA sequences; comprehensive studies of protein families can involve analysis of up to thousands of protein sequences as well. In such cases it would be better to use an automated tool that can return a large number of sequences based on criteria you specify.

NCBI provides just such a tool in the form of Batch Entrez. Batch Entrez is one of the tools that allows the user to select sequences by source organism, by an Entrez query (using the query structure described in the section on PubMed), or by a list of accession numbers (provided by the user in the form of a text file). The results of a Batch Entrez search are then packaged in a file that is downloaded to the user's computer, where the complete result set can be edited manually or using a script.

At this time, all the public databases have at least FTP sites that allows to download the entire database on the computer. That can take up a lot of space on the hard disk, but is more easier to handle a large set of results in comparison to the interactive web site. When having a local copy of the big databases of interest, a script can be written that can processes the database, looking for particular keyword of choice, and writing out the desired information from a file.

PDB

Unlike NCBI, the Protein Data Bank (PDB) contains only one type of molecular data: molecular structures of molecules and, to a growing extent, the underlying raw data sets from which the molecular structures were modeled. It offers a number of services for submitting and retrieving three-dimensional structure data. The home page of the RCSB site provides links to services for depositing three-dimensional structures, information on how to obtain the status of structures undergoing processing for submission, ways to download the PDB database, and links to other relevant sites and software.

Figure 6. PDB features

The main information stored in the PDB consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. They are available in several formats (PDB, mmCIF, XML). A typical PDB file contains a text that describes the protein, citation information, and the details of the structure solution, followed by the sequence and a list of the atoms and their coordinates. The PDB files can be viewed directly using a text editor. Online tools, such as the ones on the RCSB PDB website, allow to search and explore the information under the PDB header, including information on experimental methods and the chemistry and biology of the protein (Fig. 7).

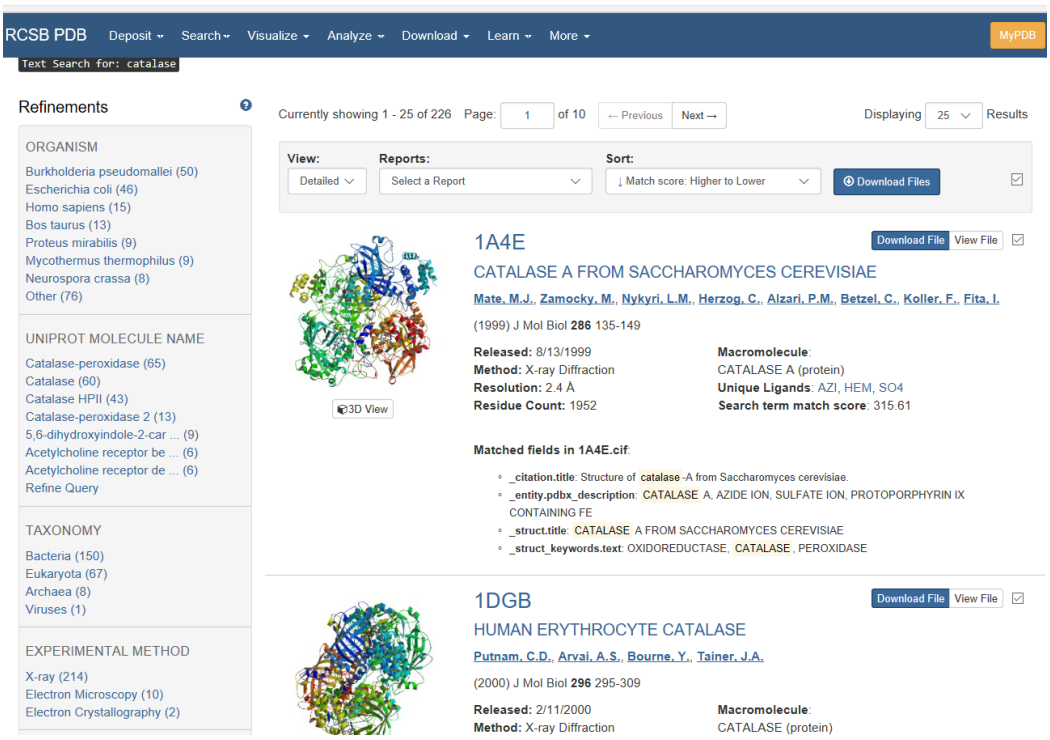


Figure 7. Query results at the PDB

The structure files may be viewed using one of several free and open source computer programs, including Jmol, Pymol, VMD, and Rasmol. Other non-free, shareware programs include ICM-Browser, MDL Chime, UCSF Chimera, Swiss-PDB Viewer, StarBiochem (a Java-based interactive molecular viewer with integrated search of protein databank), Sirius, and VisProt3DS (a tool for Protein Visualization in 3D stereoscopic view in anaglyph and other modes), and Discovery Studio. The RCSB PDB website contains an extensive list of both free and commercial molecule visualization programs and web browser plugins, as shown in Figure 8.

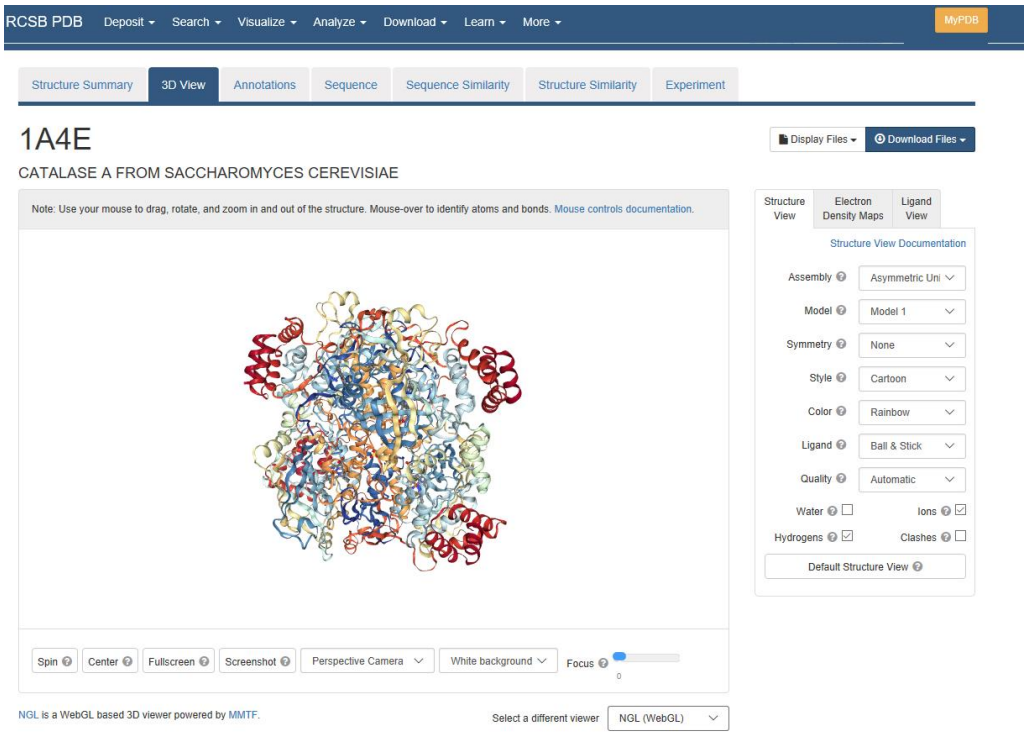


Figure 8. Viewing a PDB file using a browser plug-in

Depositing Data into the Public Databases

In addition to downloading information from the public databases, you may also submit your own results.

GenBank Deposition

Deposition of sequences to GenBank has been made extremely simple by NCBI. Users depositing only a few sequences can use the web-based BankIt tool, which is a self-explanatory form-based interface accessible from the GenBank main page at NCBI. NCBI has recently established two special submission paths: EST sequences should be submitted through dbEST, rather than to GenBank, and genome survey sequences through dbGSS.

PDB Deposition

Deposition of structures to the PDB are done using the wwPDB OneDep System that integrates data validation software with the deposition process so that the user can receive feedback on data quality during the deposition process. wwPDB OneDep System is tied in with the curation tools the PDB uses to prepare structure data for inclusion in the data bank.

Finding Software

Bioinformatics is a broad field, attracting researchers from many disciplines, and articles about new research developments in bioinformatics are widely distributed in the literature. If you're looking for cutting-edge developments, journals such as *Bioinformatics*, *Nucleic Acids Research*, *Journal of Molecular Biology*, and *Protein Science* often publish papers describing innovations in computational biology methods.

If you're looking for proven software for a particular application, there are a number of reliable web resource lists that link to computational biology software sites. Most of the major biological databases have software resource listings and the necessary motivation to keep their listings up-to-date. The PDB links to the best free software packages for macromolecular structure refinement, visualization, and dynamics. ExPASy and NCBI portals provide links to many tools for protein and DNA sequence analysis.

Judging the Quality of Information

The ability to judge the quality of information and software will improve as you continue to learn the field. One of the first things to consider when evaluating software, data, or information found on the Internet is the source. If you don't know the authors presenting the information by reputation, search for information about their affiliation and credentials available on the web site. Their expertise related to the topic or purpose of the web site is also important. An individual academic researcher's site doesn't always have the same need to be all-inclusive as a publicly funded database does. There is nothing inherently wrong with these offerings, but you should be aware of whether or not they are comprehensive, whether all their features are available to the casual user, and why.

Even data and software from national or international public sites are not necessarily entirely correct. It has been estimated that any given sequence in GenBank is likely to contain at least one error. While these errors generally don't render the data meaningless, it's always best to be aware of such issues even when using top-of-the-line public resources. Like any other software you find on the Web, software offered by public agencies such as NCBI and the PDB may still be under development. You can use this software, and much of it is of good quality. If you're basing your research on a beta version (a version still under development) of a software package, just read the documentation carefully so that you know what problems still remain to be worked out.

References

1. Baxevanis A.D., Ouellette B. F. F. (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition, John Wiley & Son, New York
2. Elloumi M., Zomaya A. Y. (2011) Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley a& Son, New York
3. Liu L., Agren R., Bordel S., Nielsen J. (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Letters 584: 2556–2564.
4. Milne C.B., Kim P.J., Eddy J.A., Price N.D. (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnol J. 4(12):1653-70
5. Pevzner P., Shamir R. (2011) Bioinformatics for Biologists, 1st Edition, Cambrage University Press
6. Ramsden J. (2015) Bioinformatics: An Introduction, Springer-Verlag, London
7. Singh G. B. (2015) Fundamentals of Bioinformatics and Computational Biology, Springer International Publishing, Switzerland

Alignments and phylogenetic trees

Advanced level

Ventsislava Petrova
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Kliment Petrov
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Contents

Introduction

The idea of using sequence alignment is to find and compare pairs of related sequences. Biologically interesting problems, however, often involve comparing more than two sequences at once. BLAST or FASTA search can yield a large number of sequences that match the query. One approach to compare all these resulting sequences with each other is to perform pairwise alignments of all pairs of sequences, then study these pairwise alignments individually. It's more efficient (and easier to comprehend), however, if you compare all the sequences at once, then examine the resulting ensemble alignment. This process is known

as *multiple sequence alignment*. Multiple sequence alignments can be used to study groups of related genes or proteins, to infer evolutionary relationships between genes, and to discover patterns that are shared among groups of functionally or structurally related sequences.

Multiple Sequence Alignment

Multiple sequence alignment techniques are generally applied to protein sequences. They are used for both evolutionary and structural similarity search among the proteins encoded by each sequence in the alignment. The proteins with closely related functions are similar in both sequence and structure from organism to organism. However, that sequence tends to change more rapidly than structure in the course of evolution. In multiple alignments generated from sequence data alone, regions that are similar in sequence are usually found to be superimposable in structure as well.

Progressive Strategies for Multiple Alignment

A common approach to multiple sequence alignment is to progressively align pairs of sequences. This strategy can be described as follows: a starting pair of sequences is selected and aligned, then each subsequent sequence is aligned to the previous alignment. Like the Needleman - Wunsch and Smith-Waterman algorithms for sequence alignment, progressive alignment is an instance of a heuristic algorithm. It decomposes a problem into pieces, then choose the best solution to each piece without paying attention to the problem as a whole. In the case of progressive alignment, the overall problem (alignment of many sequences) is decomposed into a series of pairwise alignment steps.

Because it is a heuristic algorithm, progressive alignment isn't guaranteed to find the best possible alignment. However, it is efficient and produces biologically meaningful results. The methods used differ in several respects: how they choose the initial pair of sequences to align, whether they align every subsequent sequence to a single cumulative alignment or create subfamilies, and how they score individual alignments and alignments of individual sequences to previous alignments.

Multiple Alignment with Clustal Omega

One commonly used program for progressive multiple sequence alignment is Clustal Omega. The heuristic used in Clustal Omega is based on phylogenetic analysis. First, a pairwise distance matrix for all the sequences to be aligned is generated, and a guide tree is created using the neighbor-joining algorithm. Then, each of the most closely related pairs of sequences are aligned to each other. Next, each new alignment is analyzed to build a sequence profile. Finally, alignment profiles are aligned to each other or to other sequences until a full alignment is built.

This strategy produces reasonable alignments under a range of conditions. For example, it's not guaranteed for distantly related sequences. Pairwise sequence alignment by dynamic programming is very accurate for closely related sequences regardless of which scoring matrix or penalty values are used. Using multiple sequences to create profiles increases the accuracy of pairwise alignment for more distantly related sequences.

There are several parameters involved in multiple sequence alignment - scoring matrices and gap penalties associated with the pairwise alignment steps, weighting parameters that alter the scoring matrix used in sequence-profile and profile-profile alignments. In Clustal Omega, these are set from the Set your parameters menu (Fig. 1).

The pairwise alignment parameters are familiar and have the same meaning in multiple alignment as they do in pairwise alignment. The multiple alignment parameters include gap opening and gap extension penalties for the multiple alignment process—to be used when fine-tuning alignments—and a maximum allowable delay, in terms of sequence length, for the start of divergent sequences at the beginning of the alignment.

One of Clustal Omega's heuristics is that, in protein sequence alignment, different scoring matrices are used for each alignment based on expected evolutionary distance. If two sequences are close neighbors in the tree, a scoring matrix optimized for close relationships aligns them. Distant neighbors are aligned using matrices optimized for distant relationships. Thus, when prompted to choose a series of matrices in the Multiple Alignment Parameters menu, it means just that: use BLOSUM62 for close relationships and BLOSUM45 for more distant relationships, rather than the same scoring matrix for all pairwise alignments.

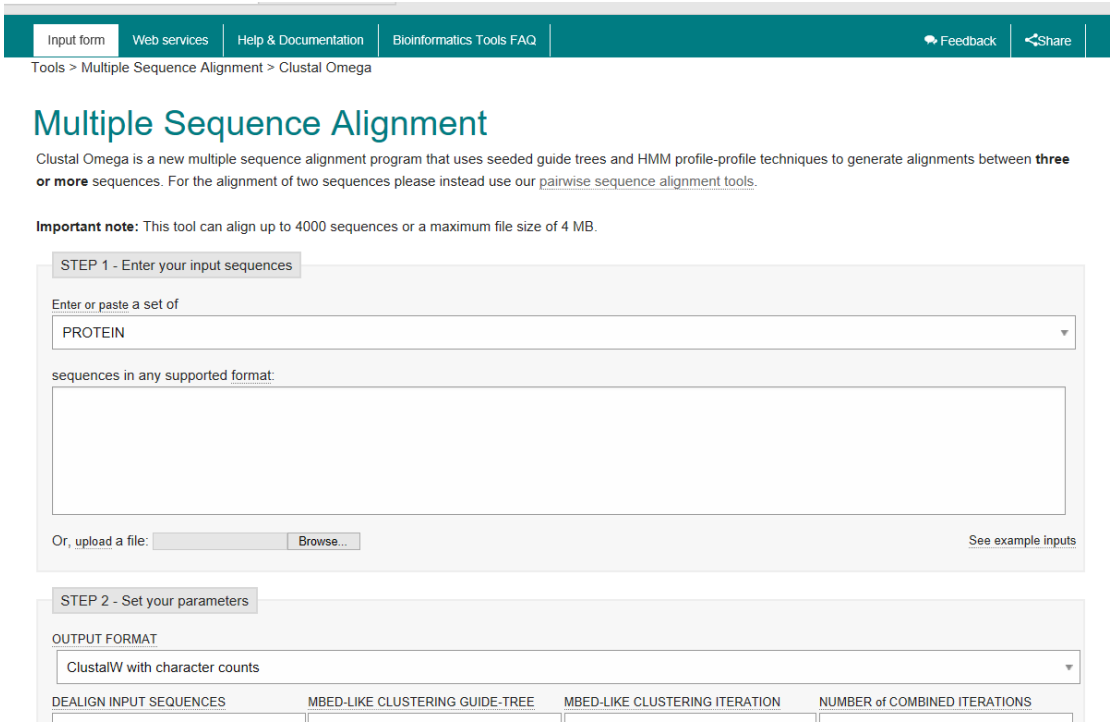


Fig. 1. Clustal Omega multiple sequence alignment program

Sequence Logos

A way to view sequence alignments, and one which has become quite popular recently, is the sequence logo format. This format is especially good for shorter sequence regions, such as protein motifs. Consensus sequences represent each position in an alignment with the residue that is most commonly found in that position. Sequence logos, as illustrated in Figure 2, are a graphical way to represent relative frequencies, information content, order of substitution preference, and other characteristics of each site in an alignment.

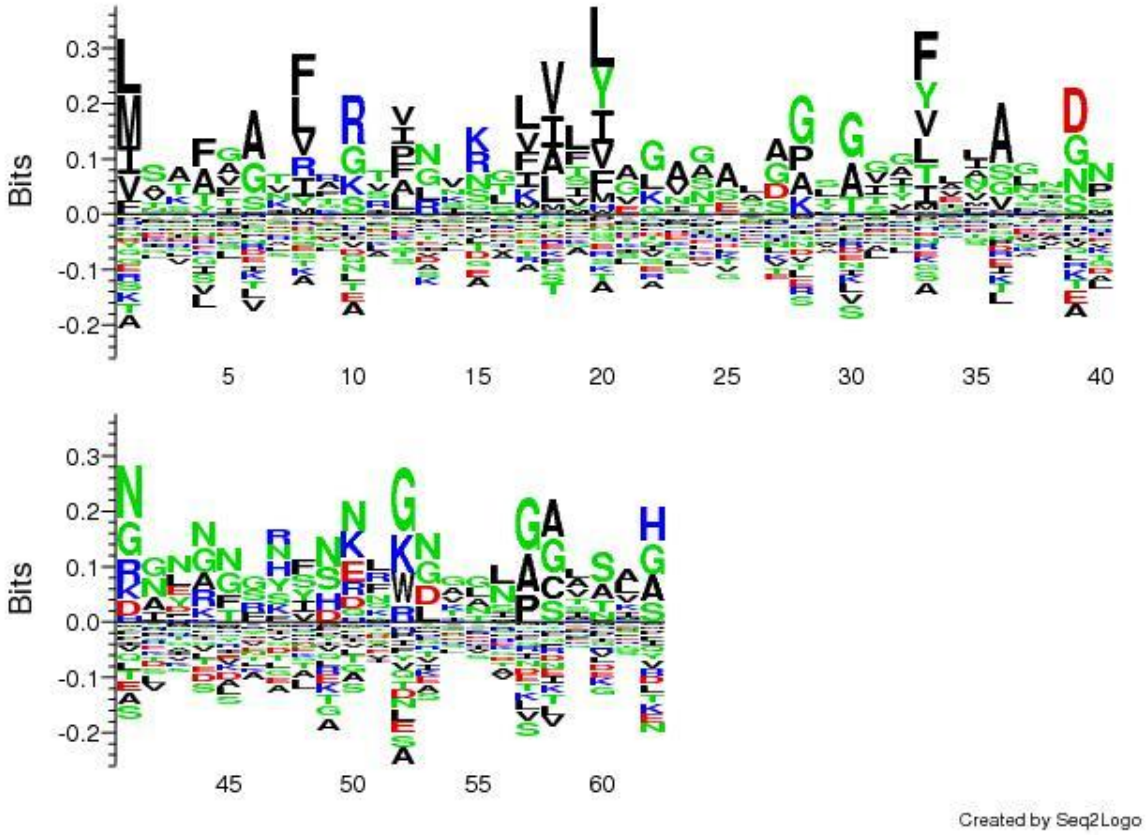


Figure 2. A sequence logo

The software for creating sequence logos is part of a larger group of programs called the DELILA package. You actually need only two of the many DELILA programs (*alpro* and *makelogo*) to create logos from aligned sequences. An easier approach for the novice is to use the Sequence logo web server. Aligned sequences can be submitted to this server in FASTA alignment format.

Phylogenetic Analysis

One of the applications of the multiple sequence alignment is the phylogenetic inference. Phylogenetic inference is the process of developing hypotheses about the evolutionary relatedness of organisms based on their observable characteristics.

While hand-drawn trees of life may branch according to what is essentially an artist's conception of evolutionary relationships, modern phylogenetic trees are strictly binary. Accordingly, at any branch point, a parent branch splits into only two daughter branches. Binary trees can approximate any other branching pattern, and the assumption that trees are binary greatly simplifies the tree-building algorithms.

The length of branches in a quantitative phylogenetic tree can be determined in more than one way. For example, the evolutionary distance between pairs of sequences is one way to assign branch length.

While a phylogeny of species generally has a root, assuming that all species have a specific common ancestor, a phylogenetic tree derived from sequence data may be rooted or unrooted. It isn't too difficult to

calculate the similarity between any two sequences in a group and to determine where branching points belong. It is much harder to pinpoint which sequence in such a tree is the common ancestor, or which pair of sequences can be selected as the first daughters of a common ancestor. While some phylogenetic inference programs do offer a hypothesis about the root of a tree, most simply produce unrooted trees. Figure 3 and Figure 4 illustrate rooted and unrooted phylogenetic trees.

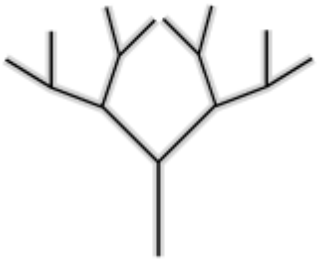


Figure 3. A rooted phylogenetic tree

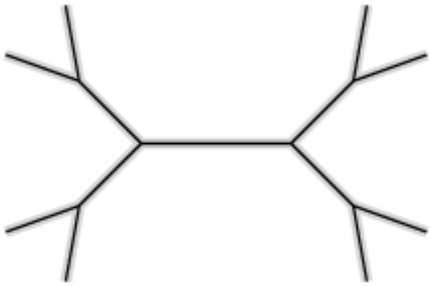


Figure 4. An unrooted phylogenetic tree

A phylogeny based on sequence alignment may be a tree, and it may describe a biological entity, but it takes far more than a single evolutionary analysis to draw conclusions about whole-organism phylogeny. Sequence-based phylogenies are quantitative. When they are built based on sufficient amounts of data, they can provide valuable, scientifically valid evidence to support theories of evolutionary history. However, a single sequence based phylogenetic analysis can only quantitatively describe the input data set. It isn't valid as a quantitative tool beyond the bounds of that data set.

It has been shown, by comparative analysis of phylogenies generated for different protein and gene families, that one protein may evolve more quickly than another, and that a single protein may evolve more quickly in some organisms than in others. Thus, the phylogenetic analysis of a sequence family is most informative about the evolution of that particular gene. Only by analysis of much larger sets of data can theories of whole-organism phylogeny be suggested.

Phylogenetic Trees Based on Pairwise Distances

One of the easiest algorithms for tree drawing is the pairwise distance method. This method produces a rooted tree. The algorithm is initialized by defining a matrix of distances between each pair of sequences

in the input set. Sequences are then clustered according to distance, in effect building the tree from the branches down to the root.

Distances can be defined by more than one measure, but one of the more common and simple measures of dissimilarity between DNA sequences is the Jukes-Cantor distance, which is logarithmically related to the fraction of sites at which two sequences in an alignment differ. The fraction of matching positions in an ungapped alignment between two unrelated DNA sequences approaches 25%. Consequently, the Jukes-Cantor distance is scaled such that it approaches infinity as the fraction of unmatched residue pairs approaches 75%.

The pairwise clustering procedure used for tree drawing (UPGMA, unweighted pair group method using arithmetic averages) is intuitive. Each sequence is assigned to its own cluster, and a branch of the tree is started for that sequence at height zero in the tree. Then, the two clusters that are closest together in terms of whatever distance measure has been chosen are merged into a single cluster. A branch point (or node) is defined that connects the two branches. The node is placed at a height in the tree that reflects the distance between the two branches that have been joined. This process is repeated iteratively, until there are only two clusters left. When they are joined, the root of the tree is defined. The branch lengths in a tree constructed using this process theoretically reflect evolutionary time.

Phylogenetic Trees Based on Neighbor Joining

Neighbor joining is another distance matrix method. It eliminates a possible error that can occur when the UPGMA method is used. UPGMA produces trees in which the branches that are closest together by absolute distance are placed as neighbors in the tree. This assumption places a restriction on the topology of the tree that can lead to incorrect tree construction under some conditions.

In order to get around this problem, the neighbor-joining algorithm searches not just for minimum pairwise distances according to the distance metric, but for sets of neighbors that minimize the total length of the tree. Neighbor joining is the most widely used of the distance-based methods and can produce reasonable trees, especially when evolutionary distances are short.

Phylogenetic Trees Based on Maximum Parsimony

A more widely used algorithm for tree drawing is called parsimony. *Parsimony* is related to a principle that states the simplest explanation is probably the correct one. Parsimony searches among the set of possible trees to find the one requiring the least number of nucleic acid or amino acid substitutions to explain the observed differences between sequences.

The only sites considered in a parsimony analysis of aligned sequences are those that provide evolutionary information — that is, those sites that favor the choice of one tree topology over another. A site is considered to be informative if there is more than one kind of residue at the site, and if each type of residue is represented in more than one sequence in the alignment. Then, for each possible tree topology, the number of inferred evolutionary changes at each site is calculated. The topology that is maximally parsimonious is that for which the total number of inferred changes at all the informative sites is minimized. In some cases there may be multiple tree topologies that are equally parsimonious.

As the number of sequences increases, so does the number of possible tree topologies. After a certain point, it is impossible to exhaustively enumerate the scores of each topology. A shortcut algorithm that finds the maximally parsimonious tree in such cases is the branch-and-bound algorithm. This algorithm establishes an upper bound for the number of allowed evolutionary changes by computing a tree using a fast

or arbitrary method. As it evaluates other trees, it throws out any exceeding this upper bound before the calculation is completed.

Phylogenetic Trees Based on Maximum Likelihood Estimation

Maximum likelihood methods also evaluate every possible tree topology given a starting set of sequences. Maximum likelihood methods are probabilistic. They search for the optimal choice by assigning probabilities to every possible evolutionary change at informative sites, and by maximizing the total probability of the tree. Maximum likelihood methods use information about amino acid or nucleotide substitution rates, analogous to the substitution matrices that are used in multiple sequence alignment.

Software for Phylogenetic Analysis

There is a variety of phylogenetic analysis software available for many operating systems. One of the most extensively is the PHYLIP package.

PHYLIP

The phylogenetic analysis package PHYLIP contains 30 programs that implement different phylogenetic analysis algorithms. Each of the programs runs separately, from the command line. By default, most of the programs look for an input file called *infile* and write an output file called *outfile*. Rather than entering parameters via command-line flags, as with BLAST, the programs have an interactive text interface that prompts you for information.

The following are frequently used the PHYLIP programs:

PROTPARS

Infers phylogenies from protein sequence input using the parsimony method

PROTDIST

Computes an evolutionary distance matrix from protein sequence input, using maximum likelihood estimation

DNAPARS

Infers phylogenies from DNA sequence input using parsimony

DNAPENNY

Finds all maximally parsimonious phylogenies for a set of sequences using a branch-and-bound search

DNAML

Infers phylogenies from DNA sequence input using maximum likelihood estimation

DNADIST

Computes a distance matrix from DNA sequence input using the Jukes-Cantor distance or one of three other distance criteria

NEIGHBOR

Infers phylogenies from distance matrix data using either the pairwise clustering or the neighbor joining algorithm

DRAWGRAM

Draws a rooted tree based on output from one of the phylogeny inference programs

DRAWTREE

Draws an unrooted tree based on output from one of the phylogeny inference programs

CONSENSE

Computes a consensus tree from a group of phylogenies

RETREE

Allows interactive manipulation of a tree by the user—not based on data

PHYLIP is a flexible package, and the programs can be used together in many ways. To analyze a set of protein sequences with PHYLIP, you can:

1. Read a multiple protein sequence alignment using PROTDIST and create a distance matrix.
2. Input the distance matrix to NEIGHBOR and generate a phylogeny based on neighbor joining.
3. Read the phylogeny into DRAWTREE and produce an unrooted phylogenetic tree.

Or, you can:

1. Read a multiple sequence alignment using PROTPARS and produce a phylogeny based on parsimony.
2. Read the phylogeny using DRAWGRAM and produce a rooted tree.

Each of the PHYLIP programs is thoroughly documented in the **.doc* files available with the PHYLIP distribution. This documentation has been converted into HTML by several groups.

Generating input for PHYLIP with Clustal Omega

The multiple sequence alignment program Clustal Omega draws phylogenetic trees with the neighbor joining method. Perhaps more importantly, it can read sequences in various input formats and then write PHYLIP - format files from multiple sequence alignments.

Profiles and Motifs

In addition to studying relationships between sequences, one of the most successful applications of multiple sequence alignments is in discovering novel, related sequences. This profile- or motif-based analysis uses data derived from multiple alignments to construct and search for sequence patterns.

Multiple sequence alignments can span the full sequence of the proteins involved or a single region of similarity, depending on their purpose. Multiple sequence alignments, such as the one shown in Figure 5, are generally built up by iterative pairwise comparison of sequences and sequence groups, rather than by explicit multiple alignment.

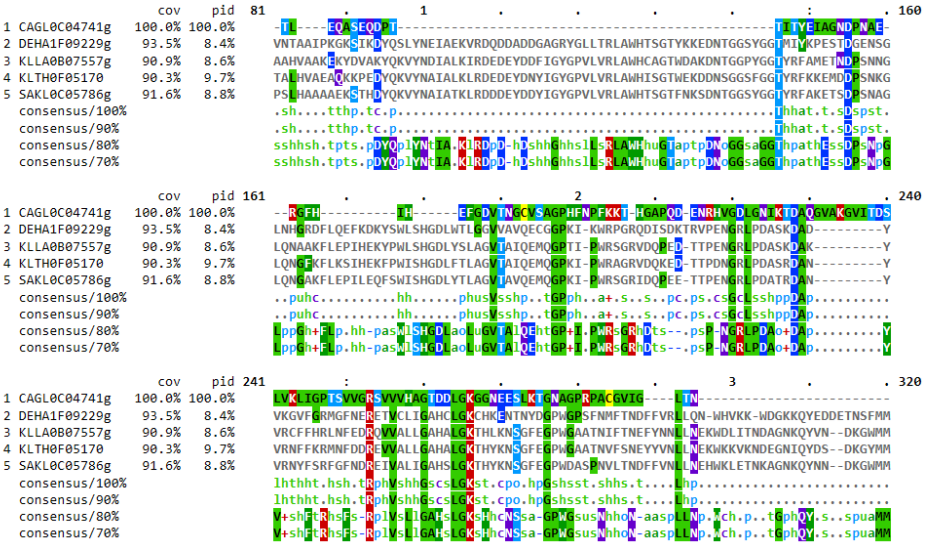


Figure 5. A multiple sequence alignment, shown using Clustal Omega

A sequence *motif* is a locally conserved region of a sequence, or a short sequence pattern shared by a set of sequences. The term "motif" most often refers to any sequence pattern that is predictive of a molecule's function, a structural feature, or family membership. Motifs can be detected in protein, DNA, and RNA sequences, but the most common use of motif-based analyses is the detection of sequence motifs that correspond to structural or functional features in proteins. Motifs are generated from multiple sequence alignments and can be displayed as patterns of amino acids (such as those in the Prosite database) or as sequence logos.

Motifs can be created for *protein families*, or sets of proteins whose members are evolutionarily related. Protein families can consist of many proteins that range from very similar to quite diverse. A sequence *profile* is a quantitative or qualitative method of describing a motif. A profile can be expressed in

its most basic form as a list of the amino acids occurring at each position in the motif. *Position-specific scoring matrix* (PSSM) is used when detecting a motif. Unlike a standard scoring matrix, the first dimension of the matrix is the length of the motif; the second dimension consists of the 20 amino acid possibilities. For each position in the matrix, there is a probability score for the occurrence of each amino acid. Most methods for developing position-specific scoring matrices normalize the raw probabilities with respect to a standard scoring matrix such as BLOSUM62.

Motif Databases

As profiles and other consensus representations of sequence families can be used to search sequence databases, it is no surprisingly that there are motif databases that can be searched using individual sequences. Motif databases contain representations of conserved sequences shared by a sequence family and their main use is in annotation of unknown sequences.

Motifs are generated by a variety of methods and with different aims. Some rely on automated analysis, but there is often a large amount of hands-on labor invested in the database by an expert curator. Because they store only those motifs that are present in reasonably large families, motif databases are small relative to GenBank, and they don't reflect the extent of the protein structure or sequence databases. An unsuccessful search against a motif database doesn't mean your sequence contains no detectable pattern. It could be part of a family that has not yet been curated or that doesn't meet the criteria of the particular pattern database you've searched. For proteins that do match defined families, a search against the pattern databases can yield a lot of homology information very quickly.

Blocks

Blocks, a service of the Fred Hutchinson Cancer Research Center, is an automatically generated database of ungapped multiple sequence alignments that correspond to the most conserved regions of proteins. Blocks is created using a combination of motif-detection methods, beginning with a step that exhaustively searches all spaced amino acid triplets in the sequence to discover a seed alignment, followed by a step that extends the alignment to find an aligned region of maximum length. The Blocks database provides several useful search services, including IMPALA (which uses the BLAST statistical model to compare a sequence against a library of profiles) and LAMA (Local Alignment of Multiple Alignments a program for comparing an alignment of your own sequences against a database of Blocks).

PROSITE

PROSITE is an expert-curated database of patterns hosted by the Swiss Institute of Bioinformatics. PROSITE uses a single consensus pattern to characterize each family of sequences. Patterns in PROSITE are carefully selected based on data published in the primary literature or on reviews describing the functionality of specific groups of proteins. PROSITE contains pattern information as well as position-specific scoring matrices that can detect new instances of the pattern.

Pfam

Pfam is a database of alignments of protein domain families. Pfam a curated database of over 2,700 gapped profiles, most of which cover whole protein domains. Its entries are generated automatically by applying a clustering method. Pfam entries begin with a *seed alignment*, a multiple sequence alignment that the curators are confident is biologically meaningful and that may involve some manual editing. From each seed alignment, a profile hidden Markov model is constructed and used to search a nonredundant database

of available protein sequences. A full alignment of the family is produced from the seed alignments and any new matches. This process can be repeated to produce more extensive families and detect remote matches. Pfam entries are annotated with information extracted from the scientific literature, and incorporate structural data when available (Fig. 6).

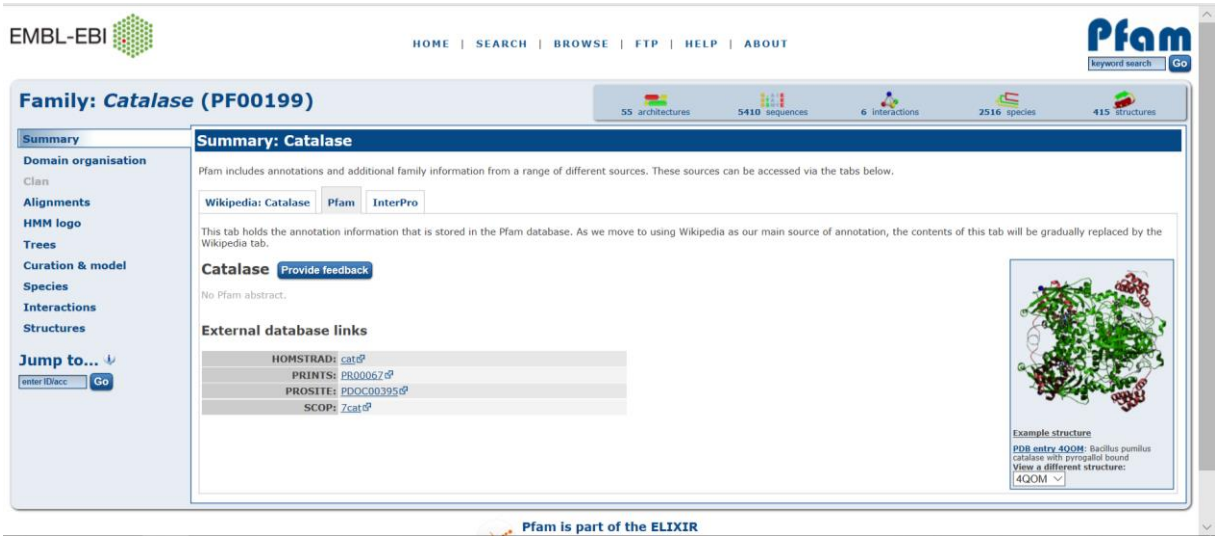


Fig. 6. Pfam entries representation

PRINTS-S

PRINTS-S is a database of protein motifs similar to PROSITE, except that it uses "fingerprints" composed of more than one pattern to characterize an entire protein sequence. Motifs are often short relative to an entire protein sequence. In PRINTS, groups of motifs found in a sequence family can define a signature for that family.

COG

NCBI's Clusters of Orthologous Groups (COG) database is a different type of pattern database. COG is constructed by comparing all the protein sequences encoded in the complete sequenced genomes. Each cluster must consist of protein sequences from at least three separate genomes. The principle of COG is that proteins that are conserved across these genomes from many diverse organisms represent ancient functions that have been conserved throughout evolution. COG entries can be accessed by organism or by functional category from the NCBI web site.

Accessing multiple databases

When analyzing a new sequence it is recommendable to use as many as possible motif databases. Blocks uses InterPro as one of the sources for its own patterns and contains only ungapped patterns, at the same time profiles contained in Pfam and PROSITE are gapped. Thus keeping track of the best matches from each database, their scores, and (if available) the significance of the hit, will provide more profound information on the performed analysis.

One service that allows integrated searching of many motif databases is the European Bioinformatics Institute's Integrated Resource of Protein Domains and Functional Sites (InterPro). InterPro allows you to compare a sequence against all the motifs from Pfam, PRINTS, ProDom, and PROSITE. InterPro motifs are annotated with the name of the source protein, examples of proteins in which the motif occurs, references to the literature, and related motifs (Fig. 7).

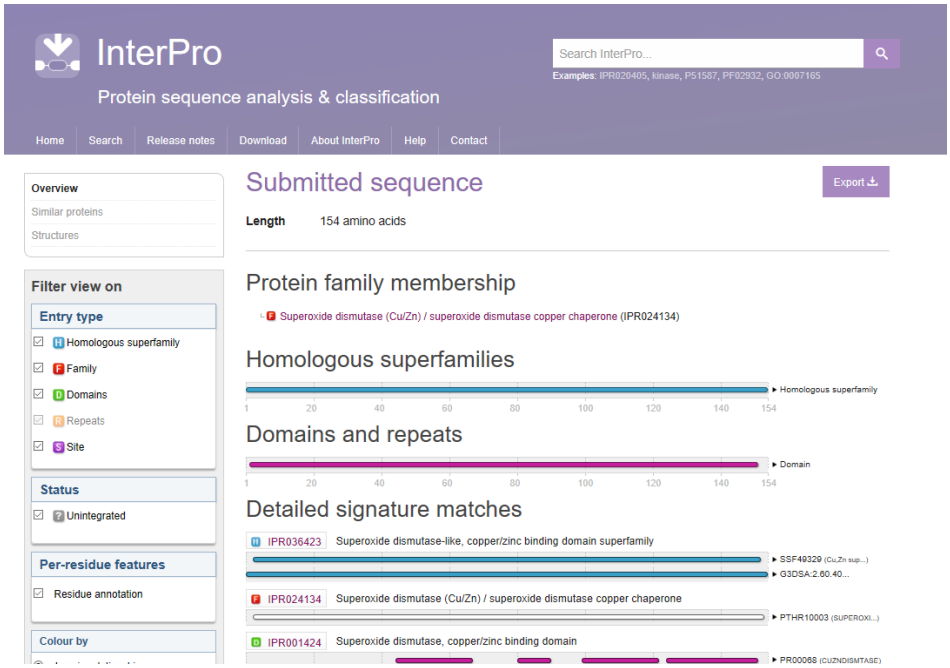


Fig. 7. Structure of InerPro database

Constructing and Using Your Own Profiles

Motif databases are useful when looking for protein families that are already well documented. However, if a new motif is found and it is intended to be used in GenBank search, or to look for patterns, it's necessary to build an own profiles. Several software packages and servers are available for *motif discovery* - a process of finding and constructing your own motifs from a set of sequences. The simplest way to construct a motif is to find a well-conserved section out of a multiple sequence alignment. A number of programs are commonly used to search for and discover motifs, like Block Maker, MEME and HMMer.

Incorporating Motif Information into Pairwise Alignment

Multiple sequence information can optimize pairwise alignments. The BLAST package contains two new modes that use multiple alignment information to improve the specificity of database searches. These modes are accessed through the *blastpgp* – a program used to run PSI-BLAST and PHI-BLAST. The last are specialized protein BLAST comparisons that are more sensitive than the standard BLASTP search.

Position Specific Iterative BLAST (PSI-BLAST) is an enhancement of the original BLAST program that implements profiles to increase the specificity of database searches. Starting with a single sequence, PSI-BLAST searches a database for local alignments using gapped BLAST and builds a multiple alignment and a profile the length of the original query sequence. The profile is then used to search the protein database again, seeking local alignments. This procedure can be restated any number of times. One caution of using

PSI-BLAST is that you need to know where to stop. Errors in alignment can be magnified by iteration, giving rise to false positives in the ultimate sequence search. The NCBI PSI-BLAST server is probably the optimal way to run a PSI-BLAST search.

Pattern Hit Initiated BLAST (PHI-BLAST) takes a sequence and a preselected pattern found in that sequence as input to query a protein sequence database. The pattern must be expressed in PROSITE syntax, which is described in detail on the PHI-BLAST server site. PHI-BLAST can also initiate a series of PSI-BLAST iterations, and can be a standalone program or a (vastly more user-friendly) web server.

References

1. Baxevanis A.D., Ouellette B. F. F. (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition, John Wiley & Son, New York
2. Elloumi M., Zomaya A. Y. (2011) Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley a& Son, New York
3. Liu L., Agren R., Bordel S., Nielsen J. (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Letters 584: 2556–2564.
4. Milne C.B., Kim P.J., Eddy J.A., Price N.D. (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnol J. 4(12):1653-70
5. Pevzner P., Shamir R. (2011) Bioinformatics for Biologists, 1st Edition, Cambrage University Press
6. Ramsden J. (2015) Bioinformatics: An Introduction, Springer-Verlag, London
7. Singh G. B. (2015) Fundamentals of Bioinformatics and Computational Biology, Springer International Publishing, Switzerland

Omics and system biology

Advanced level

Ventsislava Petrova
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Kliment Petrov
BULGAP Ltd
Sofia, Bulgaria
http://www.bggap.eu

Zlatyo Uzunov
JST Corporation Ltd.
Sofia, Bulgaria
https://jst.bg/

Contents

Comparative genome analysis

The first complete genome sequences of living organism have become available not long ago. In 1995, the genomes of the first two bacteria, *Haemophilus influenzae* and *Mycoplasma genitalium*, were reported. One year later, the first archaeal (*Methanococcus jannaschii*) and the first eukaryotic (yeast *Saccharomyces cerevisiae*) genomes were completely sequenced. Next, in 1997 the sequencing of the genomes of the two best-studied bacteria, *Escherichia coli* and *Bacillus subtilis* was done. Many more bacterial and archaeal genomes, as well as the genomes of a multicellular eukaryotes, like the nematode *Caenorhabiditis elegans*, have been sequenced since then.

An outstanding outcome of these first genome projects is that at least one-third of the genes encoded in each genome had no known or predictable function. The prediction of the general function for many

of the remaining genes have been appeared possible. The depth of our ignorance becomes particularly obvious on examination of the genome of *Escherichia coli* K12, debatably the most extensively studied organism among both prokaryotes and eukaryotes. Even in this well-known model organism of molecular biologists, at least 40% of the genes have unknown function. On the other hand, it turned out that the level of evolutionary conservation of microbial proteins is rather uniform, with ~70% of gene products from each of the sequenced genomes having orthologs in distant genomes. Thus, the functions of many of these genes can be predicted simply by comparing different genomes and by transferring functional annotation of proteins from better-studied organisms to their orthologs from lesser-studied organisms. This makes comparative genomics a powerful tool for achieving a better understanding of the genomes and, subsequently, of the biology of the respective organisms.

Progress in genome sequencing

By the beginning of 2000, genomes of 23 different unicellular organisms (5 archaeal, 17 bacterial, and 1 eukaryotic) had been completely sequenced. Up to 2018 thousands of microbial and eukaryotic genomes were in different stages of completion with respect to sequencing. Periodically updated lists of both finished and unfinished publicly funded genome sequencing projects are available in the GenBank Entrez Genomes. A complete list of sequencing centers world-wide can be found at the NHGRI Web site. One can retrieve the actual sequence data from the NCBI FTP site or from the FTP sites of each individual sequencing center. A convenient sequence retrieval system is maintained also at the DNA Data Bank of Japan. In the framework of the Reference Sequences (RefSeq) project, NCBI has started to increase the lists of gene products with some valuable sequence analysis information, such as the lists of best hits in different taxa, predicted functions for uncharacterized gene products, frame-shifted proteins, etc. On the other hand, sequencing centers like TIGR regularly updates their sequence data, correct some of the sequencing errors and, accordingly, their sites may contain more recent data on unfinished genome sequences.

General-Purpose Databases for Comparative Genomics

Because the Web makes genome sequences available to anyone with Internet access, there exists a variety of databases that offer more or less convenient access to basically the same sequence data. However, several research groups, specializing in genome analysis, maintain databases that provide important additional information, such as operon organization, functional predictions, three-dimensional structure, and metabolic reconstructions.

PEDANT

This useful Web resource provides answers to most standard questions in genome comparison. PEDANT provides an easy way to ask simple questions, such as finding out how many proteins in *H. pylori* have known (or confidently predicted) three-dimensional structures or how many NAD^+-dependent alcohol dehydrogenases (EC 1.1.1.1) are encoded in the *C. elegans* genome. The list of standard PEDANT queries includes EC numbers, PROSITE patterns, Pfam domains, BLOCKS, and SCOP domains, as well as PIR keywords and PIR superfamilies (Fig.1.). Although PEDANT does not allow the users to enter their own queries, the variety of data available at this database makes it a convenient entry point into the field of comparative genome analysis.

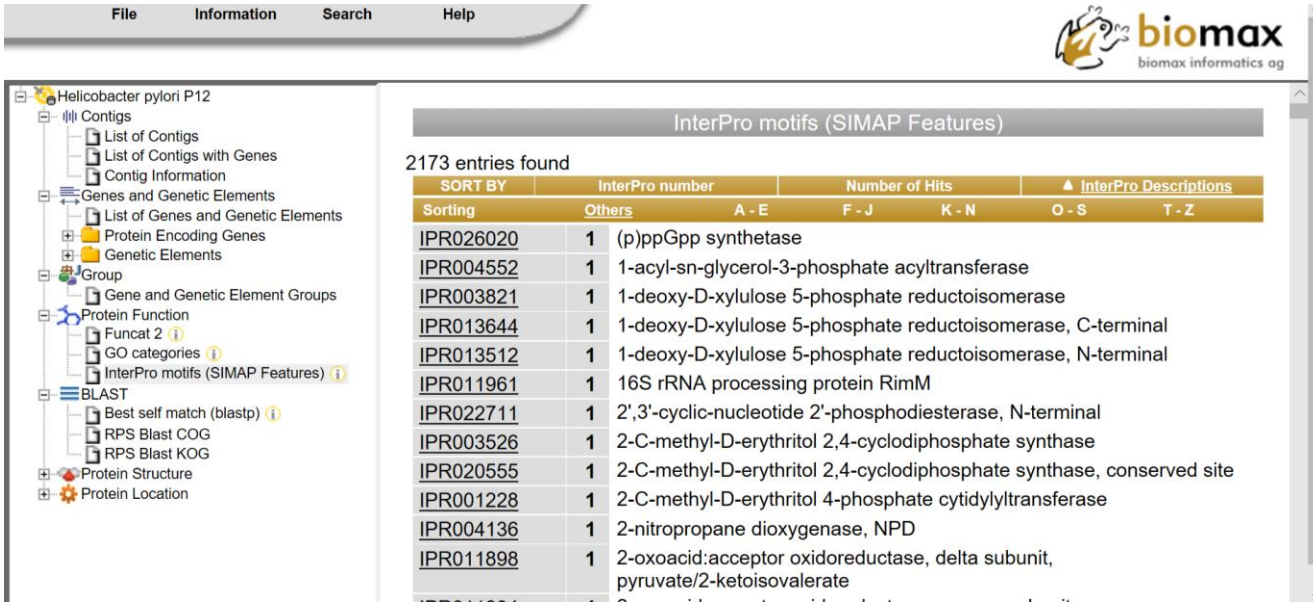
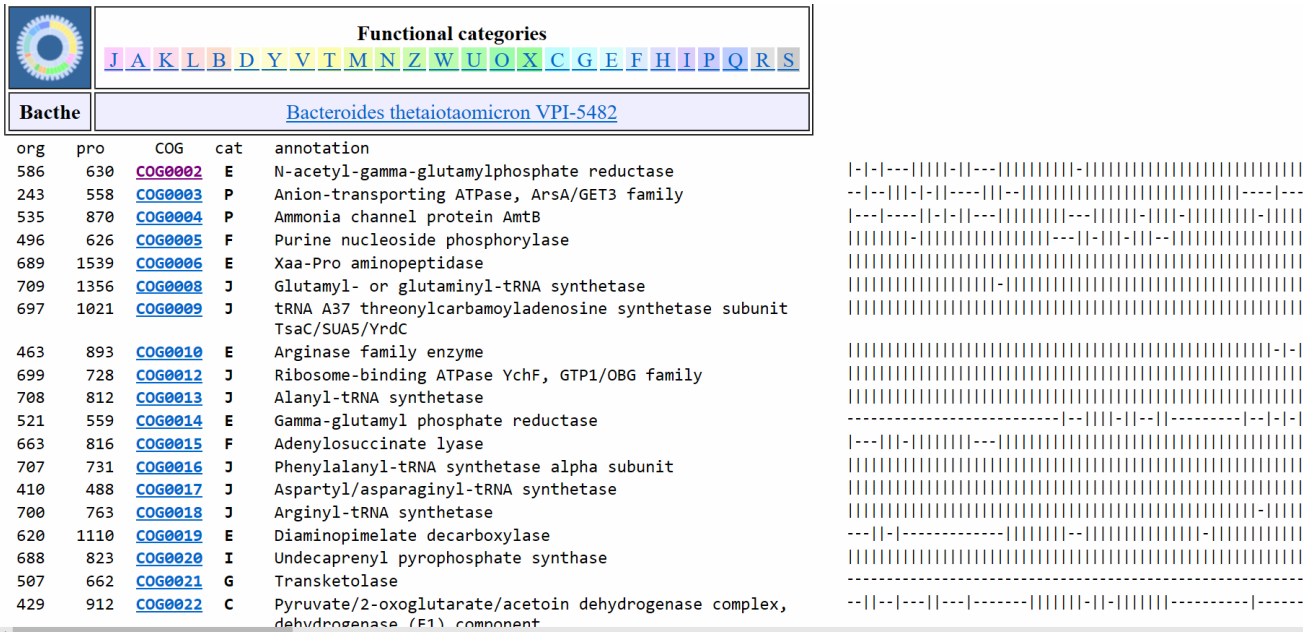


Fig. 1. *Helicobacter pylori P12* in PENDANT database

COGs

The Clusters of Orthologous Groups (COGs) database has been intended to simplify evolutionary studies of complete genomes and improve functional projects of individual proteins. It consists of more than 4,800 conserved families of proteins (COGs) from each of the completely sequenced genomes. Each COG contains orthologous sets of proteins from at least three phylogenetic lineages, which are assumed to have evolved from an individual ancestral protein. By definition, *orthologs* are genes that are connected by vertical evolutionary descent (the ‘‘same’’ gene in different species) as opposed to *paralogs*—genes related by duplication *within* a genome. Because orthologs typically perform the same function in all organisms, delineation of orthologous families from diverse species allows the transfer of functional annotation from better-studied organisms to the lesser-studied ones. The protein families in the COG database are separated into 25 functional groups that include a group of uncharacterized

yet conserved proteins, as well as a group of proteins for which only a general function prediction only has been performed (Fig.2). This site is particularly useful for functional predictions in disputed cases, where protein similarity levels are fairly low. Due to the diversity of proteins in COGs, sequence similarity searches against the COG database can often suggest a possible function for a protein that otherwise has no clear database hits.



Functional categories

J A K L B D Y V T M N Z W U O X C G E F H I P Q R S

Bacthe — Bacteroides thetaiotaomicron VPI-5482

org	pro	COG	cat	annotation
586	630	COG0002	E	N-acetyl-gamma-glutamylphosphate reductase
243	558	COG0003	P	Anion-transporting ATPase, ArsA/GET3 family
535	870	COG0004	P	Ammonia channel protein AmtB
496	626	COG0005	F	Purine nucleoside phosphorylase
689	1539	COG0006	E	Xaa-Pro aminopeptidase
709	1356	COG0008	J	Glutamyl- or glutaminyl-tRNA synthetase
697	1021	COG0009	J	tRNA A37 threonylcarbamoyladenosine synthetase subunit TsaC/SUA5/YrdC
463	893	COG0010	E	Arginase family enzyme
699	728	COG0012	J	Ribosome-binding ATPase YchF, GTP1/OBG family
708	812	COG0013	J	Alanyl-tRNA synthetase
521	559	COG0014	E	Gamma-glutamyl phosphate reductase
663	816	COG0015	F	Adenylosuccinate lyase
707	731	COG0016	J	Phenylalanyl-tRNA synthetase alpha subunit
410	488	COG0017	J	Aspartyl/asparaginyl-tRNA synthetase
700	763	COG0018	J	Arginyl-tRNA synthetase
620	1110	COG0019	E	Diaminopimelate decarboxylase
688	823	COG0020	I	Undecaprenyl pyrophosphate synthase
507	662	COG0021	G	Transketolase
429	912	COG0022	C	Pyruvate/2-oxoglutarate/acetoin dehydrogenase complex, dehydrogenase (E1) component

Fig.2. *Bacteroides thetaiotaomicron* VPI-5482 functional categories in GOG

KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is focused on cellular metabolism. This database presents a comprehensive set of metabolic pathway charts, both general and specific, for each of the completely-sequenced genomes, as well as for *Schizosaccharomyces pombe, Arabidopsis thaliana, Drosophila melanogaster*, mouse, and human. Enzymes that are already identified in a particular organism are color-coded, so that one can easily trace the pathways that are likely to be present or absent in a given organism (Fig. 3). For the metabolic pathways covered in KEGG, lists of orthologous genes that code for the enzymes participating in these pathways are also provided. It is also indicated whenever these genes are adjacent, forming likely operons. A very convenient search tool allows the user to compare two complete genomes and identify all cases in which conserved genes in both organisms are adjacent or located relatively close (within 5 genes) to each other. The KEGG site is continuously updated and serves as an ultimate source of data for the analysis of metabolism in various organisms.

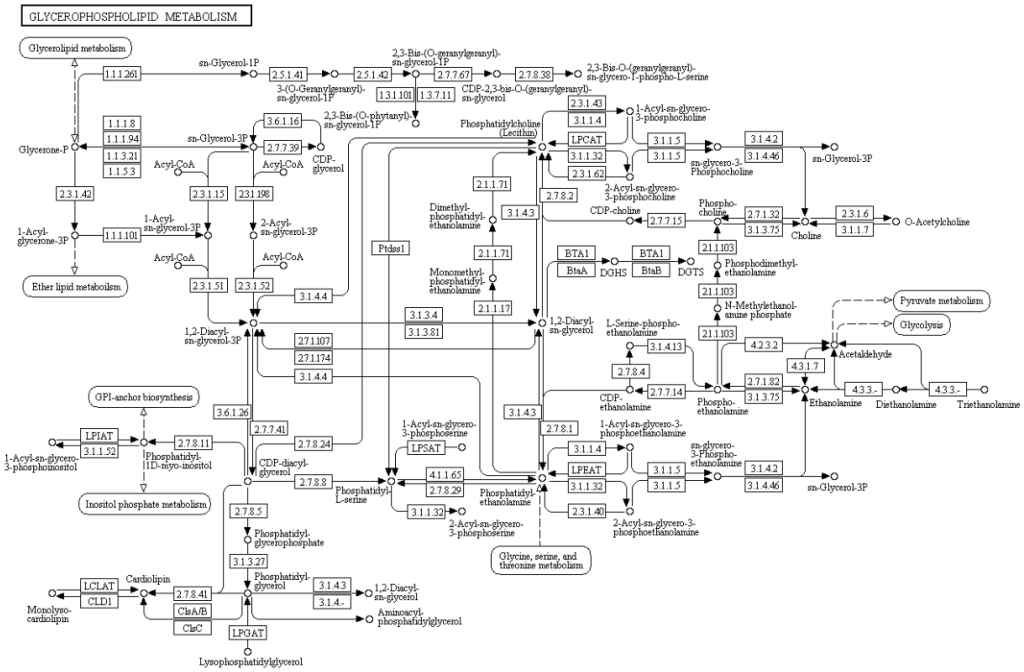


Fig. 3. Metabolic pathway chart of glycerophospholipid metabolism

MBGD

The Microbial Genome Database (MBGD) offers another convenient tool for comparative analysis of completely sequenced microbial genomes, the number of which is now growing rapidly (Fig. 4). Here, the homology relationships are based only on sequence similarity (BLASTP values of 10^{-2} or less). MBGD permits to submit several sequences at once (up to 2,000 residues) for searching against all of the completely sequenced genomes. The result is displayed as color-coded functions of the detected homologs, and shows their location on a circular genome map. The output of MBGD's BLAST search also shows the degree of overlap between the query and target sequences. For each sequenced genome, MBGD provides convenient lists of all recognized genes that are involved in a particular function, e.g., the biosynthesis of branched-chain amino acids or the degradation of aromatic hydrocarbons.

Fig. 4. MBGD database

Organism-Specific Databases

In addition to general genomics databases, exist a variety of databases for particular organism or a group of organisms. Although all of them are useful for specific purposes, those devoted to *E. coli*, *B. subtilis*, and yeast are probably the ones most widely used for functional assignments in other, less studied organisms.

Escherichia coli. The importance of *E. coli* for molecular biology is reflected in the large number of databases dedicated to this organism. One of them is maintained at the University of Wisconsin-Madison, the research groups that carried out the actual sequencing of the *E. coli* genome (Fig. 5). The Wisconsin group is also involved in sequencing the enteropathogenic *E. coli* O157:H7 and other enterobacteria, so their database is also very useful for analysis of enteric pathogens. Another useful database on *E. coli*, EcoCyc. It lists all experimentally studied *E. coli* genes and provides comprehensive coverage of the metabolic pathways identified in *E. coli*. The aim of another *E. coli* database, Bacteriome, is to provide an integrated protein interaction database for a high quality functional interaction dataset of *E. coli* proteins together with experimental datasets generated through tandem affinity purification screens.. Finally, Colibri and GenExpDB are the databases of choice for those interested in regulatory networks of *E. coli*. The *E. coli* Genetic Stock Center (CGSC) Web site also provides gene and function information.

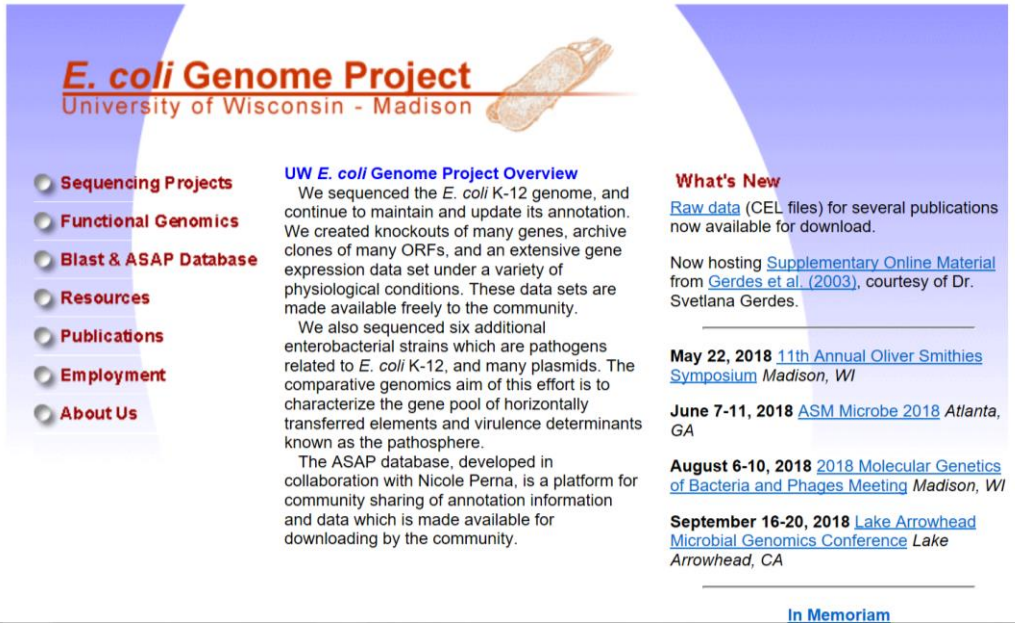


Fig.5. *E.coli* Genome Project

Mycoplasma genitalium. Mycoplasma has the smallest genome of all known cellular life forms, which offers some hints as to what is the lower limit of genes necessary to sustain life (the ‘‘minimal genome’’). Its comparison to the second smallest known genome, that of *Mycoplasma pneumoniae*, is available online. Recent data from VFDB provides insight into the range of *Mycoplasma* genes that can be mutated without loss of viability (Fig. 6). From both computational analysis and mutagenesis studies, it appears that 250–300 genes are absolutely essential for the survival of mycoplasmas.

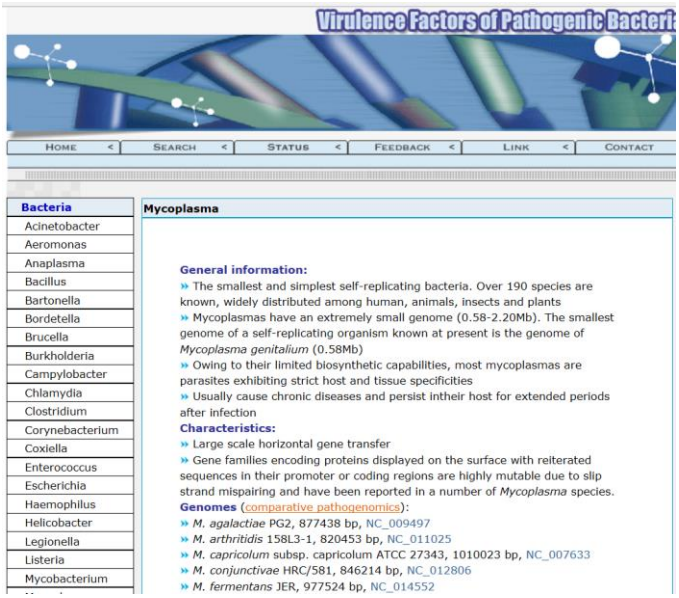
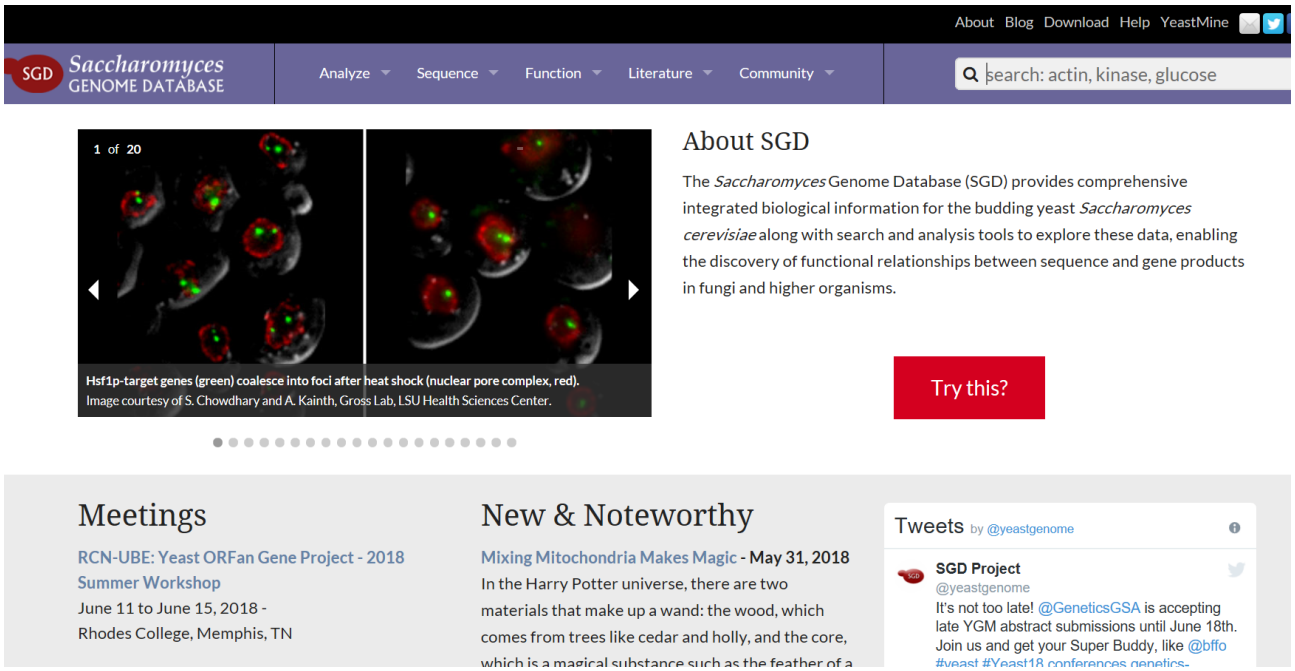


Fig. 6. *Mycoplasma* Genome Database at VFDB

Bacillus subtilis. The *B. subtilis* genome also attracts considerable attention from biologists and, like that of *E. coli*, is being actively studied from the functional perspective. The SubtiList World-Wide Web Server, maintained at the Institute Pasteur, is constantly updated to include the most recent information on functions of new *B. subtilis* genes. In addition, a DBTBS contains comprehensive database of the transcriptional regulation in *Bacillus subtilis* and contains upstream intergenic conservation information.

Saccharomyces cerevisiae. The major databases specifically devoted to the functional analysis of yeast *S. cerevisiae* genome is the Saccharomyces Genome Database (SGD) (Fig. 7). It provides regurally updated lists of yeast proteins with known or predicted functions, appropriate references, and mutant phenotypes and reflect the ongoing efforts aimed at complete characterization of all yeast proteins. SGD is probably the largest and most comprehensive source of information on the current status of the yeast genome analysis and includes the *Saccharomyces* Gene Registry.

Other useful sites for yeast genome analysis include *Saccharomyces cerevisiae* Promoter Database, listing known regulatory elements and transcriptional factors in yeast; and the *Saccharomyces* Cell Cycle Expression Database, presenting the first results on changes in mRNA transcript levels during the yeast cell cycle.



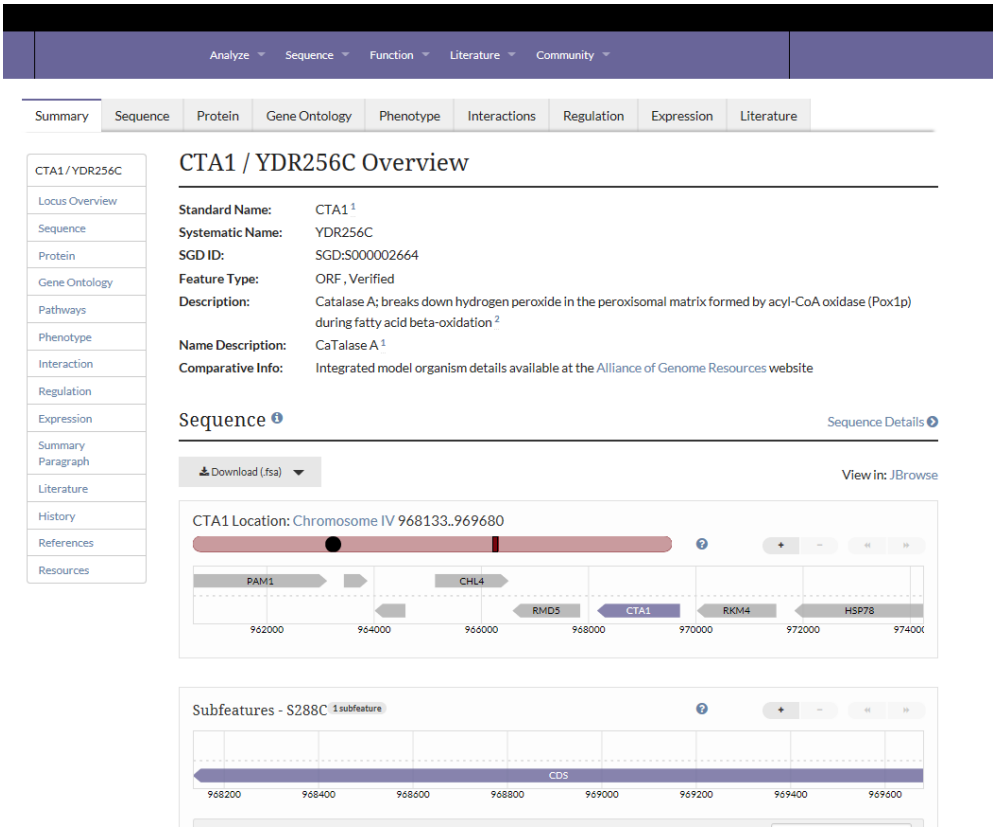


Fig. 7. *Saccharomyces* Genome Database

Genome analysis and annotation

One of the limiting steps in the most genome projects are the sequence analysis and annotation of the complete genomes. This task is particularly discouraging given the lack of functional information for a large number of genes even in the best-understood model organisms. The standard stages involved in the structural-functional annotation of uncharacterized proteins includes:

- ✓ sequence similarity searches using programs such as BLAST, FASTA, or the Smith-Waterman algorithm;
- ✓ identifying functional motifs and structural domains by comparing the protein sequence against PROSITE, BLOCKS, SMART, or Pfam;
- ✓ predicting structural features of the protein, such as likely signal peptides, transmembrane segments, coiled-coil regions, and other regions of low sequence complexity; and
- ✓ generating a secondary (and, if possible, tertiary) structure prediction.

All these steps have been automated in several software packages, such as GeneQuiz, MAGPIE, PEDANT, Imagene, and others. Of these, however, MAGPIE and PEDANT do not allow outside users to submit their own sequences for analysis and display only the authors' own results. GeneQuiz offers a limited number of searches (up to 100 a day) to general users but is still a good entry point for comparative genome analysis. It relies on unrealistically high cutoff scores to deduce homology, which results in relatively low sensitivity. One such package that is currently available for free downloading is SEALS, developed at NCBI. It consists of a number of UNIX-based tools for retrieving sequences from GenBank, running database search programs such as BLAST, viewing and analyzing search outputs, searching for sequence motifs, and predicting protein structural features. A similar package, called Imagene, has been developed at Universite´ Paris VI.

Genome Comparison for Prediction of Protein Functions

Analysis of the first sequenced bacterial, archaeal, and eukaryotic genomes using the sequence comparison methods failed to predict protein function for at least one-third of gene products in any given genome. In these cases, other approaches can be used that take into consideration all other available data, putting them into ‘‘genome context’’. These approaches rely on the same basic principle, that the organization of the genetic information in each particular genome reflects a long history of mutations, gene duplications, gene rearrangements, gene function divergence, and gene acquisition and loss that has produced organisms uniquely adapted to their environment and capable of regulating their metabolism in accordance with the environmental conditions. In this respect the cross-genome similarities can be assumed as meaningful in the *evolutionary* sense and thus are potentially useful for functional analysis. The most applicable comparative methods specifically employ information derived from multiple genomes thus achieving reliability and sensitivity that are not easily attainable with standard tools. Some of these new approaches are briefly reviewed below.

Transfer of Functional Information

The simplest and the most common way to exploit the information embedded in multiple genomes is the transfer of functional information from well-characterized genomes to poorly-studied ones. Indirectly, this is done through making a prediction for a newly sequenced gene on the basis of a database hit(s). There are, however, many pitfalls that tend to hamper accurate functional prediction on the basis of such hits. The most important ones relate to the lack of sufficient sensitivity, leading to error broadcast. Main reasons for that are due to the dependence on incorrect or imprecise annotations already present in the databases, and the difficulty in distinguishing orthologs from paralogs. The issue of orthology vs. paralogy is critical because transfer of functional information could be assumed as reliable for orthologs (direct evolutionary counterparts) but may not be quite consistent for the paralogs (products of gene duplications). All these problems are, in part, avoided in the COG system, which consists of carefully annotated sets of likely orthologs and does not rely on arbitrary cutoffs for assigning new proteins to them.

The COGs can be employed for annotation of newly-sequenced genomes using the COGNITOR program. This program allocates new proteins to COGs by comparing them to protein sequences from all genomes included in the COG database and detecting genome-specific best hits (BeTs). When three or more BeTs fall into the same COG, the query protein is considered a likely new COG member. The requirement of multiple BeTs for a protein to be assigned to a COG serves, to some extent, as a safeguard against the propagation of errors that might be present in the COG database itself. Indeed, if a COG contains one or even two false-positives, this will not result in a false assignment by COGNITOR under the three-BeT cutoff rule.

Phylogenetic Patterns (Profiles)

The COG-type analysis applied to multiple genomes provides for the root of *phylogenetic patterns*, which are potentially useful in many aspects of genome analysis and annotation. The phylogenetic pattern for each protein family (COG) is defined as the set of genomes in which the family is represented. The COG database is accompanied by a pattern search tool that allows the user to select COGs with a particular pattern. On this basis, tit is considered that the genes that are functionally related presumably have the same phylogenetic pattern. Because of these features, phylogenetic patterns can be used to improve functional predictions in complete genomes. When a particular genome is represented in the COGs for a subset of components of a particular complex or pathway but is missing in the COGs for other components, a focused search for the latter is justified. The same applies to cases in which a gene is found in one of two closely related genomes, but not the other.

Use of Phylogenetic Patterns for Differential Genome Display

The phylogenetic pattern approach and, specifically, the pattern search tool associated with the COGs can be used to perform systematic logical operations (AND, OR, NOT) on gene sets — an approach

called ‘‘differential genome display’’. This type of genome comparison permits to delineate subsets of gene products that are likely to contribute to the specific characteristics of the studied organisms, for example, thermophily. The use of this approach is of particular interest when identifying candidate drug targets in pathogenic bacteria. It seems logical to look for such targets among those genes that are shared by several pathogenic organisms, but are missing in eukaryotes. On the other hand, it is appealing to suggest that the best targets for new broad-spectrum antimicrobial agents would be genes that are shared by all pathogenic microbes, but not by any other organisms. However, such genes do not seem to exist. In this respect, it seems that the best solution when searching for such potentially universal antimicrobial agents is to isolate the genes that are present in most of the pathogens, but not in eukaryotes.

Study of Gene (Domain) Fusions

Another recently developed comparative genomic approach involves systematic analysis of protein and domain fusion (and fission). The basic hypothesis is that fusion would be maintained by selection only when it facilitates functional interaction between proteins, for example, kinetic coupling of consecutive enzymes in a pathway. Thus, proteins that are fused in some species can be expected to interact, perhaps physically or at least functionally, in other organisms. A straightforward example of functional inferences that can be drawn from domain fusion is seen in the histidine biosynthesis pathway, which in *E. coli* and *H. influenzae* includes two two-domain proteins, HisI and HisB. The two domains of HisI catalyze two consecutive steps of histidine biosynthesis and thus represent subunits that are likely to physically interact even when produced as separate proteins. In contrast, the two domains of HisB catalyze the seventh and ninth steps of the pathway and hence are not likely to physically interact. The COG database includes about 700 distinct multidomain architectures. Thus, using domain fusion for functional prediction has considerable empirical potential although this approach will not work for ‘‘promiscuous’’ domains such as, for example, the DNA-binding helix-turn-helix domain, which can be found in combination with a wide variety of other domains.

In addition, several databases have recently been developed for detecting domains and exploring architectures of multidomain proteins: Pfam, ProDom, and SMART.

From all of them, SMART seems to be the most advanced, combining high sensitivity of domain detection with accuracy, high speed, and extremely informative presentation of domain architectures. Rapid searches for protein domains, based on a modification of the PSI-BLAST program is now also available through the Conserved Domains Database (CDD) at NCBI.

Analysis of Operons

An approach that is conceptually similar to the analysis of gene fusions, but is more general, involves systematic analysis of gene ‘‘neighborhoods’’ in genomes. Because functionally linked genes frequently form operons in bacteria and archaea, gene adjacency may provide important functional suggestions. However, many functionally related genes never form operons, and, in many instances, adjacent genes are not connected in any way. Due to the lack of overall conservation of gene order in

prokaryotes, the presence of a pair of adjacent orthologous genes in three or more genomes or the presence of three orthologs in a row in two genomes can be considered a statistically meaningful event and can be used to infer potential functional interaction for the products of these genes. The simplest current tool for identification of conserved gene strings in any two genomes is available as part of KEGG. It allows the user to select any two complete genomes (e.g., *B. burgdorferi* and *R. prowazekii*) and look for all genes whose products are similar to each other and are located within a certain distance from each other (for example, separated by 0–5 genes). The results are displayed in a graphical format illustrating the gene order and the presumed functions of gene products. The conservation of gene position in phylogenetically distant bacteria suggests a functional connection.

Application of comparative genomics—reconstruction of metabolic pathways

To illustrate the genome analysis tools discussed above, a reconstruction of the glycolytic pathway in the archaeon *Methanococcus jannaschii* is presented. Metabolic reconstruction is one of the crucial final steps of all genome analyses and a convergence point for the data produced by different methods. Glycolysis is one of the central pathways of cellular biochemistry as it becomes obvious from a cursory exploration of the general scheme of biochemical pathways, available in the interactive form on the KEGG Web site (Fig. 8).

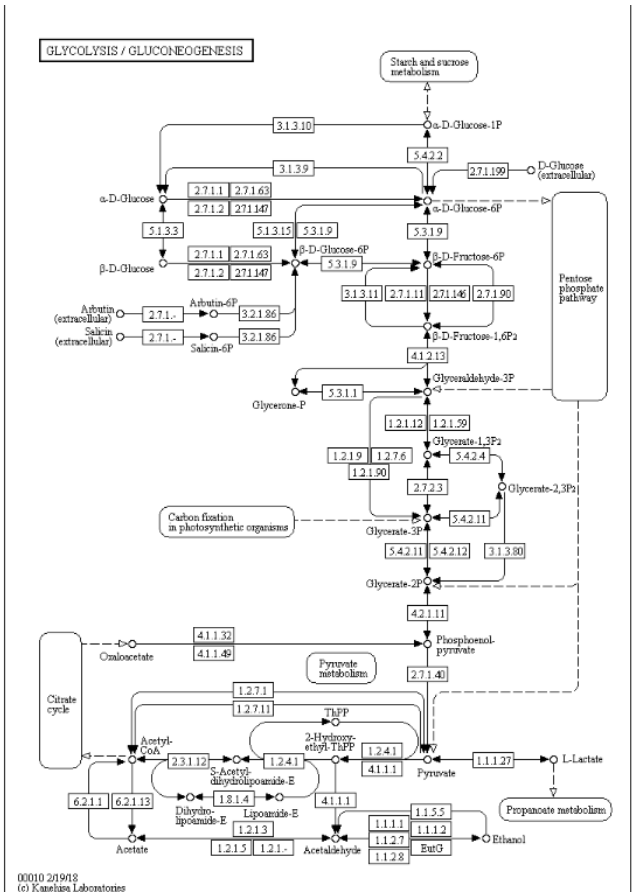


Fig. 8. Glycolysis in KEGG

The names of all the enzymes and metabolites on this map are hyperlinked and searchable. The enzyme names are hyperlinked to the enzyme information. It lists the names and catalyzed reactions, the official Enzyme Commission (EC) numbers, whether or not their protein sequences are known. Thus, clicking on the name ‘‘hexokinase” will bring up the corresponding page (Fig. 9).

ORTHOLOGY: K00844

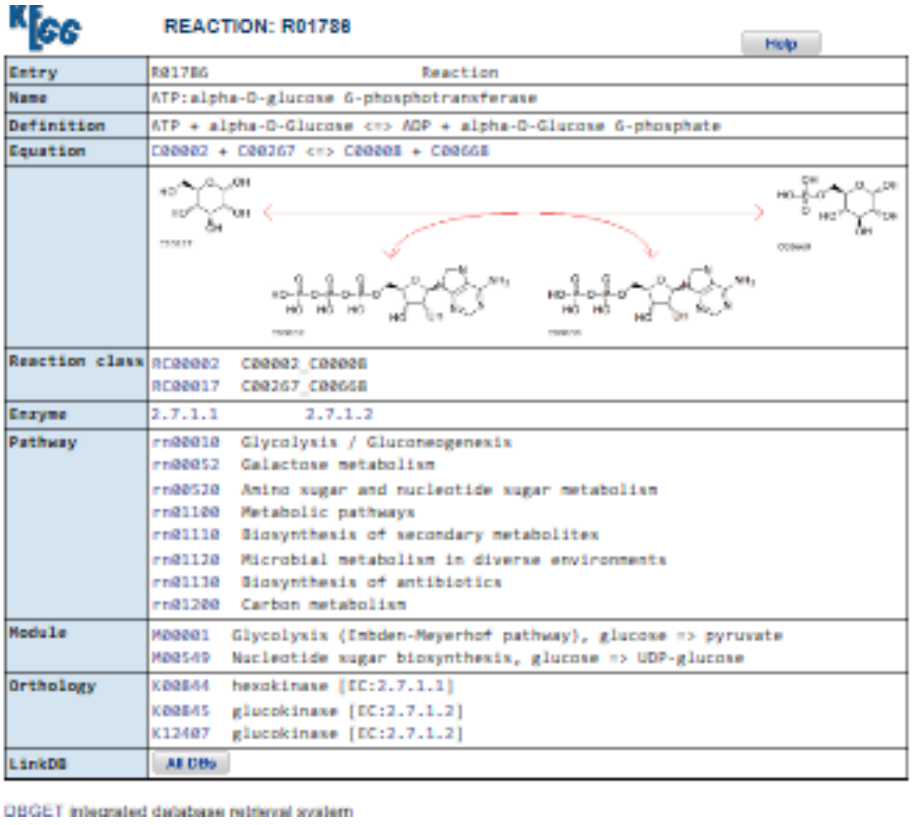
Help

Entry	K00844 KO
Name	HK
Definition	hexokinase [EC:2.7.1.1]
Pathway	ko00010 Glycolysis / Gluconeogenesis ko00051 Fructose and mannose metabolism ko00052 Galactose metabolism ko00500 Starch and sucrose metabolism ko00520 Amino sugar and nucleotide sugar metabolism ko00521 Streptomycin biosynthesis ko00524 Neomycin, kanamycin and gentamicin biosynthesis ko01100 Metabolic pathways ko01110 Biosynthesis of secondary metabolites ko01120 Microbial metabolism in diverse environments ko01130 Biosynthesis of antibiotics ko01200 Carbon metabolism ko04066 HIF-1 signaling pathway ko04910 Insulin signaling pathway ko04930 Type II diabetes mellitus ko04973 Carbohydrate digestion and absorption ko05230 Central carbon metabolism in cancer
Module	M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate M00549 Nucleotide sugar biosynthesis, glucose => UDP-glucose
Disease	H00664 Anemia due to disorders of glycolytic enzymes
Brite	KEGG Orthology (KO) [BR:ko00001] Metabolism Carbohydrate metabolism 00010 Glycolysis / Gluconeogenesis K00844 HK; hexokinase 00051 Fructose and mannose metabolism K00844 HK; hexokinase 00052 Galactose metabolism K00844 HK; hexokinase 00500 Starch and sucrose metabolism K00844 HK; hexokinase 00520 Amino sugar and nucleotide sugar metabolism K00844 HK; hexokinase Biosynthesis of other secondary metabolites 00521 Streptomycin biosynthesis K00844 HK; hexokinase 00524 Neomycin, kanamycin and gentamicin biosynthesis K00844 HK; hexokinase Environmental Information Processing Signal transduction 04066 HIF-1 signaling pathway K00844 HK; hexokinase Organismal Systems Endocrine system 04910 Insulin signaling pathway K00844 HK; hexokinase Digestive system 04973 Carbohydrate digestion and absorption K00844 HK; hexokinase Human Diseases Cancers 05230 Central carbon metabolism in cancer K00844 HK; hexokinase Endocrine and metabolic diseases 04930 Type II diabetes mellitus K00844 HK; hexokinase KEGG modules [BR:ko00002] Pathway module Carbohydrate and lipid metabolism Central carbohydrate metabolism M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate K00844 HK; hexokinase Other carbohydrate metabolism M00549 Nucleotide sugar biosynthesis, glucose => UDP-glucose K00844 HK; hexokinase Enzymes [BR:ko01000] 2. Transferases 2.7 Transferring phosphorus-containing groups 2.7.1 Phosphotransferases with an alcohol group as acceptor 2.7.1.1 hexokinase K00844 HK; hexokinase Membrane trafficking [BR:ko04131] Autophagy Mitophagy Other mitophagy associated proteins K00844 HK; hexokinase BRITE hierarchy
Other DBs	RN: R00299 R00760 R00867 R01326 R01600 R01786 R01961 R02920 GO: 0004396
Genes	HSA: 3098(HK1) 3099(HK2) 3101(HK3) 80201(HKDC1) PTR: 450504(HKDC1) 450505(HK1) 462208(HK3) 741291(HK2) PPS: 100969639(HKDC1) 100969975(HK1) 100983149(HK2) 100990081(HK3) GGO: 101125395(HK2) 101127052(HKDC1) 101131029(HK1) 101146050(HK3) PON: 100172246(HK1) 100433183(HKDC1) 100458288(HK3) 100460834(HK2) NLE: 100591323 100591401(HK3) 100593006(HK2) 100595352(HKDC1) 100595917 (HK1) MCC: 698120(HK3) 710479(HK2) 711922(HK1) 711995(HKDC1) MCF: 102121518(HK2) 102145864(HK1) 102147228(HKDC1) 107126374(HK3) CSAB: 103216074(HK1) 103216076(HKDC1) 103220012(HK2) 103245036(HK3) RRO: 104661961(HKDC1) 104661963(HK1) 104674833(HK2) 104680441(HK3) » show all Taxonomy KOALA UniProt
Reference	PMID:16233797
Authors	Kawai S, Mukai T, Mori S, Mikami B, Murata K
Title	Hypothesis: structures, evolution, and ancestor of glucose kinases in the hexokinase family.
Journal	J Biosci Bioeng 99:320-30 (2005) DOI:10.1263/jbb.99.320
LinkDB	All DBs

ENZYME: 2.7.1.1

Help

Entry	EC 2.7.1.1 Enzyme
Name	hexokinase; hexokinase type IV glucokinase; hexokinase D; hexokinase type IV; hexokinase (phosphorylating); ATP-dependent hexokinase; glucose ATP phosphotransferase
Class	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group as acceptor BRITE hierarchy
Sysname	ATP:D-hexose 6-phosphotransferase
Reaction(IUBMB)	ATP + D-hexose = ADP + D-hexose 6-phosphate [RN:R01786]
Reaction(KEGG)	R02848 > R00299 R00760 R00867 R01326 R01600 R01786 R01961 R02865 R02920; (other) R00725 R00876 R01139 R01140 R01327 R01330 R01964 R01965 R02849 R02850 R02867 R02868 Reaction
Substrate	ATP [CPD:C00002]; D-hexose [CPD:C00738]
Product	ADP [CPD:C00008]; D-hexose 6-phosphate [CPD:C02965]
Comment	D-Glucose, D-mannose, D-fructose, sorbitol and D-glucosamine can act as acceptors; ITP and dATP can act as donors. The liver isoenzyme has sometimes been called glucokinase.
History	EC 2.7.1.1 created 1961
Pathway	ec00010 Glycolysis / Gluconeogenesis ec00051 Fructose and mannose metabolism ec00052 Galactose metabolism ec00500 Starch and sucrose metabolism ec00520 Amino sugar and nucleotide sugar metabolism ec00521 Streptomycin biosynthesis ec00524 Neomycin, kanamycin and gentamicin biosynthesis ec01100 Metabolic pathways ec01110 Biosynthesis of secondary metabolites ec01120 Microbial metabolism in diverse environments ec01130 Biosynthesis of antibiotics
Orthology	K00844 hexokinase
Genes	HSA: 3098(HK1) 3099(HK2) 3101(HK3) 80201(HKDC1) PTR: 450504(HKDC1) 450505(HK1) 462208(HK3) 741291(HK2) PPS: 100969639(HKDC1) 100969975(HK1) 100983149(HK2) 100990081(HK3) GGO: 101125395(HK2) 101127052(HKDC1) 101131029(HK1) 101146050(HK3) PON: 100172246(HK1) 100433183(HKDC1) 100458288(HK3) 100460834(HK2) NLE: 100591323 100591401(HK3) 100593006(HK2) 100595352(HKDC1) 100595917 (HK1) MCC: 698120(HK3) 710479(HK2) 711922(HK1) 711995(HKDC1) MCF: 102121518(HK2) 102145864(HK1) 102147228(HKDC1) 107126374(HK3) CSAB: 103216074(HK1) 103216076(HKDC1) 103220012(HK2) 103245036(HK3) RRO: 104661961(HKDC1) 104661963(HK1) 104674833(HK2) 104680441(HK3) » show all Taxonomy
Reference	1 [PMID:16748250]
Authors	Bailey K, Webb EC.
Title	Purification of yeast hexokinase and its reaction with betabeta'-dichlorodiethyl sulphide.
Journal	Biochem J 42:60-8 (1948)
Reference	2
Authors	Berger, L., Slein, M.W., Colowick, S.P. and Cori, C.F.
Title	Isolation of hexokinase from baker's yeast.
Journal	J Gen Physiol 29:379-391 (1946)
Reference	3
Authors	Kunitz, M. and McDonald, M.R.
Title	Crystalline hexokinase (heterophosphatase). Method of isolation and properties.
Journal	J Gen Physiol 29:393-412 (1946)
Reference	4 [PMID:7048863]
Authors	Pollard-Knight D, Cornish-Bowden A.
Title	Mechanism of liver glucokinase.
Journal	Mol Cell Biochem 44:71-80 (1982)
Reference	5 [PMID:233226]
Authors	Ureta T, Radojkovic J, Lagos R, Guixe V, Nunez L.
Title	Phylogenetic and ontogenetic studies of glucose phosphorylating isozymes of vertebrates.
Journal	Arch Biol Med Exp (Santiago) 12:587-604 (1979)
Reference	6 [PMID:6477520]
Authors	Cardenas ML, Rabajille E, Niemeyer H.
Title	Fructose is a good substrate for rat liver 'glucokinase' (hexokinase D).
Journal	Biochem J 222:163-70 (1984)
Other DBs	ExplorEnz - The Enzyme Database: 2.7.1.1 IUBMB Enzyme Nomenclature: 2.7.1.1 ExPASy - ENZYME nomenclature database: 2.7.1.1 UM-BBD (Biocatalysis/Biodegradation Database): 2.7.1.1 BRENDA, the Enzyme Database: 2.7.1.1 CAS: 9001-51-8
LinkDB	All DBs



KEGG

REACTION: R01788

Help

Entry	R01786 Reaction
Name	ATP:alpha-D-glucose 6-phosphotransferase
Definition	ATP + alpha-D-Glucose <=> ADP + alpha-D-Glucose 6-phosphate
Equation	C00002 + C00267 <=> C00008 + C00668
Reaction class	RC00002 C00002_C00008 RC00017 C00267_C00668
Enzyme	2.7.1.1 2.7.1.2
Pathway	rn00010 Glycolysis / Gluconeogenesis rn00052 Galactose metabolism rn00520 Amino sugar and nucleotide sugar metabolism rn01100 Metabolic pathways rn01110 Biosynthesis of secondary metabolites rn01120 Microbial metabolism in diverse environments rn01130 Biosynthesis of antibiotics rn01200 Carbon metabolism
Module	M00001 Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate M00549 Nucleotide sugar biosynthesis, glucose => UDP-glucose
Orthology	K00844 hexokinase [EC:2.7.1.1] K00845 glucokinase [EC:2.7.1.2] K12407 glucokinase [EC:2.7.1.2]
LinkDB	All DBs

OBGET integrated database retrieval system

Fig. 9. Hexokinase information

Error Propagation and Incomplete Information in Databases

Sequence databases are predisposed to error propagation, whereby wrong annotation of one protein causes multiple errors as it is used for annotation of new genomes. Furthermore, database searches have the potential for noise amplification, so that the original annotation could have involved a minor inaccuracy or incompleteness, but its transfer on the basis of sequence similarity worsens the problem and eventually results in outright false functional assignments. These aspects of sequence databases make the common practice of assigning gene function on the basis of the annotation of the best database hit (or even a group of hits with compatible annotations) highly error-prone. Although time- and labor-consuming, the adequate genome annotation requires that each gene be considered in the context of both its phylogenetic relationships and the biology of the respective organism, hence the rather disappointing performance of automated systems for genome annotation. There are numerous reasons why functional annotation may be wrong in the first place, but two main groups of problems are due to the database search methods and to the complexity and diversity of the genomes themselves.

False Positives and False Negatives in Database Searches

It is usual in genome annotation to use a cutoff for ‘‘statistically significant’’ database hits. It can be expressed in terms of the false-positive expectation (*E*) value for the BLAST searches and is set routinely at values such as $E = 0.001$ or $E = 10^{-5}$. The problem with this approach is that the distribution of similarity scores for evolutionarily and functionally relevant sequence alignments is very broad and that a considerable fraction of them fail the *E*-value cutoff, resulting in undetected relationships and missed opportunities for functional prediction (*false negatives*). On the contrary, spurious hits may have *E*-values lower than the cutoff, resulting in false positives. The latter is most frequently caused by compositional bias (low-complexity regions) in the query sequence and in the database sequences. Clearly, there is a trade-off between *sensitivity* (false-negative rate) and *selectivity* (false-positive rate) in all database searches, and it is particularly difficult to optimize the process in genome-wide analyses. There is no simple decision to circumvent these problems. To minimize the false-positive rate, appropriate procedures for filtering low-complexity sequences are critical. Filtering using the SEG program is the default for Web-based BLAST searches, but additional filtering is justified for certain types of proteins. For example, filtering of predicted nonglobular domains using SEG with specifically adjusted parameters and filtering for coiled-coil domains using the COILS2 program is one way to minimize the false positive rate. Minimizing the false-negative rate (that is, maximizing sensitivity) is an open-ended problem. It should be kept in mind that a standard database search (e.g., using BLAST) with the protein sequences encoded in the given genome as queries is insufficient for an adequate annotation. To increase the sensitivity of genome analysis, it should be supplemented by other, more powerful methods such as screening the set of protein sequences from the given genome with preformed profile libraries.

Genome, Protein, and Organismal Context as a Source of Errors

As discussed above, protein domain architecture, genomic context and an organism’s biology may serve as sources of important, even if indirect, functional information. However, those same context features, if misinterpreted, may become one of the major sources of error and confusion in genome annotation. Standard database search programs are not equipped with the means to clearly address the implications of the multidomain organization of proteins. Therefore, unless specialized tools such as SMART or COGs are employed and/or the search output is carefully examined, assignment of the function of a single-domain protein to a multidomain homolog and vice versa becomes frequent in genome annotation. For example, mobile domains could cause chaos in the annotation process, as demonstrated, for example, by the proliferation of ‘‘IMP-dehydrogenase-related proteins’’ in several genomes. In reality, most or all of these proteins (depending on the genome) share with IMP dehydrogenase the mobile CBS domain but not the enzymatic part.

As discussed above, it is also critical for reliable genome annotation that the biological context of the given organism is taken into account. For example, it is undesirable to annotate archaeal gene products as nucleolar proteins, even if their eukaryotic homologs are correctly described as such. As a general

guide to functional annotation, it should be kept in mind that current methods for genome analysis, even the most powerful and sophisticated of them, facilitate, but do not replace the work of a human expert.

Final remarks

With an increasing number of complete genome sequences becoming available and specialized tools for genome comparison being developed, the comparative approach is becoming the most powerful strategy for genome analysis. It seems that the future should belong to databases and tools that consistently organize the genomic data according to phylogenetic, functional, or structural principles and explicitly take advantage of the diversity of genomes to increase the resolution power and robustness of the analysis. Many tasks in genome analysis can be automated, and, given the rapidly growing amount of data, automation is critical for the progress of genomics. This being said, the ultimate success of comparative genome analysis and annotation critically depends on complex decisions based on a variety of inputs, including the unique biology of each organism. Therefore, the process of genome analysis and annotation taken as a whole is, at least at this time, not automatable, and human expertise is necessary for avoiding errors and extracting the maximum possible information from the genome sequences.

References

1. Baxevanis A.D., Ouellette B. F. F. (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition, John Wiley & Son, New York
2. Elloumi M., Zomaya A. Y. (2011) Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley a& Son, New York
3. Liu L., Agren R., Bordel S., Nielsen J. (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Letters 584: 2556–2564.
4. Milne C.B., Kim P.J., Eddy J.A., Price N.D. (2009) Accomplishments in genome-scale *in silico* modeling for industrial and medical biotechnology. Biotechnol J. 4(12):1653-70
5. Pevzner P., Shamir R. (2011) Bioinformatics for Biologists, 1st Edition, Cambrage University Press
6. Ramsden J. (2015) Bioinformatics: An Introduction, Springer-Verlag, London
7. Singh G. B. (2015) Fundamentals of Bioinformatics and Computational Biology, Springer International Publishing, Switzerland

Health Bioinformatics

Zoi Georgiou
Biognosis
Volos, Greece
http://www.biognosis.gr

Maria Lambrou
Biognosis
Volos, Greece
http://www.biognosis.gr

Anastassios Hadjicrystallis
Biognosis
Volos, Greece
http://www.biognosis.gr

Contents

Health Bioinformatics

Bioinformatics includes the integration of computers, software tools, and databases in an effort to address biological questions. Bioinformatics approaches are often used for major initiatives that generate large data sets. Two important large-scale activities that use bioinformatics are genomics and proteomics. Genomics refers to the analysis of genomes. A genome can be thought of as the complete set of DNA sequences that codes for the hereditary material that is passed on from generation to generation. These DNA sequences include all of the genes (the functional and physical unit of heredity passed from parent to offspring) and transcripts (the RNA copies that are the initial step in decoding the genetic information) included within the genome. Thus, genomics refers to the sequencing and analysis of all of these genomic entities, including genes and transcripts, in an organism. Proteomics, on the other hand, refers to the analysis of the complete set of proteins or proteome. In addition to genomics and proteomics, there are many more areas of biology where bioinformatics is being applied (i.e., metabolomics, transcriptomics). Each of these important areas in bioinformatics aims to understand complex biological systems.

Many scientists today refer to the next wave in bioinformatics as systems biology, an approach to tackle new and complex biological questions. Systems biology involves the integration of genomics, proteomics, and bioinformatics information to create a whole system view of a biological entity.

For instance, how a signaling pathway works in a cell can be addressed through systems biology. The genes involved in the pathway, how they interact, and how modifications change the outcomes downstream, can all be modeled using systems biology. Any system where the information can be represented digitally offers a potential application for bioinformatics. Thus, bioinformatics can be applied from single cells to whole ecosystems. By understanding the complete "parts lists" in a genome, scientists are gaining a better understanding of complex biological systems. Understanding the interactions that occur between all of these parts in a genome or proteome represents the next level of complexity in the system. Through these approaches, bioinformatics has the potential to offer key insights into our understanding and modeling of how specific human diseases or healthy states manifest themselves.

Translational bioinformatics

Translational bioinformatics, a field in the study of health informatics that emerged after the first human genome mapping, focuses on the convergence of molecular bioinformatics, biostatistics, statistical genetics and clinical informatics. The field is evolving at a tremendously fast pace, and many related areas have been proposed. Amongst them, pharmacogenomics is a branch of genomics concerned with individuals' variations to drug response due to genetic differences. The area is important for designing precision medicine in future. Though a relatively young field, translational bioinformatics has become an important discipline in the era of personalized and precision medicine.

> According to the American Medical Informatics Association (AMIA), translational bioinformatics (TBI) is "the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data, into proactive, predictive, preventive, and participatory health".
> (http://www.amia.org/applications-informatics/translational-bioinformatics)

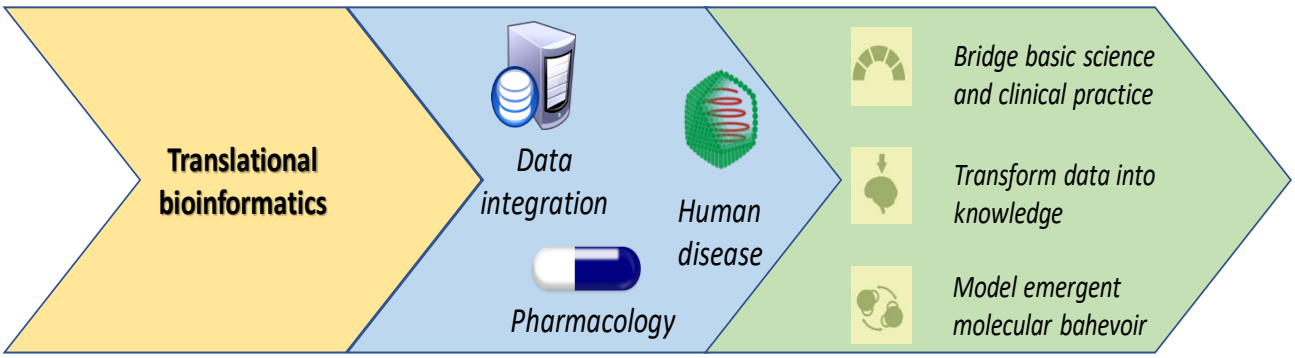


Figure 1. Translational Bioinformatics.

A 2014 review article categorized recent themes in the field of TBI into four major categorizations:

1. clinical ''big data", or the use of electronic health record (EHR) data for discovery (genomic and otherwise);
2. genomics and pharmacogenomics in routine clinical care;
3. omics for drug discovery and repurposing; and
4. personal genomic testing, including a number of ethical, legal, and social issues that arise from such services.

The importance of translational bioinformatics may be best understood in the things it is teaching us, things not previously knowable. For example, it is identifying flawed science, improving estimates of relative pathogenicity of human genetic variants, inferring new insights about underlying genetic mechanisms of disease, and identifying promising new drug indications based on curating large volumes of scientific literature. While,

sequencing an exome for a clinical diagnosis can be a routine task, the interpretation of the data to make an actual diagnosis or treatment plan is much more complex. Out of the many thousands of variants identified, many of them will have to be evaluated for their clinical utility. At times, for perhaps a simple Mendelian disorder this may be as simple, as only a single variant will need to be identified and considered. But for more complex diseases (e.g. cancers, diabetes, or neurodegenerative diseases) multiple variants will need to be identified. It is only by asking the correct questions about the patient and the disease, along with employing the right computational tools that correct answers can be achieved.

New discoveries, resulting from the Human Genome Project, are now frequently applied to develop improved diagnostics, prognostics, and therapies for complex diseases, which is known as "translational genomics". In particular, the sequencing cost per genome has markedly reduced over the last decade, according to the data presented by the National Institutes of Health (NIH) Human Genome Research Institute as shown in Figure 2. This further gives rise to new opportunities for personalized treatment and risk stratification.

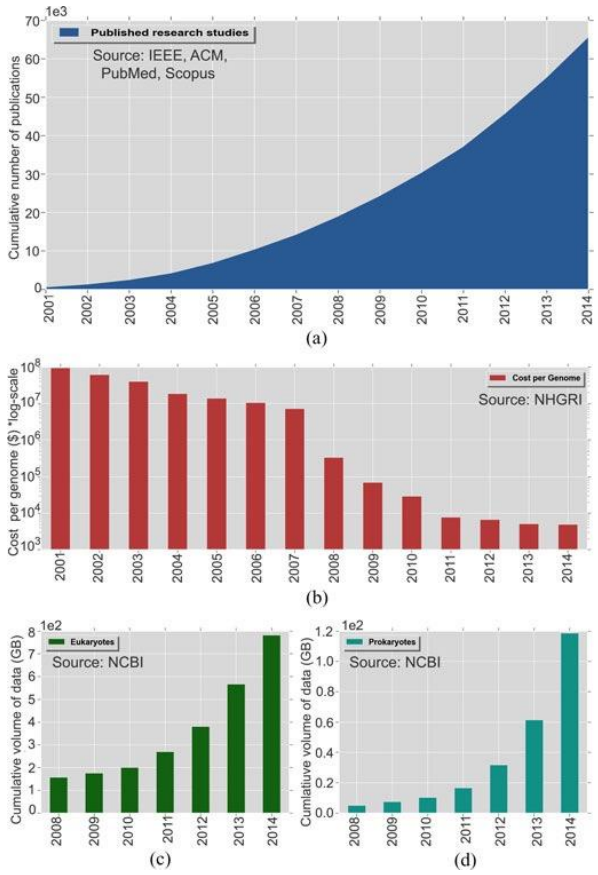


Figure 2. a) Number of research studies sequencing DNA or genomes (source: PubMed, Web of Science, Scopus, IEEE, ACM). b) Sequencing cost per human-sized genome (source: National Human Genome Research Institute, NHGRI). Total volume of genomic data per year reported by completed studies for c) eukaryotes and d) prokaryotes in 1e2 GB (source: National Center for Biotechnology Information) (Andreu-Perez, Poon, et al. 2015).

On the other hand, research in bioinformatics has broadened from solely sequencing the genome of an individual to also measuring epigenomic data (i.e., above the genome), which include processes that alter gene expression other than changes of primary DNA sequences, such as DNA methylation and histone modifications.

Information technologies for acquiring and analyzing biological molecules other than the genome, for example, transcriptome (the total mRNA in a cell or organism), proteome (the set of all expressed proteins in a cell, tissue, or organism), and metabolome (the total quantitative collection of low molecular weight compounds, metabolites, present in a cell or organism that participate in metabolic reactions) are also needed for future advances in the field. To summarize, OMICS aims at collectively characterizing and quantifying groups of biological molecules that translate into the structure, function, and dynamics of an organism. The OMICS profile of each individual should eventually be linked up with phenotypes obtained from clinical observations, medical images, and physiological signals (see Figure 3).

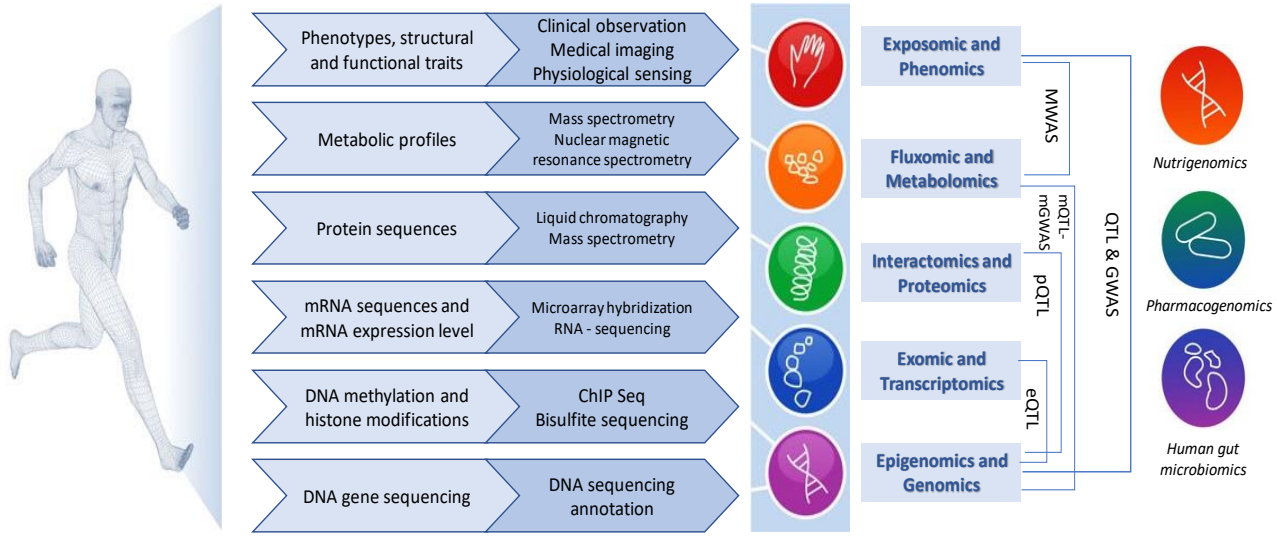


Figure 3. Outline of the "OMICS" approach for studying disease mechanisms. OMICS aims at collectively characterizing and quantifying groups of biological molecules that translate into the structure, function, and dynamics of an organism. The OMICS profile of each individual, including the genome, transcriptome, proteome, and metabolome, should be eventually linked up with phenotypes obtained from clinical observations, medical images, and physiological signals. Different acquisition technologies are required to collect data at each biological level. Interaction within each level and across different levels as well as with the environment, including nutrition, food, drugs, traditional Chinese medicine, and gut microbiome presents grand challenges in future bioinformatics research.

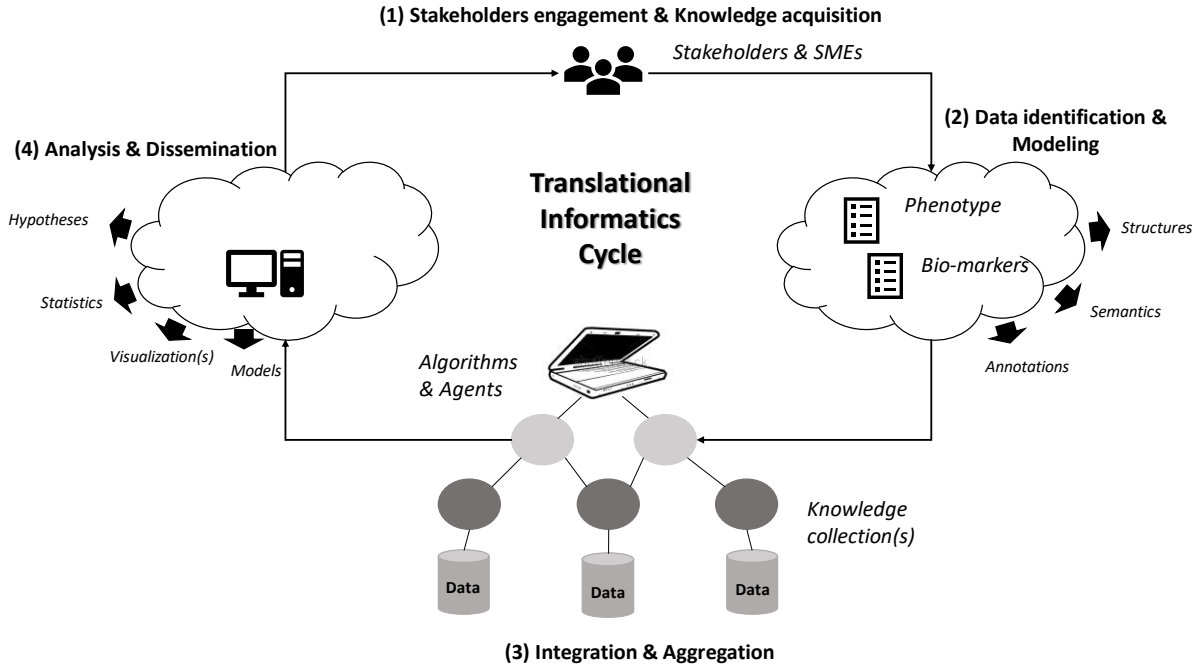


Figure 4. Practical model for the design and execution of translational informatics projects, illustrating major phases and exemplary input or output resources and data sets (Payne et al. 2009).

Genomics in clinical care (Translational Genomics)

While genetics focuses on DNA coding for single functional genes, *genomics* is the study of the entirety of our DNA, recognizing the crucial regulatory role of non-coding DNA and the complex interactions between multiple genes and the environment. Genomics is fundamental to precision medicine which, through its four components of predictive, preventive, personalized, and participatory medicine, aims to promote wellness as well as to more precisely treat disease. Currently, there is a great amount of genomic discovery research occurring that includes new genomic variants, biomarkers and other basic science discoveries. Thus, many foresee rapid advances in genetic testing and genome sequencing over the next decade, with inevitable implementation into clinical practice.

GPs will play an important role within a genomics medicine service both in supporting patients through diagnostic and treatment processes and in using knowledge of genomics for disease prevention. Also, decreasing costs and increased availability of genetic testing and genome sequencing mean many physicians will consider using these services over the next few years, with some projecting that sequencing will become fully integrated into standard medical care within 10 years.

A tumour's genomic signature may be used to make a precise diagnosis, enabling more accurate prognosis and better tailored treatment. Examples include Herceptin® (trastuzumab) in breast cancer treatment and BRAF inhibitors in malignant melanoma. Treatment can also be based on germline genomic information; PARP inhibitors are more efficacious in the treatment of ovarian cancer in individuals who carry a *BRCA* gene mutation.

Although comprehensive genotyping is still relatively recent, it has a high potential for genetic stratification in patient screening, for instance, in the case of factors arising from genotyping, such as high-risk DNA mutations, milk and gluten intolerance, and muscovisciosis. Genetics combined with phenotypic

information provided by EHR may help to provide greater insights into low penetrant alleles. For example, it is well known that mutations of fibrillin 1 (FBN1) cause MFS. Nevertheless, the etiology of the disease leads to marked clinical variability of MFS patients of the same family as well as different families. Combining genetic tests of FBN1 and a series of related genes (TGFBR1, TGFBR2, TGFB2, MYH11, MYLK1, SMAD3, and ACTA2) will help to screen out patients who are more likely to develop aortic aneurysms that lead to dissections. Further studies on these high-risk patients based on morphological images of the aorta may provide insight into the rate of disease development.

Another potential area for translational genomics is to study the gene networks of different syndromes of the same person in order to better understand how these syndromes are interrelated. For example, this has been used to study different genes on chromosome 21 (HSA21) and their role in Down's Syndrome (DS), as well as to understand the underlying reason why nearly half of DS patients exhibit an overprotection against cardiac abnormalities related to the connective tissue. One hypothesis is based on the recent evidence that there is an overall upregulation of FBN1 in DS (which is normally down regulated in MFS). The construction of genetic networks will, therefore, provide a clearer picture of how these syndromes are related. By understanding the gene networks of the related syndromes, it may be possible to provide specific gene therapy for the related diseases.

Another example took place at Stanford's Lucile Packard Children's Hospital, where a newborn presented with a condition known as long QT syndrome. In this specific case, the manifestation was unusually severe-the baby's heart stopped multiple times in the hours after its birth. Long QT syndrome can be caused by mutations in a number of different genes. It is necessary to know which gene harbors the mutation in order to know how to treat the condition. In this case, a whole-genome sequencing (WGS) was performed enabling identification of a previously-studied mutation, as well as a novel copy number variation in the TTN gene that would not otherwise have been detectable through targeted genotyping alone. Moreover, NGS enabled the answer to be obtained in a matter of hours to days instead of weeks.

Pharmacogenomics

Pharmacogenomics can be defined as the study of how genetic factors affect a person's response to drugs. This relatively new field combines pharmacology (the science of drugs) and genomics (the study of genes and their functions) to develop effective, safe medications and doses that will be tailored to a person's genetic makeup.

Many drugs that are currently available are "one size fits all," but they don't work the same way for everyone. It can be difficult to predict who will benefit from a medication, who will not respond at all, and who will experience negative side effects (called adverse drug reactions). Adverse drug reactions are a significant cause of hospitalizations and deaths. Once a patient takes a drug, the drug must travel through the body to its target(s), act on its target(s), and then leave the body. The first and last of these processes is facilitated by pharmacokinetic (PK) genes, which may affect a drug in the ''ADME'' processes: to be *absorbed* into and *distributed* through the body, *metabolized* (either to an active form or broken down into an inactive form), and *excreted*. With the knowledge gained from the Human Genome Project, researchers are learning how inherited differences in genes affect the body's response to medications. These genetic differences will be used to predict whether a medication will be effective for a particular person and to help prevent adverse drug reactions.

Pharmacogenomics focuses on the identification of genome variants that influence drug effects, typically via alterations in a drug's pharmacokinetics or via modulation of a drug's pharmacodynamics (e.g., modifying a drug's target or perturbing biological pathways that alter sensitivity to the drug's pharmacological effects). For

diseases other than cancer and infectious diseases, the genome variations of interest are primarily in the germline DNA, either inherited from parents or *de novo* germline sequence changes that alter the function of gene products. In cancer, both inherited genome variations and somatically acquired genome variants can influence response to anticancer agents.

Whole genome sequencing by NGS is important to the study of complex diseases such as cancer. It has been a long-standing problem in cancer treatment that drugs often have heterogeneous treatment responses even for the same type of cancer, and some drugs only show profound sensitivity in a small number of patients. Currently, large-scale personal genomics and pharmacogenomics datasets have been generated to uncover unique signaling patterns of individual patients and discover drugs that target these unique patterns. These include cancer cell line databases of nonspecific cancer cell types or a specific cancer cell type such as breast cancer. The Cancer Genome Atlas Project of the NIH has tested the personal genomic profiles of over 10000 individuals across over 20 types of cancer and uncovered new cancer subtypes based on those profiles. Patients with distinct genomics aberrations are believed to be responsible for the variability of drug response. Large-scale datasets as such can be used to enable drug repositioning, predict drug combinations, and delineate mechanisms of action. They are becoming an important component in drug development. It is, therefore, possible to design precision medicine for individual patients based on their genomics profiles.

Pharmacogenomics has gone beyond studying individuals' drug response based on genome characteristics (e.g., copy number variations and somatic mutations) and now incorporates additional transcriptomic and metabolic features such as gene expression, considering factors that influence the concentration of a drug reaching its targets and factors associated with the drug targets. Since the gene expression profiles of cell lines are known to vary considerably in the process of prolonged culture under different culture conditions and techniques, the use of gene expression from cell lines for prediction of drug response in the patient is currently controversial. A recent algorithm for predicting *in vivo* drug response with the patient's baseline gene expression profile achieved 60%– 80% predictive accuracy for different cases. Other research studied drug response using immunodeficient mice xenografted with human tumors, which have the advantage of potentially studying both genetic and nongenetic factors that affect cancer growth and therapy tolerance.

The field of pharmacogenomics is still in its infancy. Its use is currently quite limited, but new approaches are under study in clinical trials. In the future, pharmacogenomics will allow the development of tailored drugs to treat a wide range of health problems, including cardiovascular disease, Alzheimer disease, cancer, HIV/AIDS, and asthma.

Omics for drugs discovery and repurposing

The cost of generating new therapeutics has risen dramatically over the past 60 years, with each new drug costing about 80-fold more in 2010 than 1960 in inflation-adjusted terms. Also, much has been said about the protracted process involved in getting a drug through the FDA approval pipeline. Estimates are that the process can take on average 12 years between lead identification and FDA approval. As a result, many are investigating high-throughput and computational approaches to drug discovery and repurposing. Recent efforts have focused on the use of the omics data, especially genomics, to discover new drug targets and search for new uses for existing drugs, referred to as drug repositioning.

Pharmacogenomics can impact how the pharmaceutical industry develops drugs, as early as the drug discovery process itself (Figure 5). First, cheminformatics and pathway analysis can aid in the discovery of suitable gene targets, followed by small molecules as ‘‘leads’’ for potential drugs. Additionally, discovery of pharmacogenomic variants for the design of clinical trials can allow for safer, more successful passage of drugs

through the pharmaceutical pipeline. As mentioned previously, cheminformatics methods can be used to identify novel drug-protein interactions. While these predicted interactions can be used to discover new small molecules for therapeutic purposes, any new drug must still go through the significant regulatory hurdles of safety and efficacy testing.

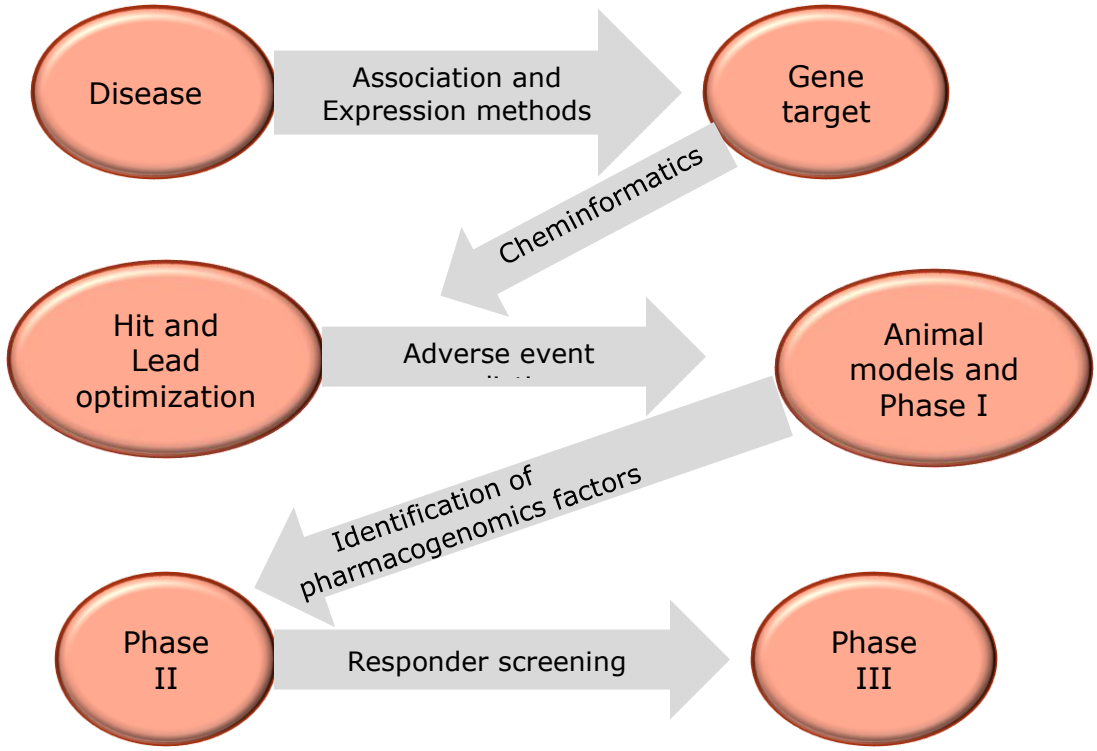


Figure 5. Drug discovery. Pharmacogenomics can be used at multiple steps along the drug discovery pipeline to minimize costs, as well as increase throughput and safety. First, association and expression methods can be used to identify potential gene targets for a given disease. Cheminformatics can then be used to narrow the number of targets to be tested biochemically, as well as identifying potential polypharmacological factors that could contribute to adverse events. After initials, pharmacogenomics can identify variants that may potentially affect dosing and efficacy. This information can then be used in designing a larger Phase III clinical trial, excluding ''non-responding'' and targeting the drug towards those more likely to respond favorably.

In addition to the Human Genome Project, several large-scale biological databases launched recently will further facilitate the study of disease mechanisms and progressions, particularly at the system level as outlined in Figure 18. The Research Collaboratory for Structural Bioinformatics Protein Data Bank is a worldwide archive of structural data of biological macromolecules, providing access to the 3-D structures of biological macromolecules, as well as integration with external biological resources, such as gene and drug databases. ProteomicsDB is another example, encompassing mass spectrometry of the human proteome acquired from human tissues, cell lines, and body fluid to facilitate the identification of organ-specific proteins and translated long intergenic noncoding RNAs, with due consideration of time-dependent expression patterns of proteins.

Parallel to these developments, the Human Metabolome Database consists of more than 40000 annotated metabolites entries in the latest version released in 2013. It provides both experimental metabolite concentration data and analyses through mass spectrometry and Nuclear Magnetic Resonance (NMR) spectrometry. Databases as such are believed to greatly facilitate the translation of information into knowledge for transforming clinical practice, particularly for metabolic-related diseases, such as diabetes and coronary artery diseases. In fact, metabolomics has emerged as an important research area that does not only include endogenous metabolites of the human body but also chemical and biochemical molecules that can interact with the human body. Specifically, ongoing efforts have been placed for fingerprinting metabolites from food and nutrition products, drugs, and traditional Chinese medicine, as well as molecules produced by the gut bacterial microbiota. These will eventually help us to better understand the interaction between the host, pathogen and environment.

The availability of the genomic, proteomic, and metabolic databases allows a better understanding of the development of complex diseases such as cancer. They also allow the search of new biomarkers using different pattern mining and clustering techniques. The clusters can be either partitional (hard) or hierarchical (tree-like nested structure). Using multicore CPU, GPU, and field-programmable gate arrays with parallel processing techniques can further accelerate these methods.

In two linked papers, Dudley et al. and Sirota et al. created disease signatures from microarray data in Gene Expression Omnibus and compared these to gene expression data from Connectivity Map to identify potentially novel therapeutics for lung cancer and inflammatory bowel disease. A similar study using this method, noted that tricyclic antidepressants might have efficacy against small cell lung cancer (but not non-small cell lung cancer).

Drug repurposing refers to taking an existing, already on the market, FDA-approved compound and using it to treat a disease or condition other than the one for which it was originally intended. In the past, inspiration for this type of ‘‘off label use” has been largely serendipitous. For example, Viagra was initially aimed at treating heart disease, and turned out to be useful for erectile dysfunction. By using a pre-approved compound, early phase clinical trials can be avoided, which can save significant time and money.

Disease-gene association data may also predict drug targets. Sanseau et al. evaluated existing GWAS hits and found that genes related to GWAS hits are significantly more likely to be targetable by small molecules or by biologic agents than other genomic regions, and that 15.6% of GWAS genes are existing drug targets (compared to 5.7% of the general genome). In support of this hypothesis, Okada et al. performed a multi-ethnic GWAS of 103,638 cases and controls for rheumatoid arthritis (RA) and noted 101 total RA risk loci; these loci identified 18 of 27 current RA drug target genes, and identified three approved cancer medications that may be active against RA. Khatri et al. analyzed eight existing organ transplant rejection datasets and found a common module of 11 genes overexpressed in all rejected organs. Using these genes, they identified two existing non-immunosuppressant drugs that could be repurposed to regulate these genes and demonstrated enhanced effect in a mouse model. Resources such as the drug-gene interaction database (DGI), which integrates data from 13 databases, and PharmGKB may facilitate translation of genomic study results to potential therapeutics. See the Table below for a listing of TBI resources.

Finally, an increasing collection of available computational and experimental methods that leverage molecular and clinical data enable diverse drug repositioning strategies. Integration of translational bioinformatics resources, statistical methods, chemoinformatics tools and experimental techniques (including medicinal chemistry techniques) can enable the rapid application of drug repositioning on an increasingly broad scale. Efficient tools are now available for systematic drug-repositioning methods using large repositories of compounds with biological activities. Medicinal chemists along with other translational researchers can play a key role in various aspects of drug repositioning.

Table 1. Public resources available for Translational Bioinformatics.

Name	URL	Comments
Pharmacogenomic Biomarkers in Drug Labels	http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm	Lists FDA-approved drugs with pharmacogenomic information in their drug labels.
PharmGKB	http://www.pharmgkb.org	PharmGKB is a curated resource about the impact of genetic variation on drug response for clinicians and researchers.
Clinical Pharmacogenetics Implementation Consortium (CPIC)	http://www.pharmgkb.org/page/cpic	Provides a list of the published guidelines for drug-gene interactions produced by CPIC.
Phenotype Knowledgebase	http://phekb.org	Online collaborative repository for building, validating, and sharing electronic phenotype algorithms and their performance characteristics.
NHGRI Catalog of GWAS studies	http://www.genome.gov/26525384	Curated list of GWAS studies, their phenotypes, and key results.
Catalog of PheWAS results	http://phewascatalog.org	Searchable, downloadable catalog of EHR PheWAS results.
Drug-Gene Interaction database	http://dgidb.genome.wustl.edu	Provides a search interface into drug-gene interactions from data derived from 13 resources.
My Cancer Genome	http://www.mycancergenome.org	Provides up-to-date data regarding cancer mutations, treatments, and relevant clinical trials.
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	It provides up-do-date relationships among human variations and phenotypes along with supporting evidence.
SHARPn	http://phenotypeportal.org	Collection of computable phenotype algorithms generated by SHARPn.

Personalized genomic testing

Personalized medicine has become important as a means to help patients receive the best possible outcomes while reducing adverse effects and high direct medical costs if a treatment will not benefit the patient.

Genetic and genomic tests each have a place in personalized medicine. Genetic tests typically focus on a specific, known gene, while genomic tests, whole-genome sequencing (WGS), focus on expression and interaction of groups of genes. Genetic tests concentrate on the presence or absence of mutations, or overexpression, of individual genes, while genomic tests provide gene signature profiles based on expression levels of specific component genes. Examples of genetic tests include BRCA-1 and -2 in breast cancer, EGFR in non-small cell lung cancer, and BRAF in melanoma. Examples of genomic tests include the Oncotype DX assays in breast, colon, and prostate cancers, and the 70-gene assay in breast cancer. Since WGS was first developed, advances in technology have made the test easier, quicker, and less expensive. So easy, in fact, that it could become a routine test offered to healthy patients during primary care visits. However, it can be difficult to determine what the results of WGS mean.

What is genetic testing?

Genetic testing is the analysis of human DNA, RNA, or proteins to detect gene variants, changes in chromosomes, or proteins associated with certain diseases or conditions; non-diagnostic uses include paternity testing and forensics. The results of a genetic test can confirm or rule out a suspected genetic condition or help determine a person's chance of developing or passing on a genetic disorder. More than 1,000 genetic tests are currently in use, and more are being developed.

Genetic testing methodology varies:

- *Molecular genetic tests* study single genes or short lengths of DNA to identify variations or mutations that lead to a genetic disorder.
- *Chromosomal genetic tests* analyze whole chromosomes or long lengths of DNA to detect large genetic changes such as an extra copy of a chromosome.
- Finally, *biochemical genetic tests* study the amount or activity level of proteins; abnormalities in either can indicate changes to the DNA that result in a genetic disorder.

The Figure 6 summarizes the various applications of genetic testing available today. Genetic testing is voluntary, and it has benefits as well as limitations and risks. Thus, the decision about whether to be tested is a personal and complex one. A geneticist or genetic counselor can help by providing information about the pros and cons of the test and discussing the social and emotional aspects of testing.

The last decade has seen an unprecedented pace of advancement in our ability to sequence the genome. As the cost of sequencing decreases, the opportunity to move from targeted sequencing to whole exome sequencing (the analysis of all a person's genes) and then to whole genome sequencing that analyzes a person's entire genetic code becomes more accessible, particularly for researcher.

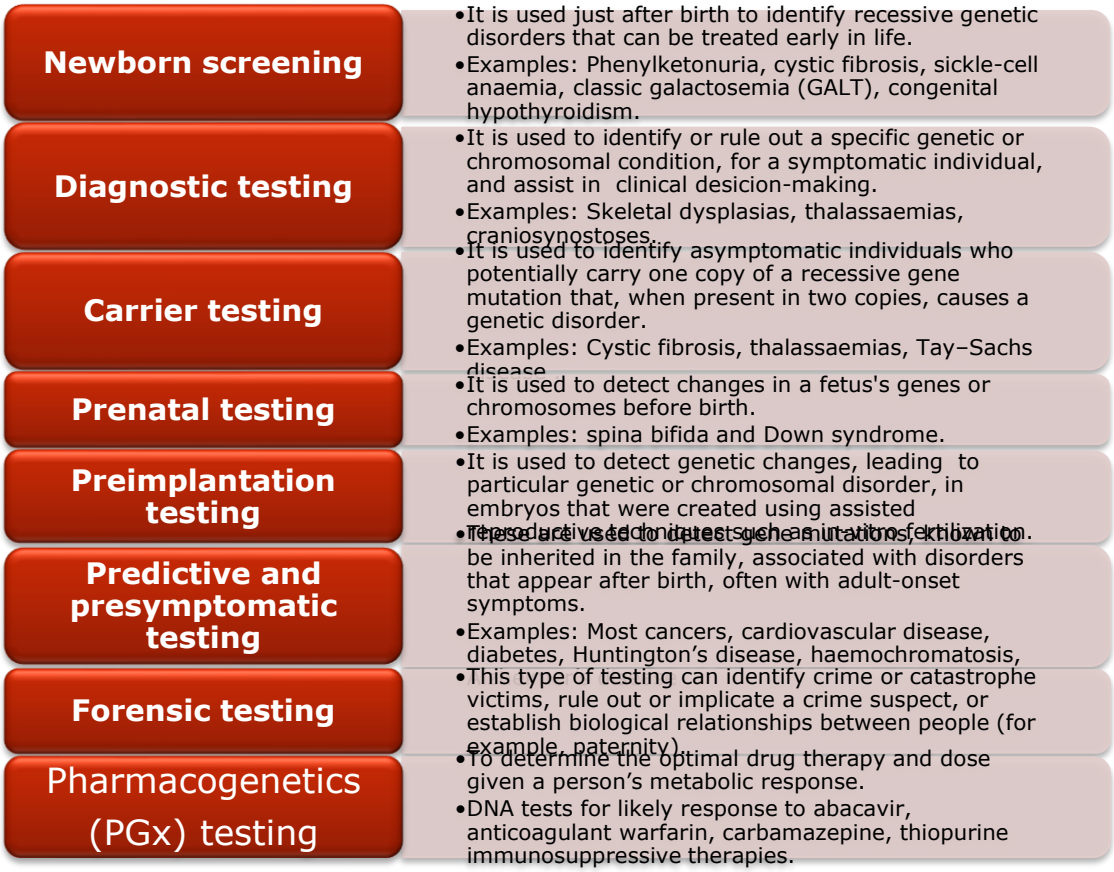


Figure 6. Available types of genetic testing.

Most medical genetic test results will directly change your medical care and those changes are based on evidence gathered through clinical trials and other medical practice. Medical genetic tests may be used to:

- *Diagnose a genetic disease.*

Example: Finding changes, called mutations, in a single gene can diagnose such genetic disorders as familial hypercholesterolemia, muscular dystrophy, Huntington disease, and other single gene diseases.

- *Assess the chance of having a child with certain genetic conditions.*

Example: Some genetic conditions are particularly common in people whose ancestors come from specific areas of the world. People who carry these genetic conditions usually have no family history and no way to know that they carry a gene that could cause a genetic condition in their children – like cystic fibrosis, Tay-Sachs disease, or sickle cell anemia.

- *Predict if a person may be more likely to have side effects or an abnormal response to a certain drug.*

Example: Variations in some genes that direct drug metabolism can cause people to metabolize, or process, certain drugs faster or slower than usual. Knowing about these variations may help your doctor avoid drugs that may be problematic for you or choose the safest, most effective dose of a drug. Examples of drugs for which genetic testing is in the early stages of usage are blood thinners, psychiatric drugs, and certain types of cancer chemotherapies.

- *Find an increased risk for a common disease.*

Example: Some people have a very high risk of a common disease like breast, ovarian, or colon cancer – often at an earlier age than usual – because of a mutation in a single gene. The actual risk may depend on the disease and the gene mutation. Knowing about this very high risk increases the chance that the disease can either be prevented or caught early when the treatment options are best.

For genomic assays to be a viable tool, they must be accurate and clinically meaningful. As below Table shows, genomic assays need to have analytic validity, clinical validity, and clinical utility. The analytic validity is the test's ability to accurately and reliably measure the genotype (or analyte) of interest in the clinical laboratory and in specimens representative of the population of interest. Regarding clinical validation, a major goal is to identify and quantify potential sources of biologic variation in the analysis of a given sample. Clinical utility is a test's ability to benefit patients by improving treatment decisions.

Table 2. Evidence Requirements for Genomic Assays:

– *Analytical validity*: Ability to accurately and reproducibly measure analyte (or genotype). Does it detect what it is supposed to detect?
– *Clinical validity*: Ability to accurately and reliably predict phenotype, clinical disease, or predisposition to disease. Does it detect information that is known to be associated with a specific disease?
– *Clinical utility*: Evidence that guides patient management and affects decision making, resulting in added value and improved outcomes. How useful is the information to improve health outcomes?

The rapid evolution of genomic sequencing technologies has decreased the cost of genetic analysis to the extent that it seems plausible that genome-scale sequencing could have widespread availability in health care across all stages of life - from preconception to adult medicine (Figure 7). Challenges to fully embracing genomics in a clinical setting remain, but some approaches are starting to overcome these barriers, such as community-driven data sharing to improve the accuracy and efficiency of applying genomics to patient care.

Early analyses comparing genomes of different individuals confirmed the remarkable similarities of sequence (99% identical), but soon gave way to expectations that the millions of nucleotide differences among different individuals would enable clinicians to not only recognize each individual's biologic uniqueness, but to translate this knowledge into more precise understanding of physiology, more refined diagnoses, better disease risk assessment, earlier detection and monitoring, and tailored treatments to the individual patient; ie, personalized (or individualized or precision) medicine.

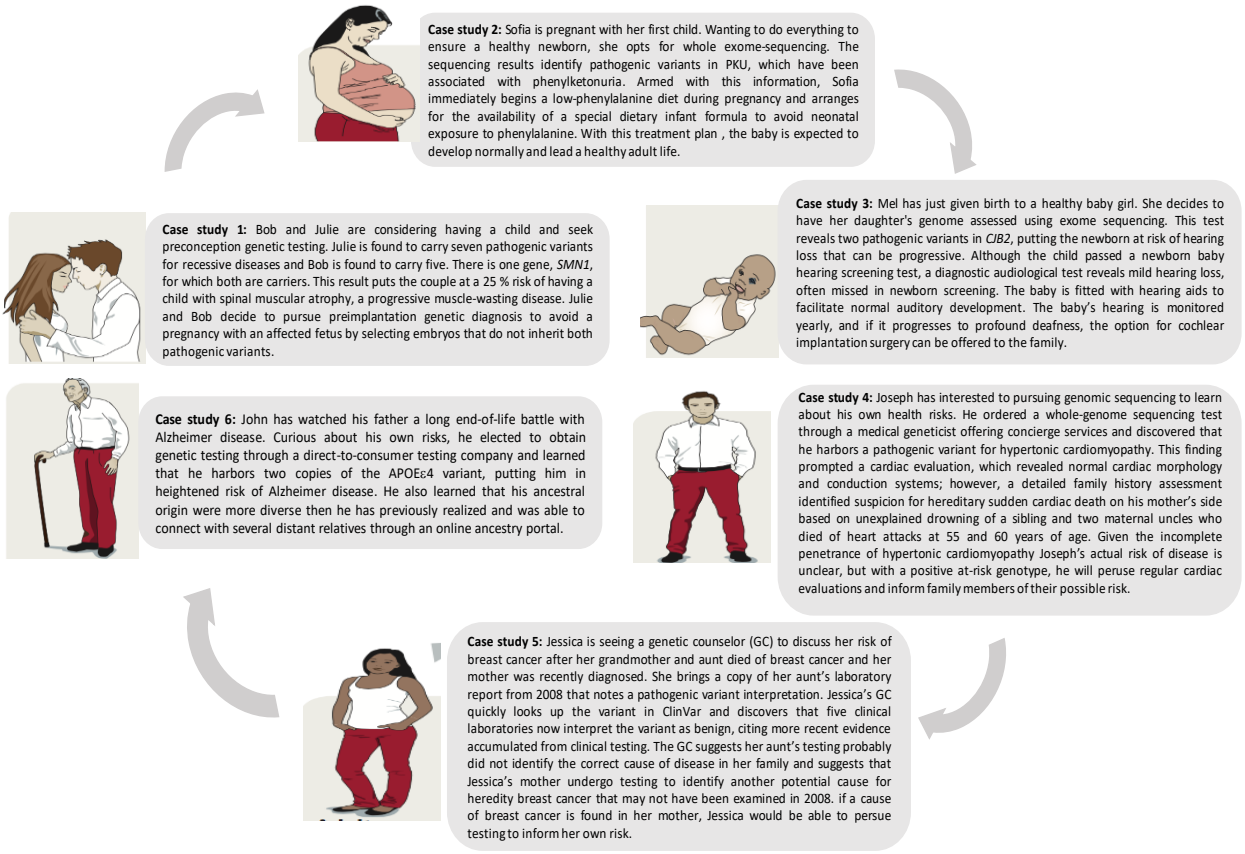


Figure 7. The use of genomics throughout an individual's lifespan. Case studies of the use of genomics to inform patient care at different stages of life. (Rehm 2017)

Value of genomics in personalized medicine

Despite the use of DNA diagnostic testing prior to 2000, it has been the exponential increase in our capacity to perform nucleotide sequencing that has been largely responsible for the relatively recent emphasis on personalized medicine. Completion of the HapMap project allowed for selection of genome wide single nucleotide variants (SNVs) that would tag common variants throughout the genome. This enabled genome-wide association studies (GWASs) for discovery of loci associated with clinical phenotypes. Advances in next-generation sequencing (NGS) have reduced the cost and time required for whole exome sequencing (WES) or whole genome sequencing (WGS), and we are continually improving our capacity for handling the storage, transfer, and analyses of huge amounts of sequence data. Also, have enabled millions of people to have their individual genomic sequence analysed, primarily within the settings of research studies or clinical care. There is

widespread recognition that access to an individual's genomic sequence and other 'omics' data can enable a more detailed understanding of our health and disease risks, and inform a more precise approach to patient care, a strategy now commonly called 'precision medicine'.

With genomic data now increasingly used to guide the individual care of patients, our health care systems are evolving, although several challenges remain. This Perspective considers how genomics is guiding health care for the individual, providing illustrative examples of how individuals are taking advantage of personal genomic information, ranging from advanced diagnostics and tumor profiling to genomic risk assessments. These examples are then interweaved with the day-to-day challenges still facing the integration of genomics into clinical practice as well as with strategies that are being developed to overcome these barriers and enable genomics to be a part of ever more aspects of everyday patient care.

Trends in Personal Genomic Testing to Guide Health Care

In 2008 saw the founding of several companies that offered direct-to-consumer (DTC) genetic testing, reporting on a variety of genes for both health and recreational purposes. Direct-To-Consumer (DTC) genetic testing through sites such as 23andMe (Mountain View, CA) has provided an avenue for patients to pursue genetic testing outside of a doctor's order. Individuals received test results and personalized information on their genetic ancestry, disease risk, and drug response for selected medications.

DTC genetic testing raises a number of interesting ethical, legal, and social issues. For several years, there was an open question as to whether or not these tests should be subject to government regulation. In November 2013, the US FDA ordered 23andMe to stop advertising and offering their health-related information services. The FDA considered these tests to be ''medical devices" and as such to require formal testing and FDA approval for each test. In February 2015, it was announced that the FDA had approved 23andMe's application for a test for Bloom syndrome (http://www.fda.gov/News Events/Newsroom/PressAnnouncements/UCM435003), and in October 2015 it was announced that the company would once again be offering health information in the form of carrier status for 36 genes. Note that a 23andMe customer is able to download his or her raw genomic data and to use information from other websites to interpret the results, including Promethease, Geneticgenie, openSNP, and Interpretome for health-related associations.

A more positive example of where genetic testing is helping patients is a case presented at the American Neurological Association conference in 2014. A patient had a history of Alzheimer's disease on her mother's side of the family. She did not know if she was a carrier, nor did she want to know. But she wanted to ensure that she did not pass that mutation to her future children. Preimplantation genetic diagnosis (PGD) testing enabled her doctors to select embryos that did not have that Alzheimer's disease gene mutation. The patient herself was never tested, nor was she informed how many (if any) of the embryos contained the mutation.

Table 3. Examples of personal genetic profiling tests for disease susceptibility.

Company	Example product	Details
23andMe	Health Edition	"Find out if you carry inheritable markers for diseases such as breast cancer, cystic fibrosis, and Tay-Sachs...Learn your genetic risk for type 2 diabetes, Parkinson's disease, and other conditions.
deCODEme	Complete Scan	"Calculate your genetic risk for 51 conditions..."
Genetic Health	Premium Male	"These are our most comprehensive test and includes all the other tests in our range... Evaluates the risk of prostate cancer as well as the risk for thrombosis, osteoporosis, metabolic imbalances of detoxification and chronic inflammation. It also evaluates the risk profile of the most common cardiovascular diseases..."
Graceful Earth	Alzheimer's genome test	"Check your future susceptibility BEFORE symptoms occur... Pre-emptive insight into one's genetic predisposition can empower and allow for pro-active prevention."
Navigenics	Health Compass	"Knowing your genetic predispositions for important health conditions and medication reactions can help motivate you to take steps towards a healthier life. By gaining insight into these risks, you can plan for what's important."

Also, Universal newborn screening (NBS) is an extraordinarily successful public health program, preventing morbidity and mortality through early diagnosis and management of conditions including rare inborn errors of metabolism. Conditions such as phenylketonuria are not clinically evident at birth but lead to significant irreversible harm or death if not treated promptly. NBS has saved countless lives and vastly improved the quality of children's lives by allowing timely therapeutic interventions, and technological advances such as the use of tandem mass spectrometry (MS/MS) have played a significant role in expansion of NBS. The capacity of genome-scale sequencing for disease gene discovery is increasingly being applied as a diagnostic test in children with suspected monogenic disorders.

The ability to analyze many or all genes in the genome simultaneously provides new opportunities for genomic medicine. The clinical utility of sequencing is recognized for certain diseases and in an increasing number of medical specialties, with genetic and genomic medicine offering the promise of improved diagnostics and treatments – and patients asking physicians about the applicability of these technologies for their own care. However, some experts caution the roadmap for translating genetics and genomics into routine clinical practice is unclear.

Computational health informatics

Computational health informatics (CHI) is an emerging research topic within and beyond the medical industry. It is a multidisciplinary field involving various sciences such as biomedical, medical, nursing, information technology, computer science, and statistics. CHI is a computer science branch that addresses how computational methods relate to providing health care. Using Information and Communication Technologies (ICTs), health informatics collects and analyzes the information from all healthcare domains to predict patients' health status. The major goal of health informatics research is to improve the quality of care provided to patients or Health Care Output (HCO). The healthcare industry has experienced rapid growth of medical and healthcare data in recent years. Figure 8 depicts the growth of both healthcare data and digital healthcare data. It is projected that the healthcare data analytics market will increase and grow 8-10 times as fast as the overall economy until 2017.

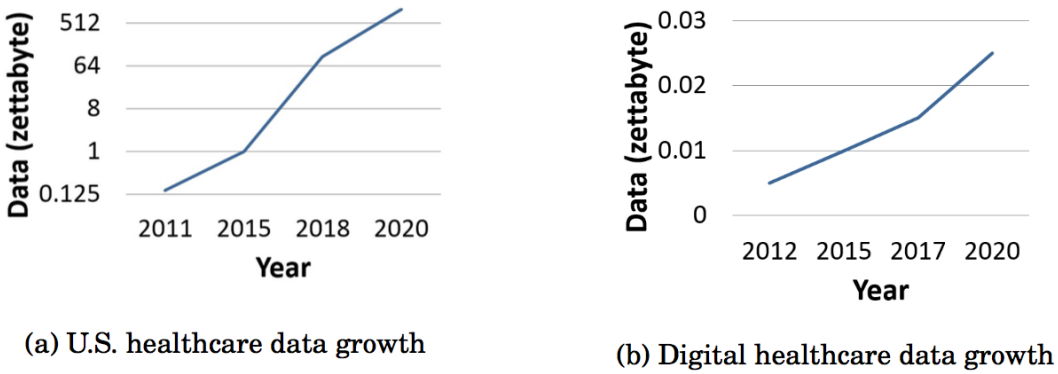


Figure 8. Healthcare data growth. (Fang et al. 2016)

The rapid growth of new technologies has led to a significant increase of digital health data in recent years. More medical discoveries and new technologies such as novel sensors, mobile apps, capturing devices, wearable technology have contributed to additional data sources. Therefore, the healthcare industry produces a huge amount of digital data by utilizing information from all sources of healthcare data such as Electronic Health Records (EHR, including electronic medical records) and personal health records (PHR, one subset of EHR including medical history, laboratory results, and medications). Based on reports, digital healthcare data from all over the world was estimated to be equal to 500 petabytes (1015) in 2012 and it is expected to reach 25 exabytes in 2020 as shown in Figure 23b.

The digital health data is not only enormous in amount, but also complex in its structure for traditional software and hardware. Some of the contributing factors to the failure of traditional systems in handling these datasets include:

- The vast variety of structured and unstructured data such as medical records, hand-written doctor notes, medical diagnostic images (MRI, CT), and radiographic films.
- Existence of noisy, heterogeneous, complex, diverse, longitudinal, and large datasets in healthcare informatics.
- Difficulties to capture, store, analyze and visualize such large and complex datasets.
- Necessity of increasing the storage capacity, computation power and the processing power.

– Improving the quality of care, security of patients' data, sharing, and the reduction of the healthcare cost.

Hence, solutions are needed in order to manage and analyze such complex, diverse and huge datasets in a reasonable time complexity and storage capacity. Big data analytics, a popular term given to datasets which are large and complex, play a vital role in managing the huge healthcare data and improving the quality of healthcare offered to patients. In addition, it promises a bright prospect for decreasing the cost of care, improving treatments, reaching more personalized medicine, and helping doctors and physicians to make personalized decisions.

Finally, the major benefits of big data analytics in healthcare are as follow:

1. It makes use of the huge volume of data and provides timely and effective treatment to patients.
2. It provides personalized care to patients.
3. It will benefit all the components of a medical system (i.e., provider, payer, patient, and management).

References

Altman, R.B., 2012. Translational Bioinformatics: Linking the Molecular World to the Clinical World. *Clinical Pharmacology & Therapeutics*, 91(6), pp.994–1000. Available at: http://doi.wiley.com/10.1038/clpt.2012.49.

Andreu-Perez, J., Poon, C.C.Y., et al., 2015. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4), pp.1193–208. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7154395.

Andreu-Perez, J., Leff, D.R., et al., 2015. From Wearable Sensors to Smart Implants--Toward Pervasive and Personalized Healthcare. *IEEE transactions on bio-medical engineering*, 62(12), pp.2750–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25879838.

Anhøj, J., 2003. Generic Design of Web-Based Clinical Databases. *Journal of Medical Internet Research*, 5(4), p.e27. Available at: http://www.jmir.org/2003/4/e27/.

Aronson, S.J. & Rehm, H.L., 2015. Building the foundation for genomics in precision medicine. *Nature*, 526(7573), pp.336–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26469044.

Bain, J.R. et al., 2009. Metabolomics applied to diabetes research: moving from information to knowledge. *Diabetes*, 58(11), pp.2429–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19875619.

Ban, T.A., 2006. The role of serendipity in drug discovery. *Dialogues in clinical neuroscience*, 8(3), pp.335–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17117615.

Baro, E. et al., 2015. Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Research International*, 2015(1), pp.1–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6137488.

Barretina, J. et al., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391), pp.603–7. Available at: http://www.nature.com/nature/journal/v483/n7391/full/nature11003.html%3FWT.ec_id%3DNATURE-

20120329.

Baskar, S. & Aziz, P.F., 2015. Genotype-phenotype correlation in long QT syndrome. *Global Cardiology Science and Practice*, 2015(2), p.26. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4614326&tool=pmcentrez&rendertype=abstract.

Bates, D.W. et al., 2014. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs*, 33(7), pp.1123–1131. Available at: http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.2014.0041.

Benson, G., 2015. Editorial: Nucleic Acids Research annual Web Server Issue in 2015. *Nucleic Acids Research*, 43(Web Server issue), pp.W1–W2. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv581.

Berg, J.S. et al., 2017. Newborn Sequencing in Genomic Medicine and Public Health. *Pediatrics*, 139(2). Available at: http://www.ncbi.nlm.nih.gov/pubmed/28096516.

Bhatia, D., 2015. *Medical Informatics* PHI Learni., PHI Learnign Private Limited, Delhi: PHI Learnign Private Limited, Delhi.

Boland, M.R. et al., 2013. Discovering medical conditions associated with periodontitis using linked electronic health records. *Journal of clinical periodontology*, 40(5), pp.474–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23495669.

Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23000897.

Cars, T. et al., 2013. Extraction of electronic health record data in a hospital setting: comparison of automatic and semi-automatic methods using anti-TNF therapy as model. *Basic & clinical pharmacology & toxicology*, 112(6), pp.392–400. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23374887.

Chen, J. et al., 2013. Translational Biomedical Informatics in the Cloud: Present and Future. *BioMed Research International*, 2013, pp.1–8. Available at: http://www.hindawi.com/journals/bmri/2013/658925/.

Choi, I.Y. et al., 2013. Perspectives on clinical informatics: integrating large-scale clinical, genomic, and health information for clinical care. *Genomics & informatics*, 11(4), pp.186–90. Available at: http://dx.doi.org/10.5808/GI.2013.11.4.186.

Christakis, N.A. & Fowler, J.H., 2008. The collective dynamics of smoking in a large social network. *The New England journal of medicine*, 358(21), pp.2249–58. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18499567.

Christensen, K.D. et al., 2016. Are physicians prepared for whole genome sequencing? a qualitative analysis. *Clinical genetics*, 89(2), pp.228–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26080898.

Collins, F.S. & Varmus, H., 2015. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9), pp.793–795. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:New+engla+nd+journal#0.

Collymore, D.C. et al., 2016. Genomic testing in oncology to improve clinical outcomes while optimizing utilization: the evolution of diagnostic testing. *The American journal of managed care*, 22(2 Suppl), pp.s20-5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26978033.

Costello, J.C. et al., 2014. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12), pp.1202–12. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/24880487.

Danciu, I. et al., 2014. Secondary use of clinical data: the Vanderbilt approach. *Journal of biomedical informatics*, 52(1), pp.28–35. Available at: http://dx.doi.org/10.1016/j.jbi.2014.02.003.

Delaney, S.K. et al., 2016. Toward clinical genomics in everyday medicine: perspectives and recommendations. *Expert review of molecular diagnostics*, 16(5), pp.521–32. Available at: https://www.tandfonline.com/doi/full/10.1586/14737159.2016.1146593.

Denny, J.C., 2014. Surveying Recent Themes in Translational Bioinformatics: Big Data in EHRs, Omics for Drugs, and Personal Genomics. *IMIA Yearbook*, 9(1), pp.199–205. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25123743.

Dudley, J.T. et al., 2011. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96), p.96ra76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21849664.

Dugas, M., 2015. Clinical Research Informatics: Recent Advances and Future Directions. *Yearbook of medical informatics*, 10(1), pp.174–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26293865%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4587057.

Eichstaedt, J.C. et al., 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2), pp.159–69. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25605707.

Embi, P.J., 2013. Clinical research informatics: survey of recent advances and trends in a maturing field. *Yearbook of medical informatics*, 8, pp.178–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23974569.

Embi, P.J. & Payne, P.R.O., 2009. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *Journal of the American Medical Informatics Association : JAMIA*, 16(3), pp.316–27. Available at: http://dx.doi.org/10.1197/jamia.M3005.

Eriksson, R. et al., 2014. Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population. *Drug safety*, 37(4), pp.237–47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24634163.

Fang, R. et al., 2016. Computational Health Informatics in the Big Data Age. *ACM Computing Surveys*, 49(1), pp.1–36. Available at: http://dl.acm.org/citation.cfm?doid=2911992.2932707.

Fernández-Suárez, X.M. & Galperin, M.Y., 2013. The 2013 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 41(D1), pp.1–7.

Forrest, G.N. et al., 2014. Use of electronic health records and clinical decision support systems for antimicrobial stewardship. *Clinical infectious diseases*, 59(Suppl 3), pp.S122-33. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25261539.

Friedl, M.A. et al., 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114(1), pp.168–182. Available at: http://dx.doi.org/10.1016/j.rse.2009.08.016.

Friedman, C. et al., 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*, 11(5), pp.392–402. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M1552.

Friedman, C.P., 2009. A “fundamental theorem” of biomedical informatics. *Journal of the American*

Medical Informatics Association : JAMIA, 16(2), pp.169–70. Available at: http://dx.doi.org/10.1197/jamia.M3092.

Galperin, M.Y. & Fernandez-Suarez, X.M., 2012. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 40(D1), pp.D1–D8. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1297.

Gao, W. et al., 2016. Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature*, 529(7587), pp.509–514. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26819044.

Garnett, M.J. et al., 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391), pp.570–5. Available at: http://www.nature.com/doifinder/10.1038/nature11005.

Geeleher, P., Cox, N.J. & Huang, R., 2014. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, 15(3), p.R47. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r47.

Griffith, M. et al., 2013. DGIdb: mining the druggable genome. *Nature methods*, 10(12), pp.1209–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24122041.

Hagar, Y. et al., 2014. Survival analysis with electronic health record data: Experiments with chronic kidney disease. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(5), pp.385–403. Available at: http://doi.wiley.com/10.1002/sam.11236.

Haghi, M., Thurow, K. & Stoll, R., 2017. Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices. *Healthcare Informatics Research*, 23(1), p.4. Available at: https://synapse.koreamed.org/DOIx.php?id=10.4258/hir.2017.23.1.4.

Hall, M.J. et al., 2015. Understanding patient and provider perceptions and expectations of genomic medicine. *Journal of Surgical Oncology*, 111(1), pp.9–17. Available at: http://doi.wiley.com/10.1002/jso.23712.

den Hartog, A.W. et al., 2015. The risk for type B aortic dissection in Marfan syndrome. *Journal of the American College of Cardiology*, 65(3), pp.246–54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25614422.

Hayes, D.F., Khoury, M.J. & Ransohoff, D., 2012. Why Hasn't Genomic Testing Changed the Landscape in Clinical Oncology? *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Meeting*, 1, pp.e52-5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24451831.

Hayward, J. et al., 2017. Genomics in routine clinical care: what does this mean for primary care? *British Journal of General Practice*, 67(655), pp.58–59. Available at: http://bjgp.org/lookup/doi/10.3399/bjgp17X688945.

He, K.Y., Ge, D. & He, M.M., 2017. Big Data Analytics for Genomic Medicine. *International journal of molecular sciences*, 18(2), p.412. Available at: http://www.mdpi.com/1422-0067/18/2/412.

Herland, M., Khoshgoftaar, T.M. & Wald, R., 2014. A review of data mining using big data in health informatics. *Journal Of Big Data*, 1(1), p.2. Available at: http://www.journalofbigdata.com/content/1/1/2.

Hersh, W., 2009. A stimulus to define informatics and health information technology. *BMC medical informatics and decision making*, 9(1), p.24. Available at: http://www.biomedcentral.com/1472-6947/9/24.

Hijmans, R.J. et al., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), pp.1965–1978. Available at: http://doi.wiley.com/10.1002/joc.1276.

Hoyt, R.E., Sutton, M. & Yoshihashi, A., 2009. *Medical Informatics Practical Guide for the Healthcare Professional*,

Huser, V. & Cimino, J.J., 2013. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2013(1), pp.648–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24551366%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3900207.

Hutchinson, S. et al., 2003. Allelic variation in normal human FBN1 expression in a family with Marfan syndrome: a potential modifier of phenotype? *Human molecular genetics*, 12(18), pp.2269–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12915484.

Iyer, G. et al., 2012. Genome sequencing identifies a basis for everolimus sensitivity. *Science (New York, N.Y.)*, 338(6104), p.221. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22923433.

Jennings, L. et al., 2009. Recommended principles and practices for validating clinical molecular pathology tests. *Archives of pathology & laboratory medicine*, 133(5), pp.743–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19415949.

Jensen, P.B., Jensen, L.J. & Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. *Nature reviews. Genetics*, 13(6), pp.395–405. Available at: http://www.nature.com/doifinder/10.1038/nrg3208.

Kahn, M.G. et al., 2012. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical care*, 50 Suppl(0), pp.S21-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22692254.

Kahn, M.G. & Weng, C., 2012. Clinical research informatics: a conceptual perspective. *Journal of the American Medical Informatics Association*, 19(e1), pp.e36–e42. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84863552740&partnerID=tZOtx3y1.

Kamesh, D.B.K., Neelima, V. & Ramya Priya, R., 2015. A review of data mining using big data in health informatics. *International Journal of Scientific and Research Publications*, 5(3), pp.1–7. Available at: http://www.ijsrp.org/research-paper-0315/ijsrp-p3913.pdf.

Karczewski, K.J., Daneshjou, R. & Altman, R.B., 2012. Chapter 7: Pharmacogenomics F. Lewitter & M. Kann, eds. *PLoS Computational Biology*, 8(12), p.e1002817. Available at: http://dx.plos.org/10.1371/journal.pcbi.1002817.

Khatri, P. et al., 2013. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. *The Journal of experimental medicine*, 210(11), pp.2205–21. Available at: http://www.jem.org/lookup/doi/10.1084/jem.20122709.

Kouskoumvekaki, I., Shublaq, N. & Brunak, S., 2014. Facilitating the use of large-scale biological data and tools in the era of translational bioinformatics. *Briefings in bioinformatics*, 15(6), pp.942–52. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt055.

Kovats, R.S. & Hajat, S., 2008. Heat stress and public health: a critical review. *Annual review of public health*, 29(1), pp.41–55. Available at: http://www.annualreviews.org/doi/10.1146/annurev.publhealth.29.020907.090843.

Kreso, A. et al., 2013. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science (New York, N.Y.)*, 339(6119), pp.543–8. Available at:

http://www.sciencemag.org/cgi/doi/10.1126/science.1227670.

Ku Jena, R. et al., 2009. Soft Computing Methodologies in Bioinformatics. *European Journal of Scientific Research*, 26(2), pp.189–203.

Laakko, T. et al., 2008. Mobile health and wellness application framework. *Methods of information in medicine*, 47(3), pp.217–22. Available at: http://www.schattauer.de/index.php?id=1214&doi=10.3414/ME9113.

Lamb, J. et al., 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)*, 313(5795), pp.1929–35. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1132939.

Larsen, M.E. et al., 2015. We Feel: Mapping Emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp.1246–1252. Available at: http://ieeexplore.ieee.org/document/7042256/.

Lee, J., Kuo, Y.-F. & Goodwin, J.S., 2013. The effect of electronic medical record adoption on outcomes in US hospitals. *BMC health services research*, 13(1), p.39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23375071%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3568047.

Lobitz, B. et al., 2000. Climate and infectious disease: use of remote sensing for detection of Vibrio cholerae by indirect measurement. *Proceedings of the National Academy of Sciences of the United States of America*, 97(4), pp.1438–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10677480.

Londin, E.R. & Barash, C.I., 2015. What is translational bioinformatics? *Applied & Translational Genomics*, 6, pp.1–2. Available at: http://linkinghub.elsevier.com/retrieve/pii/S2212066115000174.

Luber, G. & McGeehin, M., 2008. Climate change and extreme heat events. *American journal of preventive medicine*, 35(5), pp.429–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18929969.

Lussier, Y.A. & Liu, Y., 2007. Computational approaches to phenotyping: high-throughput phenomics. *Proceedings of the American Thoracic Society*, 4(1), pp.18–25. Available at: http://pats.atsjournals.org/cgi/doi/10.1513/pats.200607-142JG.

MacKenzie, S.L. et al., 2012. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *Journal of the American Medical Informatics Association*, 19(e1), pp.e119–e124. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000508.

Marcos, M. et al., 2013. Interoperability of clinical decision-support systems and electronic health records using archetypes: a case study in clinical trial eligibility. *Journal of biomedical informatics*, 46(4), pp.676–89. Available at: http://dx.doi.org/10.1016/j.jbi.2013.05.004.

Mccauley, M.P. et al., 2017. Genetics and Genomics in Clinical Practice : The Views of Wisconsin Physicians. *WMJ*, 116(2), pp.69–75.

McMurry, A.J. et al., 2013. SHRINE: enabling nationally scalable multi-site disease studies. *PloS one*, 8(3), p.e55811. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23533569.

Moltchanov, S. et al., 2015. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *The Science of the total environment*, 502, pp.537–47. Available at: http://dx.doi.org/10.1016/j.scitotenv.2014.09.059.

Murphy, S.N. et al., 2010. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association : JAMIA*, 17(2), pp.124–30. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.2009.000893.

Nadkarni, P.M. & Brandt, C., 1998. Data Extraction and Ad Hoc Query of an Entity--Attribute--Value Database. *Journal of the American Medical Informatics Association*, 5(6), pp.511–527. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/jamia.1998.0050511.

Nagalla, S. & Bray, P.F., 2016. Personalized medicine in thrombosis: back to the future. *Blood*, 127(22), pp.2665–2671. Available at: http://linkinghub.elsevier.com/retrieve/pii/S2468171717300029.

Nunes, M. et al., 2015. Evaluating patient-derived colorectal cancer xenografts as preclinical models by comparison with patient clinical data. *Cancer research*, 75(8), pp.1560–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25712343.

Okada, Y. et al., 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488), pp.376–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24390342.

Oyelade, J. et al., 2015. Bioinformatics, Healthcare Informatics and Analytics: An Imperative for Improved Healthcare System. *International Journal of Applied Information Systems*, 8(5), pp.1–6. Available at: http://research.ijais.org/volume8/number5/ijais15-451318.pdf.

Pandey, A.S. & Divyasheesh, V., 2016. Applications of Bioinformatics in Medical Renovation and Research. *International Journal of Advanced Research in Computer Science and Software Engineering*, 6(3), pp.56–58.

Payne, P.R.O., Embi, P.J. & Sen, C.K., 2009. Translational informatics: enabling high-throughput research paradigms. *Physiological Genomics*, 39(3), pp.131–140. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2789669&tool=pmcentrez&rendertype=abstract.

Plunkett-Rondeau, J., Hyland, K. & Dasgupta, S., 2015. Training future physicians in the era of genomic medicine: trends in undergraduate medical genetics education. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(11), pp.927–34. Available at: http://www.nature.com/doifinder/10.1038/gim.2014.208.

Poon, C.C.Y. & Zhang, Y.-T., 2008. Perspectives on high technologies for low-cost healthcare. *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society*, 27(5), pp.42–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18799389.

Prahallad, A. et al., 2012. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387), pp.100–3. Available at: http://www.nature.com/doifinder/10.1038/nature10868.

Raghupathi, W. & Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1), p.3. Available at: http://www.hissjournal.com/content/2/1/3.

Ram, S. et al., 2015. Predicting Asthma-Related Emergency Department Visits Using Big Data. *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp.1216–1223. Available at: http://ieeexplore.ieee.org/document/7045443/.

Ramachandran, A. et al., 2013. Effectiveness of mobile phone messaging in prevention of type 2 diabetes by lifestyle modification in men in India: a prospective, parallel-group, randomised controlled trial. *The lancet. Diabetes & endocrinology*, 1(3), pp.191–8. Available at: http://dx.doi.org/10.1016/S2213-8587(13)70067-6.

Ramírez, M.R. et al., 2018. Big Data and Health "Clinical Records." In *Innovation in Medicine and Healthcare 2017*. Springer International Publishing AG 2018, pp. 12–18. Available at: http://link.springer.com/10.1007/978-3-319-39687-3.

Rehm, H.L., 2017. Evolving health care through personal genomics. *Nature reviews. Genetics*, 18(4),

pp.259–267. Available at: http://www.nature.com/doifinder/10.1038/nrg.2016.162.

Relling, M. V. & Evans, W.E., 2015. Pharmacogenomics in the clinic. *Nature*, 526(7573), pp.343–350. Available at: http://www.nature.com/doifinder/10.1038/nature15817.

Richesson, R.L. & Andrews, J.E., 2012. Introduction to Clinical Research Informatics. In *Health Informatics*. Springer-Verlag London Limited 2012, pp. 3–16. Available at: http://link.springer.com/10.1007/978-1-84882-448-5_1.

Rose, P.W. et al., 2011. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, 39(Database), pp.D392–D401. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq1021.

Rose, P.W. et al., 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic acids research*, 43(Database issue), pp.D345-56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25428375.

Ross, M.K., Wei, W. & Ohno-Machado, L., 2014. "Big data" and the electronic health record. *Yearbook of medical informatics*, 9(1), pp.97–104. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25123728%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4287068.

Sáez, C. et al., 2017. A Standardized and Data Quality Assessed Maternal-Child Care Integrated Data Repository for Research and Monitoring of Best Practices: A Pilot Project in Spain. *Studies in health technology and informatics*, 235, pp.539–543. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28423851.

Safran, C. et al., 2007. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, 14(1), pp.1–9. Available at: http://jamia.oxfordjournals.org/content/14/1/1.full.

Sanseau, P. et al., 2012. Use of genome-wide association studies for drug repositioning. *Nature biotechnology*, 30(4), pp.317–20. Available at: http://www.nature.com/doifinder/10.1038/nbt.2151.

Scanfeld, D., Scanfeld, V. & Larson, E.L., 2010. Dissemination of health information through social networks: twitter and antibiotics. *American journal of infection control*, 38(3), pp.182–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20347636.

Schadt, E.E., 2012. The changing privacy landscape in the era of big data. *Molecular Systems Biology*, 8(612), pp.1–3. Available at: http://msb.embopress.org/cgi/doi/10.1038/msb.2012.47.

Schaffer, J.D., Dimitrova, N. & Zhang, M., 2006. Chapter 26 BIOINFORMATICS. In *Advances in Healthcare Technology*. pp. 421–438.

Semenza, J.C. & Menne, B., 2009. Climate change and infectious diseases in Europe. *The Lancet. Infectious diseases*, 9(6), pp.365–75. Available at: http://dx.doi.org/10.1016/S1473-3099(09)70104-5.

Shameer, K. et al., 2017. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Briefings in Bioinformatics*, 18(1), pp.105–124. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv118.

Shameer, K., Readhead, B. & Dudley, J.T., 2015. Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Current topics in medicinal chemistry*, 15(1), pp.5–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25579574.

Sheehan, J. et al., 2016. Improving the value of clinical research through the use of Common Data Elements. *Clinical Trials*, 13(6), pp.671–676. Available at: http://journals.sagepub.com/doi/10.1177/1740774516653238.

Signorini, A., Segre, A.M. & Polgreen, P.M., 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS one*, 6(5), p.e19467. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21573238.

Silva, B.M.C. et al., 2015. Mobile-health: A review of current state in 2015. *Journal of biomedical informatics*, 56, pp.265–72. Available at: http://dx.doi.org/10.1016/j.jbi.2015.06.003.

Simon, R., 2005. Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(29), pp.7332–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16145063.

Sirota, M. et al., 2011. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, 3(96), p.96ra77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21849665.

Sun, J. et al., 2014. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association*, 21(2), pp.337–344. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-002033.

Taglang, G. & Jackson, D.B., 2016. Use of "big data" in drug discovery and clinical trials. *Gynecologic oncology*, 141(1), pp.17–23. Available at: http://dx.doi.org/10.1016/j.ygyno.2016.02.022.

Tenenbaum, J.D., 2016. Translational Bioinformatics: Past, Present, and Future. *Genomics, proteomics & bioinformatics*, 14(1), pp.31–41. Available at: http://dx.doi.org/10.1016/j.gpb.2016.01.003.

Teutsch, S.M. et al., 2009. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genetics in medicine : official journal of the American College of Medical Genetics*, 11(1), pp.3–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18813139.

The Eurowinter Group, 1997. Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. The Eurowinter Group. *Lancet (London, England)*, 349(9062), pp.1341–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9149695.

Toh, S. et al., 2011. Comparative-effectiveness research in distributed health data networks. *Clinical pharmacology and therapeutics*, 90(6), pp.883–7. Available at: http://doi.wiley.com/10.1038/clpt.2011.236.

Toubiana, L. & Cuggia, M., 2014. Big Data and Smart Health Strategies: Findings from the Health Information Systems Perspective. *IMIA Yearbook*, 9(1), pp.125–127. Available at: http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/imia-yearbook/archive/issue/1973/manuscript/22305.html.

Vilardell, M., Civit, S. & Herwig, R., 2013. An integrative computational analysis provides evidence for FBN1-associated network deregulation in trisomy 21. *Biology open*, 2(8), pp.771–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3744068&tool=pmcentrez&rendertype=abstract.

Vodopivec-Jamsek, V. et al., 2012. Mobile phone messaging for preventive health care. *The Cochrane database of systematic reviews*, 12(12), p.CD007457. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/23235643.

Wade, T.D. et al., 2014. Using patient lists to add value to integrated data repositories. *Journal of biomedical informatics*, 52, pp.72–7. Available at: http://dx.doi.org/10.1016/j.jbi.2014.02.010.

Wade, T.D., Hum, R.C. & Murphy, J.R., 2011. A Dimensional Bus model for integrating clinical and research data. *Journal of the American Medical Informatics Association*, 18(Supplement 1), pp.i96–i102. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000339.

Walker, K.L. et al., 2014. Using the CER Hub to ensure data quality in a multi-institution smoking cessation study. *Journal of the American Medical Informatics Association*, 21(6), pp.1129–1135. Available at: http://jamia.oxfordjournals.org/cgi/doi/10.1136/amiajnl-2013-002629%5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84929044127&partnerID=40&md5=2c0c1e46853824a8779ebce39d9aabd8.

Wang, X. & Liotta, L., 2011. Clinical bioinformatics: a new emerging science. *Journal of clinical bioinformatics*, 1(1), p.1. Available at: http://www.jclinbioinformatics.com/content/1/1/1.

Weber, G.M. et al., 2011. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. *Journal of the American Medical Informatics Association*, 18(Supplement 1), pp.i157–i160. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000200.

Weiner, M.G. & Embi, P.J., 2009. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Ann Intern Med*, 151(5), pp.359–360.

Weinstein, J.N. et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), pp.1113–20. Available at: http://www.nature.com/ng/journal/v45/n10/abs/ng.2764.html.

Weiskopf, N.G. & Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1), pp.144–151. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000681.

Westfall, J.M., Mold, J. & Fagnan, L., 2007. Practice-based research--"Blue Highways" on the NIH roadmap. *JAMA*, 297(4), pp.403–6. Available at: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.297.4.403.

Wilhelm, M. et al., 2014. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp.582–7. Available at: http://www.nature.com/doifinder/10.1038/nature13319.

Wishart, D.S. et al., 2013. HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic acids research*, 41(Database issue), pp.D801-7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23161693.

Wynden, R. et al., 2010. Ontology mapping and data discovery for the translational investigator. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2010, pp.66–70. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041530&tool=pmcentrez&rendertype=abstract.

Xu, H. et al., 2015. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association : JAMIA*, 22(1), pp.179–91. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2014-002649.

Zheng, Y.-L. et al., 2014. Unobtrusive Sensing and Wearable Devices for Health Informatics. *IEEE Transactions on Biomedical Engineering*, 61(5), pp.1538–1554. Available at:

http://ieeexplore.ieee.org/document/6756983/.

Zlotta, A.R., 2013. Words of wisdom: Re: Genome sequencing identifies a basis for everolimus sensitivity. *European urology*, 64(3), p.516. Available at: http://dx.doi.org/10.1016/j.eururo.2013.06.031.

Bioinformatics in food production and engineering

Anna Kujumdzieva
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
https://www.uni-sofia.bg/

Trayana Nedeva
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
https://www.uni-sofia.bg/

Alexander Savov
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
https://www.uni-sofia.bg/

Contents

Introduction

Food and nutrition have an important role in regulation of human body processes. The introduction of advanced techniques like "omics" in food science and practice causes serious difficulties in interpretation of accumulated great biological data sources. A decision of this problem is implementation of bioinformatics approach giving an excellent ground for successful development of food production and engineering.

Food acts as important regulating factor on different processes within the body, like metabolic, mental etc. Tentative growth of various chronic disease is also linked with food. Considerable endeavor is ensured to prompt and improve the nutritional potential and quality of food sources. Recently, food science has grown notably applying various smart techniques like "omics" series. In order to overcome the vast variety of data and difficulties in their interpretation a database is necessary. It can store and keep updating the comprehensive amount of biological data and resources, important for food and nutritional sciences. Thus, the development of bioinformatics in food will help in providing the simple and convenient ways for improving the food research and technologies.

Bioinformatics benefits the food production and nutrition

Bioinformatics strongly depends upon tuneful software solutions, disposable through electronic telecommunications to the individual scientist. The massive computing power of the modern computer systems is facing less and less limitations in storage of space and calculation time. Thus, the only limiting factor is the lack of information on specific topics. Since industrial food processes are based on food-grade organisms like bacteria, molds and yeasts, the advance in the number of complete genomic sequences of organisms leads to rapid increase in valuable knowledge to compensate this lack. This knowledge can be used in many different fields like metabolic engineering, cell performance as a micro-process factory and elaboration of new methods for preservation. Moreover, genomic knowledge food-grade microorganisms will innovate pre- and probiotic research in order to describe the broad range of bacterial properties from growth to stress responses, to multi-species microbial ecology within the human host.

Applied bioinformatics in nutrition food research: usage and examples

In order to realize the mechanisms of nutrients action, the investigators need to use a reductionist strategy. It poses the problem to the level of cells, proteins, genes, etc. Then, the knowledge gained is transferred to the level of human body to evaluate the nutrient effects. In this way, nutrition researchers regularly generate and interpret data at the molecular level. The serious and predictive understanding of metabolism needs nutrients and metabolites to be studied in the context of their associated regulatory mechanisms. For example, the peroxisome-proliferator activated receptors (PPARs) represent complex of molecules that directly link nutrient intake to organism response. PPARs are transcription factors that sense different metabolites, like fatty acids and their derivatives at the cellular level. After that, a launching of specific metabolic program by regulating the expression of a variety of target genes happens. As an answer to this complete mechanistic understanding of PPARs, a recent bioinformatics study was performed to predict PPAR gene targets on a genome-wide basis. In fact, this study gave the first library of nutrient-sensitive genes and showed for the first time how databases and software can be integrated to investigate nutritionally relevant logical questions.

It answers to the following questions:

i) Which genes are directly regulated by PPARs and, thus, by fatty acids and fatty acid derivatives?

ii) What are the biological functions of these fatty acid responsive genes?

iii) What other transcription factors regulate these fatty acid-responsive genes"

A simplified flow-chart in Fig. 1 illustrates how databases and software were integrated to answer these questions.

The diagram in Fig 1 illustrates the basic steps in predicting the regulatory effect of PPARs on gene expression. They can be summarized as follows:

- Search of literature in the PubMed database for manuscripts containing experimental evidence for DNA binding sites of PPARs;
- Use of these sites to build probability matrices with different probabilistic assumptions with the use of the CONSENSUS and GMMPS programs;
- Extract relevant genomic information (all known human genes, DNA regions upstream from their transcription start site, conserved elements within these upstream regions, and homologous genes in the mouse and rat genomes) using some custom programs;
- Scoring the probability matrices against DNA sequence upstream from known PPAR target genes and randomly selected genes in the genome using custom program and software;
- Application of techniques that minimized the number of false-negative and false-positive results in the detection of PPAR binding sites and identification of putative PPAR target genes on a genome-wide basis
- Analyzing the sets of genes (PPAR targets) by using a gene ontology analysis tool, along with custom software to determine the biological functions represented by each group.

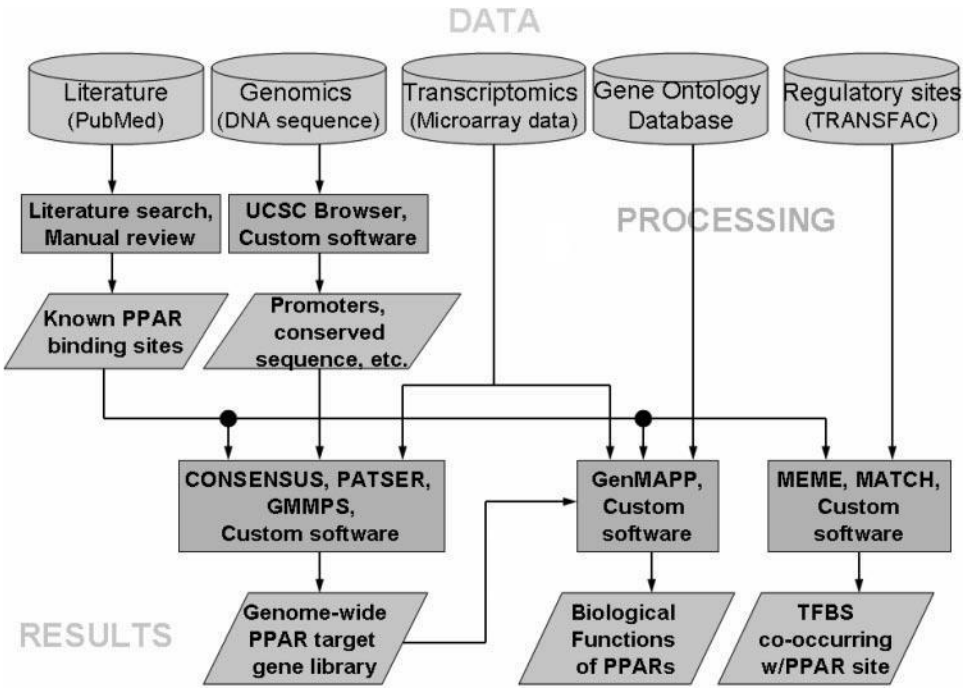


Figure 1. Integration of databases and software to predict genes regulated by Peroxisome-Proliferator (according to Lemay et al., Am J Clin Nutr 2007;86:1261–9)

Bioinformatics in reconstruction of metabolic pathways

Microbial metabolism has been the ground of a major part of food processing for centuries. Fermentation of food takes advantage of the ability of desirable microbes to convert substrates (usually carbohydrates) to organic tailor-made compounds contributing to the flavor, structure, texture, stability and safety of the food product.

Due to its fundamental importance to a wide variety of foods: breads, cheeses, wines, sausages etc., over a century of research has focused on understanding microbial metabolism. The potential to transform this knowledge into even greater value in foods has been dramatically expanded by the availability of tools to understand and control microbial metabolism using modern genomic and bioinformatics approaches. In fact, the tremendous information flow on microbial metabolism is only being converted into usable knowledge because of the arrival of the massive computing power and the bioinformatics' tools that are apply to large data sets generated by nutrition-related research.

This knowledge will not only drive a new generation of foods with additional values but also will change dramatically the ability of foods to influence individual quality of life.

Application of gene expression arrays

The ability of the nutrients to control directly the expression of specific genes is at the core of a new generation of nutritional science, which gives opportunity of researchers to use genomic information to develop technologies, able to measure the number of transcribing genes in any cell at any time (*i.e.* gene expression arrays). In this way, scientists are finding the intimate relationships between organisms and their environment.

Studies on the integrative metabolism of animals and humans are associated with food and nutrition as a multidisciplinary field center. Currently, the apparent strong relationship between diet and health is finding its mechanistic basis through understanding the interaction of nutrients with metabolic pathways. Since most nutrients affect a wide range of biochemical pathways, the food exerts multiple effects: pleiotropic dysfunctions in the relative absence of define nutrient, i.e. deficiencies, and pleiotropic benefits when they return to appropriate, optimal levels.

The classical biochemical approaches describe very well the effects of a single nutrient on a single target; however, the multiplicity of metabolic effects on the entire organism is difficult to be explained. The modern genomics uses the reverse approach: it measures everything. Genomic-based investigations reveal the pleiotropic behavior of exogenous nutrients through describing the full spectrum of transcriptional responses to any variable, including nutrients. These global experimental designs are possible due to the ability of bioinformatics tools to adequately manage and analyze the vast volume of accumulated data.

Genetic variability

After the sequencing of human genome the mapping of its polymorphic regions that control individual phenotypic differences among the population are going on. The established by this approach variations were thought at the beginning only as the key to the discovery of genetic diseases. However,

it is known now that they are also the keys to individual variation in diet and health. Sequence variation in particular gene (even in particular nucleotide, the so called Single Nucleotide Polymorphism - SNP) can influence the quantitative need for and physiological response to various nutrients. There are examples of polymorphism that influence nutrition and disease: the phenylketonuria, in which the inability to metabolize phenylalanine renders this nutrient toxic; the lactose intolerance due to polymorphism both in the structure of the lactase gene, which produce dysfunctional enzyme and in regulatory regions of the genome that prevent perfectly functional lactase enzyme from being produced in adults.

With genomics will come the knowledge of predicting health. The potential of bioinformatics to deliver knowledge about the integrative nature of multiple genes to the individual consumer will help in predicting its health leading to individualized dietary choices. This will be possible in close future due to the bioinformatics tools, capable of managing the volume of data implied by quantitatively assessing individual metabolism and intervening in an that individual's metabolism using foods to improve their health.

Genomic and bioinformatics tools will improve human nutrition trials. During their performance, it is not easy to find statistically significant positive effects of various nutrients and food because the magnitude of the benefit is quite small relative to the overall variability in a sample of humans chosen at random from the population and because humans do not respond homogeneously to even the most straightforward nutritional variables.

To overcome this obstacle, clinical and epidemiological trials are now being analyzed using SNP data as independent input variables. Most clinical trials build catalogues of SNPs of genes whose variation in function have shown to be important for manifestation of example cancer, autoimmunity and heart disease. Such approach has been successful not only in identifying the causes of statistical variation among individuals but also in identifying the potential biochemical mechanisms responsible for the variation in response.

Genetic polymorphism and nutrient requirements

Polymorphisms in the various genes encoding enzymes, transporter proteins and regulatory proteins affect the absolute quantities of essential nutrients (incl. vitamins, minerals, etc.) that are needed to satisfy the cell requirements for sufficiency. Consequently, the variation in the population's nutrient status is a complex value. It is a result of variations in food intakes plus inherent variations amongst individuals within the population in their genetically defined abilities to absorb, metabolize and utilize these nutrients. The figures for the recommended daily allowances of each nutrient are shaped on the basis of experimentally determined data for the needs of a statistically representative segment of the population. However, the range of responses to both micro-and macronutrients in the population as a whole is much larger. Specifically individual food choices, genetics and nutrition are linked in s complex way that was highlighted quite recently with the help of genomic tools. Thus, polymorphism in a recently identified sweet receptor protein has been proposed to be the basis for the varying intakes of caloric-rich foods, i.e. the famous sweet tooth.

Based on the information genomics succeeds to reveal for food preference and the corresponding roles of genetics and environment, the food science in now able to make nutritional superior foods that are more attractive (organolepticall) to that subset of the population for whom they are most appropriate. However, now the technologies to describe the effects of diet on individuals experimentally are used at broad basis only in clinical trials. They are not included yet in the routine consumer assessment. Therefore, consumers cannot benefit from nutritional knowledge about

themselves, because they simply do not have it. This lack of knowledge is the most important factor that influences negatively the widespread improvement in nutritional health in the consumer population.

Genetic variation and the response to variations in overall diet

The basic metabolism of macronutrients, especially of carbohydrates and fats in humans is strongly affected by genetic differences. For instance, polymorphisms in the apo-protein genes (apoE, apoAIV) or lipoprotein catalysts (lipoprotein lipase) have been shown to directly affect the clearance of dietary lipids. That is why polymorphisms in lipid metabolic genes command the response of the individuals to dietary fat in a different way. apoE protein clears liver-derived lipoproteins (VLDL and LDL) from blood. This functionality of the protein is influenced by the polymorphism in the genes encoding for it. In addition. health outcomes beyond heart disease including Alzheimer's disease have been shown to be correlated to apoE phenotypes. Apparently, diet plays a differential role in the development of these diseases according to genotype through the role of diet in influencing the quantitative flux of hepatic lipoprotein metabolism.

Many consumers consider the application of genomic testing in the population as useless or inappropriate. This is because they do not see any direct benefit for themselves. Nevertheless, acquiring knowledge about individual variation in diet-responsive genes is of great values, since this knowledge can be used for successful intervention. There are evidence that genotype predicts a difference in postprandial lipid metabolism of dietary fat. The translation of this discovery into practical recommendations how to alter the intakes of dietary fat for those affected is of great practical value. Thus, the information of how an individual responds to foods provides that individual with the means to change their diet to improve their health. Practically, each new discovery of genetic polymorphisms linked to health, is making the complexity of the science bigger. However, thanks to modern bioinformatics tools that are integrative by nature, each new discovery is added to the rapidly expanding coherent database of diet and health of individual consumers.

Bioinformatics approaches refine the food production

Biomass and metabolites yields

Optimization of biomass yield is by a topic of continuous attention in respect to improvement of the food production process. The genome-scale metabolic modelling is a technique applied to rationally improve fermentation yield. Within this technique, the genome sequence of the organism is used as a catalogue of the metabolic potential of a given strain. Using this technique, metabolic models have been made for many microorganisms, including several food-grade microbes. A limiting factor in the correctness of the metabolic models can be the quality of the genome sequence. For instance, a gene can be missed due to poor sequencing coverage. However, the metabolic model can be finalized by identifying those metabolic reactions that are missing in the model, but are likely to present because they are part of metabolic reaction cascade or pathway. The full genome-scale metabolic models allow the *in silico* simulation of growth of the organism under the (metabolic) restrictions provided by the substrate availability in the medium. These simulations can be used to optimize medium composition to better fit the organism requirements. Moreover, the models can suggest alternative or cheaper substrates for fermentation, and improve the production of essential compounds, taking into account

possible changes in activity with respect to flavour or texture activity of the strain. These models have also been implemented in complex (multistrain) fermentation processes, providing insight in the interactions between different species/strains in a complex fermentation.

A second factor that improves the overall yield is the robustness of the strains. This factor can be influenced largely by changing fermentation conditions under which starter cultures are prepared. For example, in *L. lactis* a number of genes that were potentially causative related to survival were identified by correlating the levels of gene expression to the survival of the species. The importance of these genes for the strains' phenotype was further proven by gene-disruption technique. It showed that not only gene itself but also its expression is important for a given phenotype. In other words, preconditioning *L. lactis* strains, followed by GTM and TTM, allows improving their survival to heat and oxidative stresses.

Texture and flavour performance

The fermentation process influences as well such important characteristics like the texture and the flavour of the food products. Since these traits are microorganism-specific, they can be altered by fermentation. For instance, addition of adjunct strains to cheese fermentation can change the product flavour or addition of exopolysaccharide-producing organisms can improve the texture of yoghurt. In a similar way, the flavour profiles of wine can be modified by either changing fermentation parameters or changing the starter cultures. Apparently, all these improvements can be made by testing a variety of experimental regimens. Thus, bioinformatics and data analytics may be used to optimize the designs of these experimental regimens.

The gene content of particular microorganisms under specific fermentation conditions may be used for deduction of their performance. Of course, such predictions based on a metabolic model must be further verified, as was the case with *L. lactis* MG1363 flavour formation. Similarly, the genomic sequence of *Lactobacillus delbrueckii* subsp. *bulgaricus* revealed how this species is adapted for the fermentation of milk and the production of yoghurt. The *Oenococcus oeni* and yeast genome analyses have been performed and their relation to wine fermentation was elucidated.

Besides these advantages of the metabolic models it is obvious that predicting more complex phenotype such as stress tolerance is less straight-forward to predict based only on gene content. For prediction of these phenotypes, information on the transcript levels of the genes might be taken into account.

The effects on taste and texture are mainly caused by the metabolites that are produced or transformed during fermentations. Predicting final sensory characteristics is possible using metabolite patterns rather than associating gene content with effects on taste texture. The quantitative descriptive analysis by a trained sensory panel is the golden standard test for sensory characteristics of a fermented product. However, these tests are elaborate and require substantial amounts of the product. In addition, the results are dependent on the panel experience. Using metabolomics' profiling techniques it is now possible to measure at the same time hundreds of metabolites in a food sample of small quantity. This has led to the development of new statistical methods that associate instrumental data (e.g. chromatographic and/or mass spectrometric ones) to sensory data.

Setting fermentations by mixed cultures

In the preparation of various fermented foods, complex fermentations take place in which strong succession of microbes (bacteria, yeasts and fungi) can occur. These are, for example the processes of obtaining cheese, malolactic wine, soy and seafood. Similar to the approaches of associating

transcription of genes to specific phenotypes, described in 2.3.2., presence and absence of (combinations of) microorganisms (or their functionality) can be associated to the characteristics of a fermentation product.

To characterize fermentation, the first essential step is to determine the microorganisms present at the different stages of the fermentation and to make correlation between these sets of microorganisms and the measurement of metabolites (making metabolomics). The functional potential encoded in their genomes determines the properties of the microbial consortia. These metagenomics studies also reveal DNA of unculturable organisms in addition to the DNA of the culturable ones. Thus, functionalities of the microorganisms can be predicted based on the sequences found in a consortium. However, there are still technical restrictions in identifying and separating the DNA of dead microbes that can be a reason for misleading results.

The mRNA-derived sequences of a complex fermentation can be profiled using metatranscriptomics approach. An advantage of metatranscriptomics over metagenomics approaches is that the gene expression measurement allows determining what genes are actually expressed in a mixed culture. Metatranscriptomics technique is using microarrays with the genomes of several species to determine global gene expression across a species. Practical application of this approach is reported for the bacterial communities involved. The advantage of this approach is that the metagenomics and metatranscriptomics profiles can be traced to their likely sources (genome sequences of isolates from the fermentation product). Thus the application of metagenomics/metatranscriptomics techniques to characterize and potentially optimize fermentations is apparent.

It is well known that bacteriophages play an important role in industrial fermentations due to the phenomenon genetic transduction via which biodiversity can be maintained. However, it is also known that phage sweeps disrupt fermentation processes with great efficiency. Currently, predicting the specificity of bacteriophages and the interactions between microorganisms in mixed-culture fermentations are time-consuming tasks. Bioinformatics techniques can be used to analyse the interaction of microbes and bacteriophages. They can contribute to knowledge-based improvements of fermentation stability. This could be achieved by performing experiments with *in situ* designed microbial consortia that are currently under development.

Bioinformatics in crop production and food processing

The progress of application of Genetically Modified Crops (GMC) as a common approach of food industry depends on genetic research of plants that contribute for successive rate of their production. The main objective of GMC production is to improve quality of raw materials of food supply to ensure their effective processing, and finally to result in costly and safety food. The identification of biosynthetic genes of plant origin that are important for health is supported by Genome sequencing projects. This genome research is directly involved in promoting efficiency and efficacy in plants breeding for their improvement.

A typical example in this direction is the Cocoa (*Theobroma cacao*) that is used as a raw material for chocolate containing food products. Selection of seeds with higher quality and good flavour has been difficult in the past. For proper seed harvesting the trees have to mature for at least 3 - 5 years. The performance of DNA fingerprinting in screening of plant markers for detection of breeds genotypic links and the availability of EST (Expressed Sequence Tags) sequences and genetic comparisons to other identified plants, all depend on bioinformatics. They will further improve selection of desired traits in early stage of plant's development based on genotype and phenotype.

As concern food processing, the most direct application of bioinformatics is in optimizing the quantitative parameters of traditional unit operations. In general, the main aim of processing food commodities is to improve storage stability and safety. Usually the processing procedures are associated with considerable excess of energy applied to ensure a large margin for error. The structural complexity of biological materials, the natural variability of living organisms and the response of the input materials to processing parameters are the three main factors that require the large error margin. With the help of bioinformatics our knowledge on biological organisms from bacteria and viruses to plants and animals is emerging progressively, facilitating the optimization of the food processes and diminishing all cost-important inputs, mainly energy.

The big challenge in modern food processing is to merge efficiently biological knowledge of living organisms with the bio-material knowledge necessary to convert them to foods.

Traditionally, during processing the biomaterials of living organisms are restructured into smaller and simpler forms of stable, relatively uniform foods. This process is strongly energy consuming and in most cases, along its performance the inherent biological properties of the living systems are lost. Bioinformatics offers detailed description of the inherent complexity of biological macromolecules within living cells, their structural properties and much of their functions, all of which make the fundamentals of functional genomics and proteomics. Although at the moment just theory, in near future it will be possible to use the inherent structural properties of natural food commodities to self-assemble new foods that retain great biological and nutritional value and that are processed with minimum energy. The biological structure–function relationships discovered through bioinformatics of living systems will be mapped into the structure–function relationships of the next generation of foods. Moreover, the vast knowledge currently being produced by the biomical sciences (genomics, proteomics, metabolomics) will improve the knowledge on ingredient characteristics and behaviours.

The natural properties of the biomaterial molecules that constitute living organisms determine the basic biomaterial properties of foods. While processing food stuffs in a traditional way, little advantage is taken of the unique properties of specific molecules. On the contrary, as a result of the classical processing methods all bio-molecules of a particular class (e.g. carbohydrates), are exposed to physical, thermal and mechanical energy to restructure them into more stable, and/or more bioavailable food systems. During this process all the unique differences (due to the characteristics inherent to biomolecules) are eliminated. Eliminated as well are the complex structure–function relationships of living organisms.

The food processing is not always necessary to the quality of foods. In fact, it is other way around: highly specific biological properties of the original living organism are a key to the processing strategy and contribute significantly to the organoleptic properties of the final food products. For instance, the treatment with rennet enzyme of bovine milk induces the natural aggregation of milk caseins leading to gelation during cheese manufacture. The texture and the organoleptic properties of the final product is due to the unique self-assembly properties of milk casein micelles that are colloidally stabilized in milk by kappa caseins but destabilized when enzymatically cleaved of their solubilizing glycomacropeptide. Another example is the leavening of bread, in which wheat seeds are ground to disassemble their biological structures through mechanical energy, and then the biological processes of yeast fermentation achieve simultaneously the enzymatic elimination of phytic acid during dough incubation and the biochemical production of CO_2 as leavening within a mechanically reworked protein gel structure. Thus, cheeses and breads provide proof of positive synergetic effect due to combination of retained biological processes of catalysis, self-assembly and restructuring. However, the functional genomics, proteomics and metabolomics are providing the knowledge necessary to readdress food processing using bimolecular activities. With the availability of such tools in hand, crops production will be organized that will result in products not simply enriched in a single valuable

component, but redesigned with a renewed purpose to increase the innumerable values of foods in providing quality of life.

Bioinformatics in food quality & safety

Food science represents a multidisciplinary research and applies area that unifies engineering, biological and physical sciences to explore the types of foods, reasons of their deterioration, mechanisms in food processing and retrieve of food quality. Bioinformatics is executing an important role during most of the processes, if only the data about them are accessible in machine-readable formats. Having in mind the important role of microorganisms in food, the use of bioinformatics tools for predicting and assessing their desired and undesired effects is of special interest. In this respect, the investigations in genomics and proteomics are performed to meet the requirements of food production, food processing, refine the quality and nutritive value of food sources and many others.

Besides, the bioinformatics approaches can also be applied in fabrication the good quality of the crop comprising high yield and disease defense. Different databases containing data on food, their constituents, nutritive value, chemistry and biology exist and can be used in food research and manufacture. A combination of bioinformatics with laboratory verification of selected findings can be outlined with the following methods: genomics-based functional predictions; genomic scale metabolic models, design of complex food properties and engineering.

The research focus in the food industry is outlined by the consumers need for high quality, convenient, tasty, safe and affordable food.

Nutrition and food quality

Modern food science and technology have provided incomparable value to consumers in the literally innumerable number of individual choices of delicious, safe and nutritious foods. This great variety of choices has been supported by scientific knowledge at all levels of the food chain from genetic improvements in agriculture production to engineering of food processes and analysis of consumer sensation. With its power to create detailed molecular knowledge of biological organisms, bioinformatics is assembling the tools to reinvent the food supply. In this way bioinformatics will produce for humans a great value contributing to the increase in the quality of their lives through the quality of the foods they eat. In particular, bioinformatics is:

- Defining which foods are safe at molecular scale;
- Developing safer to the consumer foods;
- Helping to understand the fundamentals of food flavours, textures and taste sensation and understanding the relevant neurophysiological processes;
- Improving the process of food making and optimizing the flavor and texture impact of foods.

Specific food characteristics effecting its quality

The following important elements characterizing food are used as indicators to develop its description through bioinformatics tools.

Food taste

There are molecular and genetic details of the taste receptors including: sour bitter, umami, sweet, salt. These taste receptors can be used to discover the next generation of taste modifiers for foods. New developments in computational algorithms and software with the available known structures of these receptors have made possible the molecular modelling and simulations. Such simulations will make possible to develop more intense tasting compounds as food additives. These also help in understanding the basis of taste persistence, antagonism and complementation. Bioinformatics sequence similarity algorithms have been used to determine homology between sweet taste receptors and brain glutamate receptors as well as in the identification of sour taste sensors in mammals. Flavor systems are becoming more complex, more attractive and more individualized to consumers.

Food flavour

The formation of flavour in dairy products strongly depends on the essential role of lactic acid. In this respect the investigation of the genetic sequences of lactic acid bacteria showed the flavour forming potential. The profile of many food products does not depend on single compounds but is due to the availability and liaison of many different molecules.

However, bioinformatics plays a serious role in connecting different flavour compounds for new product development on the ground of knowledge, taste and needs of the consumer. Bioinformatics has a considerable cue in providing food quality taste flavour maintaining also its safety. Running in accordance with the molecular evolution, bioinformatics has a pivotal cue in study of evolution of receptors for taste.

With various studies being conducted primarily focusing on the taste receptors with the link between the glucose regulation and bitter taste receptors established. Recently, electronic database was established which include the chemical properties of various compounds related to their taste and flavour. Moreover, study of genetic sequence of lactic acid bacteria played an important role in uncovering the formation of specific flavouring potential helping in giving flavour to many fermented foodstuffs.

In addition to the taste receptors the odor receptors (exceeding the taste ones by 100 X) are being identified as well and the full olfactory complement of genes has been published. This bioinformatics approach to both taste and odor receptors study allows design of sophisticated flavor systems that optimize flavor perception in highly nutritious foods that are currently organoleptically undesirable although their great health value.

Food borne pathogens

Recently, it is admitted that a growing appreciation for bioinformatics exists in the area of food quality and safety. A major problem of food industry are food borne pathogens and the genome sequencing projects are now focusing on innovative tools helping to determine the source of the food borne diseases. Thus, the notification of the specific molecular markers can help in determination of spoilage and pathogenic bacteria and prediction of thermal preservation stress resistance.

A very important output of bioinformatics is the design of tool for detecting and identifying bacterial food pathogens. This tool has been developed by FDA (Food and Drug Administration) for molecular characterization of bacterial food borne pathogens using microarrays.

Due to its potential many genomic sequencing projects are targeting on the food-borne pathogens. With the development of genomic sequencing technologies bioinformatics has propose an innovative way which will help in determining the source of the food-borne diseases. For instance, recently developed approach by the FDA (Food and Drug Administration) helps in detection of the bacterial food pathogen and these computer based tools are focusing on microbial growth prediction on a given food source. To ensure food quality progress it is necessary to use bioinformatics tools that allow detection of various properties of food automatically.

Detection of food allergens

Bioinformatics give efficient approach to evaluate allergenic potential of normal proteins in food and have an important role in safety assessment of genetically modified crops as it is crucial to have safety from food allergy. These tools are acting for prediction of functionality and allergenicity of food products studying the protein sequence of their ingredients. Practically, a comparative genomics technique of bioinformatics has been used to characterize many food related pathogens associate with food and sources linked to their production. They have been an object of many sequencing and comparative genomic research projects. The results obtained showed that such studies can have significant cue in prevention of crop related disease and food poisoning. Crops are major part of food industry and for this reason must be of good quality (i.e. high yielding and disease resistant).Using bioinformatics approach genes identification in the commercially important crops can be used in development of transgenic crops and new genes can increase quality and quantity of food products. Such technique can be useful in elaboration of agro-chemicals based on the idea of signal transduction pathways for specific targets and finding of compounds applicable as pesticide, herbicide or insecticide. Because of the very distinct origin of allergens they possess very large sequence similarity in the structure causing equivalent responses of IgE. The use of these methodologies has incited WHO to involve sequences similarity search as rules of the feature for evaluating allergenicity of genetically modified food. Recently, various techniques of bioinformatics have been performed for allergen diagnostic development to predict the peanut allergy with the help of machine learning.

At present, different databases dedicated to the food allergens exist, like AllerMatch, Informall, FARRP Allergen database and SDAP.

Bioinformatics in food quality and safety

There is a growing appreciation for bioinformatics in the area of good quality and safety. Many food products undergo some form of processing before they reach the consumer, ranging from fermentation to packaging. In many of these processes, microorganisms play important roles, either in transforming the food into the desired end product or in spoiling or contaminating the food.

Bioinformatics plays an increasing role in predicting and assessing the desired and undesired effects of microorganisms on food. I respect to the desired properties, bioinformatics methods can be used to improve the microbial production of fermented food products, such as genomics-based functional predictions, the creation of genome-scale metabolic models and prediction of complex food properties (e.g. taste and texture), and properties of complex fermentations.

For deduction of a specific gene function, correlating analysis of the presence and absence of

the gene in organisms with the presence and absence of a certain phenotypic trait in the same set of organisms (the so called gene–trait matching; GTM) is applied. For instance, a set of proteins was predicted to be involved in the degradation of plant (oligo-)saccharides by linking isolation source of bacteria to gene presence/absence.

In the light of food safety, comparative analysis of the genome sequences of a species where some strains have a positive impact (e.g. flavour enhancement) while others are detrimental (e.g. spoilage) can be used to identify genetic elements potentially underlying these differences.

Tools that can be used to link -omics data to phenotypes are PhenoLink and DuctApe. Techniques like multiple displacement amplification can be used to amplify DNA from a single cell, and a range of genome assembly tools can be used to assemble the reads obtained from single-cell sequencing.

And finally, mobile elements such as transposons, plasmids or phages can transfer functionality from one bacterial strain to another. An example is the galactose utilization operon transfer between *Lactococcus lactis* strains. Identifying potential transposon insertion sites is crucial and can be facilitated by bioinformatics tools such as transposon insertion finder

Risk assessment

The identification of potential health or safety risks of microbial strains present in the food is an important step is risk assessment of food products consumption. Bioinformatics contribute to this issue with the performance of selectively screening microbial genome sequences for genes with specific functionalities - a highly sensitive and computationally efficient way of identification of potential health hazards.

The potential of a specific bacterium for antibiotic resistance or virulence can be investigated by comparing its genome sequence to a reference database containing known resistance genes and virulence factors. Similar approaches have been described for the identification of persistence of bacteria in food products, anaerobic spore-forming organisms in food and potential pathogens using metagenomics data. This (meta)genomics-based methodology can be applied to a wide range of functionalities, e.g. production of antimicrobial peptides.

Tracing and detection of food microorganisms

Food production and food consumption both take place in complex environments. There, besides the microorganisms present in the natural environment, many other sources of biomolecules (proteins, fats and carbohydrates) are present. This complexity is causing difficulties in detection and tracing of specific microorganisms, either potential food pathogens or beneficial probiotic strains added to the food product to enhance its functionality.

Next to classical detection DNA-based techniques such as (q)PCR, new methods based on genomic data have been developed that allow for a fast and precise tracking or detection of specific species or even strains among the natural microflora. For instance, specific amplification and sequencing of a locus that was identified to be discriminatory between different *L. plantarum* strains was performed and the data obtained showed that this is a useful approach to quantify the relative presence of different strains through the passage of the GIT. The same approach can be followed to design specific primers to distinguish between pathogenic and non-pathogenic populations of specific species and to detect a strain of interest in food products, allowing this specific product to be branded.

The metagenome approaches for dedicated tracing of a single strain can reveal their potential in the detection of harmful bacteria as well. The main advantages of these methods that do not require culturing stage, overcome the concern of creating bias in the results due to failure of detecting low

abundant microbes that might be overgrown in culture-dependent detection methods.

The role of toxicogenomics in foods' quality guarantee

Food safety is becoming more and more a major area of concern for consumers and the food industry has developed a coherent research programme to ensure food safety with well-established classical methodologies but also new state-of-the-art research tools. The goal here is to ensure that the inactivation or inhibition of undesired microbes is possible using the minimum treatment of foods necessary, to increase the understanding on the ecology of food-born microbial populations, to find-out how these populations respond to environmental factors like stress and last but not least the toxicological evaluation of foods and food compounds.

A branch of genomics, toxicogenomics, is an emerging field that contributes to evaluation of toxicological effects of specific compounds. Toxicogenomics utilizes DNA arrays (tox-chips) to test the toxicological effects of a particular compound. The DNA arrays techniques is based on the DNA microchip methodology and it probes human or animal genetic material printed on micro-devices to profile gene expression in cells exposed to test compounds. This technique avoids the study of animal pathology to define illness. The advantages of the test are speed and ease of use, typical for DNA expression analysis, and reduced animal testing. The application of this technique presently faces the challenge of accumulation of massive amounts of data, which are produced through the DNA arrays and their sophisticated analysis and interpretation. Nevertheless, the integration of tox-chip data must into the knowledge basis of the research institutions is a question of near future.

Perspectives

Bioinformatics is increasingly applied in food production, engineering and safety. Some future trends of its potential implementation are as follow:

- Sequence-based prediction of microbial functionality. An inventory is needed of the functionalities, for which bacteria can reliably be determined using sequence data. New publicly available data sets with genotype/phenotype/transcriptome such as those available for *L. lactis* and *L. plantarum* could help to develop new sequence-based functional prediction strategies such as further specified protein domains to more specifically screen for, e.g., carbohydrate active enzymes and relating promoters or regulatory binding sites to phenotype.
- Establishment of culture collections for desired traits on the basis of knowledge-based *in silico* screening. This would require databases that integrate data from genomics, systems biology, phenotypes, ingredient information, properties of batches of foods, on-line measuring of parameters during the food making process and 'biomarkers' for functionality in specific taxa (based on, e.g., GTM). Specific emphasis should be put in propagating the FAIR (findable, accessible, interoperable, re-usable; http://datafairport.org/) principle in storing data. The future software and databases can be consolidated in a virtual machine that can subsequently be run in the cloud. First steps in this direction are being made in the EU-funded project GenoBox (www.genobox.eu) that aims to create a database that consolidates genotype and phenotype data that allow screening microbial genomes for functionality and safety risk factors.
- Creation of database to assess risks of the presence of certain microbes/functionality in a given food product. The idea is to determine minor levels of microbial components in many food products across the world through sequencing of the food supply chain. The project is already established by a consortium of IBM and MARS (http://www.research.ibm.com/client-programs/foodsafety/). The ambition is into this data base sufficient biodiversity to be recorded and

further use for branding products based on unique microbiota paterns present in fermented products or foods that contain a microbiome.

- Directing fermentations performance through studying the interactions between microbes and their environment. These approach use systems biology beyond genome-scale metabolic models and kinetic models to describe interactions between microbes and their matrix. To be realized these studies require a substantial knowledge base on both the properties of the microorganisms and the physical properties of the matrix in which the organism operate. The consolidation of the information and expanding amount of data on food fermentation and safety in databases and its combination with appropriate experimental design, algorithms, expertise and follow-up experiments should allow enhancing the prediction of fermentation performance and safety.

References

1. Abee T. Van Schaik W. & Siezen R. J. (2004). Impact of genomics on microbial food safety. Trends in Biotechnology 22, 653-660.

2. Alkema W. Boekhorst J. Wels M. & S. A. F. T. Van Hijum. (2015). Microbial bioinformatics for food safety and production. Briefings in Bioinformatics.

3. Brul S. Schuren F. Montijn R. Keijser B. J. F. Van Der Spek H. & Oomes S. J. C. M. (2006). The impact of functional genomics on microbiological food quality and safety. International Journal of Food Microbiology 112 195-199.

4. Carrasco-Castilla J. Hernandez-Alvarez A. J. Jimenez-Martınez C. Gutierrez-Lopez G. F. & Davila-Ortiz G. (2012). Use of proteomics and peptidomics methods in food bioactive peptide science and engineering. Food Engineering Reviews 4 224-243.

5. Chibuike C. Udenigwe Bioinformatics approaches prospects and challenges of food bioactive peptide research Trends in Food Science & Technology Volume 36 Issue 2 April 2014 Pages 137-143 ISSN 0924-2244.

6. Desiere F., German B., Watzke H., Pfeifer A., Saguy S. (2001). Bioinformatics and data knowledge: the new frontiers for nutrition and foods. Trends in Food Science & Technology 12 (7): 215-229; ISSN0922244 http://dx.doi.org/10.1016/S09242244(01)00089-9.

7. FAO/WHO. (2001). Evaluation of allergenicity of genetically modified foods. Report of a joint FAO/WHO expert consultation on 14 T.A. Holton et al. / Trends in Food Science & Technology 34 (2013) 5-17 allergenicity of foods derived from biotechnology. Rome: Food and Agriculture Organization of the United Nations (FAO).

8. Lemay D. G. Zivkovic A. M. & German J. B. (2007). Building the bridges to bioinformatics in nutrition research. The American Journal of Clinical Nutrition 86 1261-1269.

9. Liu M. Nauta A. Francke C. & Siezen R. J. (2008). Comparative genomics of enzymes in flavor-forming pathways from amino acids in lactic acid bacteria. Applied and Environmental Microbiology 74 4590-4600.

10. Mari A. Scala E. Palazzo P. Ridolfi S. Zennaro D. & Carabella G. (2006). Bioinformatics applied to allergy: allergen databases from collecting sequence information to data integration. The allergome platform as a model. Cellular Immunology 244 97-100.

11. Mochida K. & Shinozaki K. (2010). Genomics and bioinformatics resources for crop improvement. Plant and Cell Physiology 51 497-523.

12. R.D Pridmore D Crouzillat C Walker S Foley R Zink M.-C Zwahlen H Brüssow V Pétiard B Mollet Genomics molecular genetics and the food industry Journal of Biotechnology Volume 78 Issue 3 31 March 2000 Pages 251-258 ISSN 0168-1656.

13. Waidha K. M., Jabalia N., Singh D., Jha A. and Kaur R., Bioinformatics Approaches in Food Industry: An Overview. Conference Paper November 2015, DOI: 10.13140/RG.2.2.27961.77926

14. Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acid Res 1996;24:238 – 41.

15. The Universal Protein Resource (UniProt). Nucleic Acid Res 2007;35: D193–7.

16. http://www.spss.com/ Clementine

17. http://www.ifst.org/fst.htm

18. http://snp.cshl.org

19. http://datafairport.org

20. http://www.research.ibm.com/client-programs/foodsafety/

The role of bioinformatics in agriculture

Tsvetina Mihailova
BULGAP Ltd.
Sofia, Bulgaria
http://bggap.eu

Ventsislava Petrova
BULGAP Ltd.
Sofia, Bulgaria
http://bggap.eu

Kliment Petrov
BULGAP Ltd.
Sofia, Bulgaria
http://bggap.eu

Contents

Bioinformatics for agriculture

Genomics, metabolomics and interactomics for sustainable agricultural development

Agriculturis not only a major occupation of a few nations, but also way of life, culture and custom. Cee reals like rice, wheat, barley, corn, sorghum, millet, sugar cane have always been considered as important food in human populations over different continents. From thousands of years, people are using breeding and selection to make domestic varieties of these crops with the wanted characteristics. Significant progress has been completed in taste, nutritional value and productivity, especially during the “Green Revolution” which took place in 1960 - 1970.

However, the Green Revolution has been also known with its unsuccess and we are no longer capable to survive by few “high yield” varieties. That's why now we need to use more advanced and modern biotechnology methods in agronomy in order to supply nutritional food to continuous increasing world population while considering three important limitations - less plow lands, depletion of energy resources and unpredictable climate change. In other word, we need to enlarge the pace of research so we can be capable to provide enough food for future generations.

The last ten years were considered to be a new era of bioinformatics and computational biology which enlarges the pace of scientific invention in life science. Involvement of computer science in the area of plant biology has change the way we usually do research related to plants in previous decades. Rapid ground breaking progress of sequencing technology during the few last years made this technology so cost-effective that nowadays it is common for any experimental lab to use sequencing methods to study genome of interest.

Including modern biotechnology progress in agriculture will definitely achieve huge dividends to the bioenergy sector, agro-based industries, agricultural by-products utilization, plant improvement and better management of the environment. Latest genome and transcriptomics sequencing of a plant species gives the opportunity to reveal the genetic architecture of many plant species, the differences in thousands of individuals within and outside population, the genes and mutations which are essential for improving the particular wanted complex traits (Fig. 1).

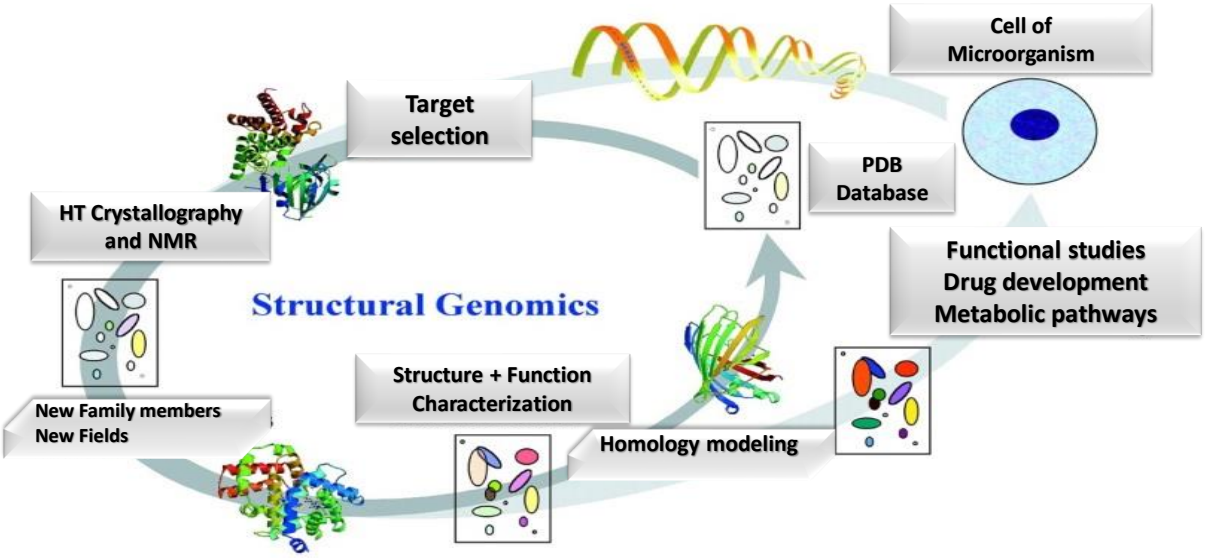


Fig. 1. Structural Genomics

Therefore, we need to use genomics resources available for many non-model and model plant species as a result of rapid technological progress in omics and bioinformatics fields which finally led us to admit new translational area of plant science well-known as 'Plant Genomics'. Within the scope of plant genomics, we will be able to do following activities:

1. Sequencing and de novo assembly of non-model plant species;
2. Making a detailed list of genes with their functional annotation and ontology;
3. Discovery of a great quantity of SNP (single nucleotide polymorphism) / InDeLs (insertion-deletion length polymorphism) markers to help in fine mapping and selection of superior breed;
4. Identify "candidate genes/mutations/alleles" in conjunction with wanted traits after differentiating underlying QTLs (quantitative trait locus) from markers generated in 2) using QTL mapping methods e.g. GWAS (genome-wide association study);
5. Creating "MarkerChip Panel" for the purpose of genotyping and selection.

In this respect, metabolomics is also fast emerging field in the world of omics, and normally used to scan all the metabolites present in sample using LC-MS, NMR-MS and GC-MS instruments. For example in human, it was used to define all the possible metabolites which directly or indirectly indicate food habit of an individuals whose urine samples were collected, analyzed in one of MS instruments and obtained data process computationally (Fig. 2).

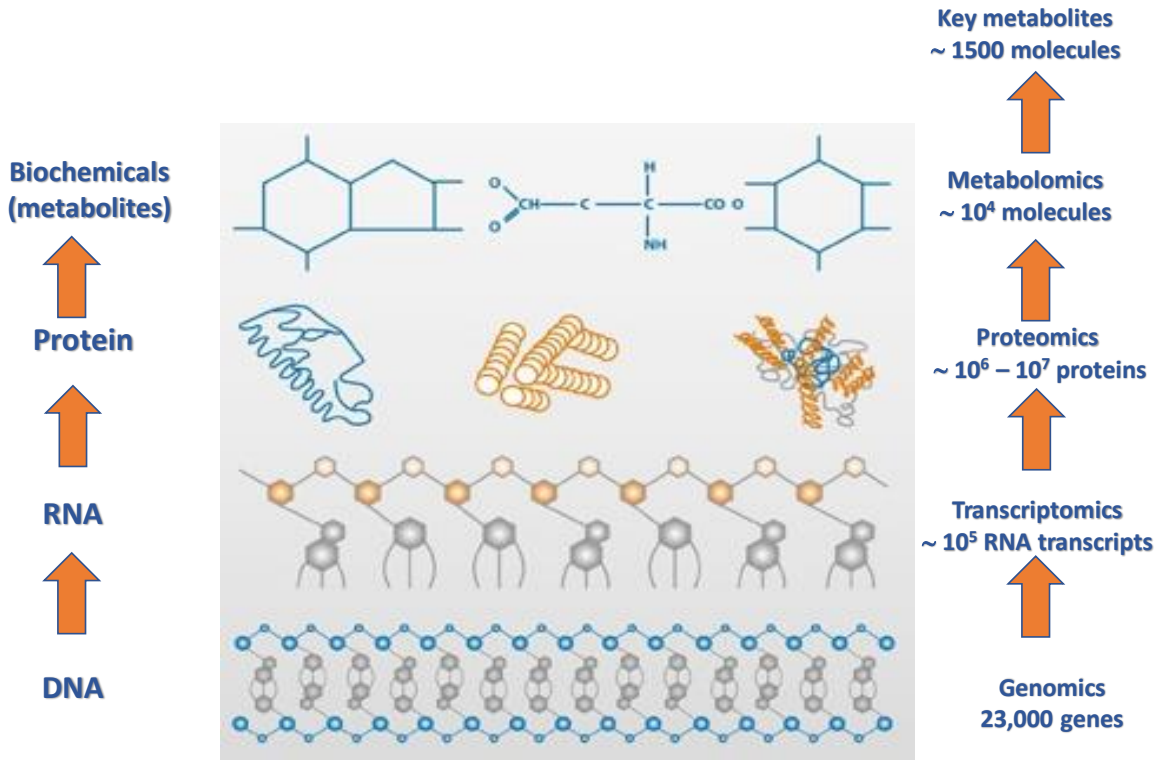


Fig. 2. Metabolomics Technology

Also the interactome is made up from a complete set of all protein–protein interactions which help to understand the molecular networks governing cellular systems. For example, the interaction map of *Arabidopsis* revealed about thousands of highly reliable relations between proteins (Arabidopsis Interactome Mapping Consortium 2011).

Impact of genome sequencing in agriculture

The term genome can be applied particularly to the whole genetic material of an organism including the full set of nuclear DNA (i.e., nuclear genome) and also to the genetic information stored within organelles, which have their own DNA - the 'mitochondrial genome' or the 'chloroplast genome'.

Some organisms have multiple copies of chromosomes, which are diploid, triploid, tetraploid, etc. In the reproducing organism (typically eukaryotes) the gamete has half of the number of chromosome of the somatic cell and the genome is a complete set of chromosomes in a gamete.

Moreover, the genome can contain non chromosomal genetic elements like viruses, plasmids or transposable elements. Most biological units which are more complex than a virus, have extra genetic material besides that which has in their chromosomes. Therefore 'genome' describes all of the genes and information on non-coding DNA that have the potential to be present.

However, in eukaryotes like plants, protozoa or animals, 'genome' is typically associated with only the information on chromosomal DNA. The genetic information contained by DNA within organelles i.e., chloroplast and/or mitochondria is not considered to be a part of the genome. Actually, mitochondria are sometimes mentioned to carry their own genome often called 'mitochondrial genome' (Fig. 3). The DNA established in the chloroplast may be called 'plastome' (Fig. 4).

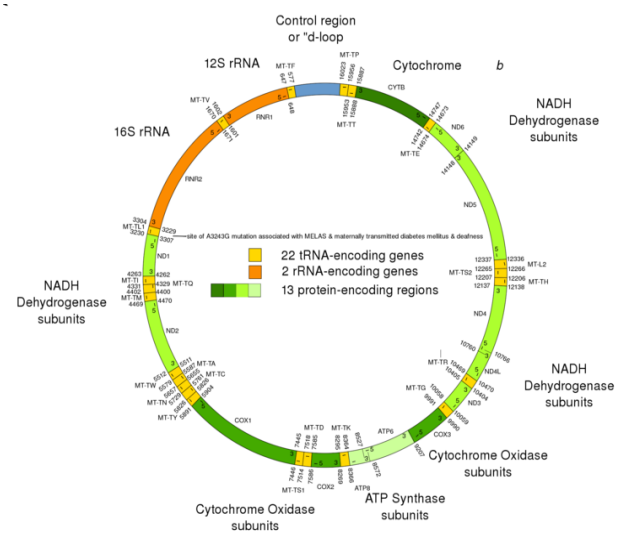


Fig. 3. Mitochondrial genome

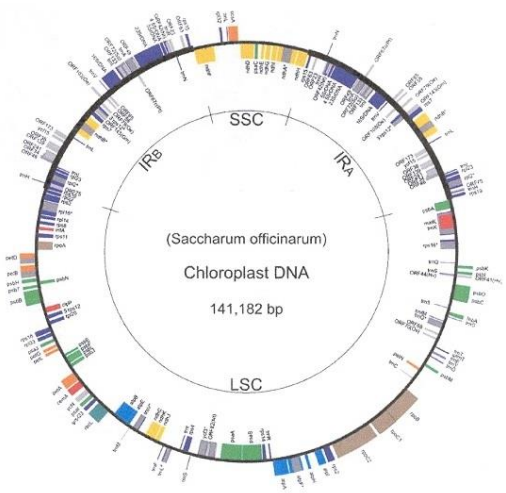


Fig. 4. Plastome

The better understanding of genome evolution comes from the comparative analysis in microbial genome which uses metabolic comparison and gene organization at metabolic reactions level with their operons using pathway, reaction, structure, compounds and gene orthologs. In this regard, the sequencing of whole genomes from various species allows determining their organization and provides the starting point for understanding their functionality, thus favoring human agriculture practice.

At this point, the contribution of genomics to agriculture includes the identification and the manipulation of genes related to particular phenotypic traits as well as genomics breeding by marker-assisted selection of variants. The name “agricultural genomics” (or agri-genomics) aims to find innovative decisions through the study of crops or livestock genomes, reaching information for protection and sustainable productivity for food industry, but also for different aspects such as energy production or design of pharmaceuticals.

Because of the fact that most bacterial species are still unknown most of the methods used for profiling microbial society and characterize their basic functional features are now accepting whole DNA extraction and the use of NGS (Next-Generation Sequencing) on the entire sample, with the objective of sequencing and characterizing DNA fragments of all the species included, i.e., the metagenome (Fig. 5).

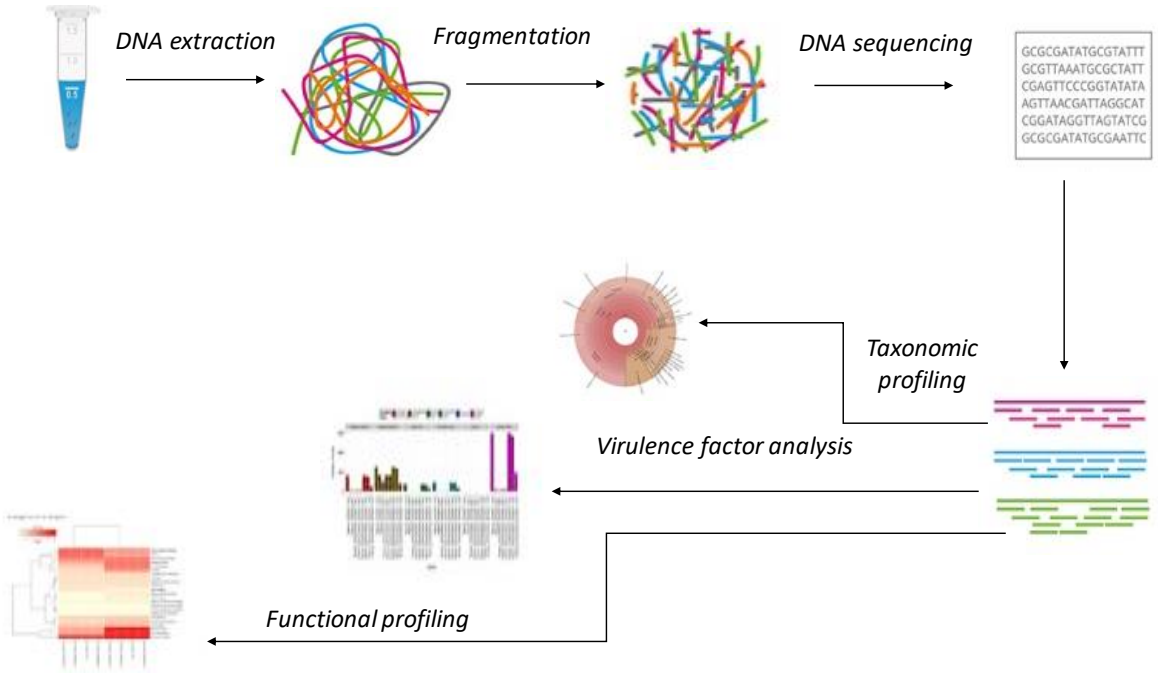


Fig. 5. Metagenome analysis

The application of metagenomics in agriculture also showed to be appropriate for representing the complex patterns of interactions occurring among microorganisms in soil and in plant rhizosphere or in particular tissues or organs. Moreover, metagenomics showed to be useful for tracing the shift in taxonomic composition and functional redundancy of microbial society in rhizosphere and in soil which are in connection to environmental changes related to fertilization and agricultural management.

Applications of agricultural bioinformatics

Collection and storage of plant genetic resource can be used to manufacture stronger, disease and insect resistant crops and improve the quality of livestock making them healthier, more resistant to diseases and more productive.

Comparative genetics of the model and non-model plant species can discover an organization of their genes with respect to each other which are used after that for transferring information from the model crop systems to other food crops. In this regard, examples of existing full plant genomes are *Arabidopsis thaliana* (water cress) and *Oryza sativa* (rice).

Also one of the resources for receiving energy by converting into biofuels such as ethanol is plant based biomass and it could be used as for vehicles and planes. In addition, biomass based crop species like maize (corn), switch grass and lignocellulosic species like bagasse and straw are widely used for biofuel production. Accordingly, the use of genomics and bioinformatics in combination with breeding would likely increase the ability of breeding crop species to be being used as biofuel feedstock and therefore keep increasing the use of renewable energy in modern society.

In addition, genes from *Bacillus thuringiensis* which can control a number of serious pests have been successfully transferred to cotton, maize and potatoes. This new ability of the plants to resist insect outbreak may decrease the number of used insecticides and therefore will increase the nutritional quality of the crops (Fig. 6).

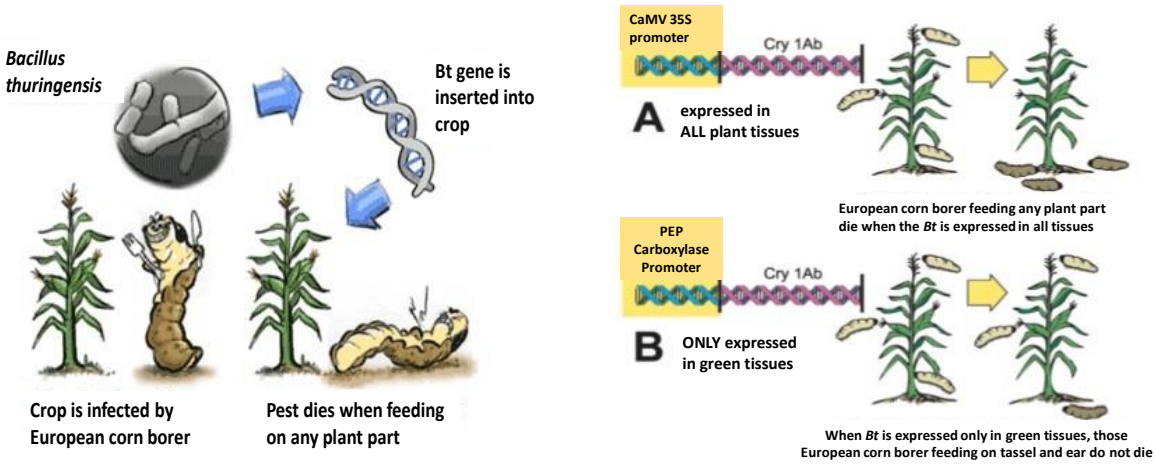


Fig. 6. *Bacillus thuringiensis* gene

Scientists have recently succeeded in transferring genes into rice to enlarge the levels of Vitamin A, iron and other micronutrients. This success could have a deep impact in reducing incidents of blindness and anemia caused by deficiencies in Vitamin A and iron respectively (Fig. 7.1, 7.2).

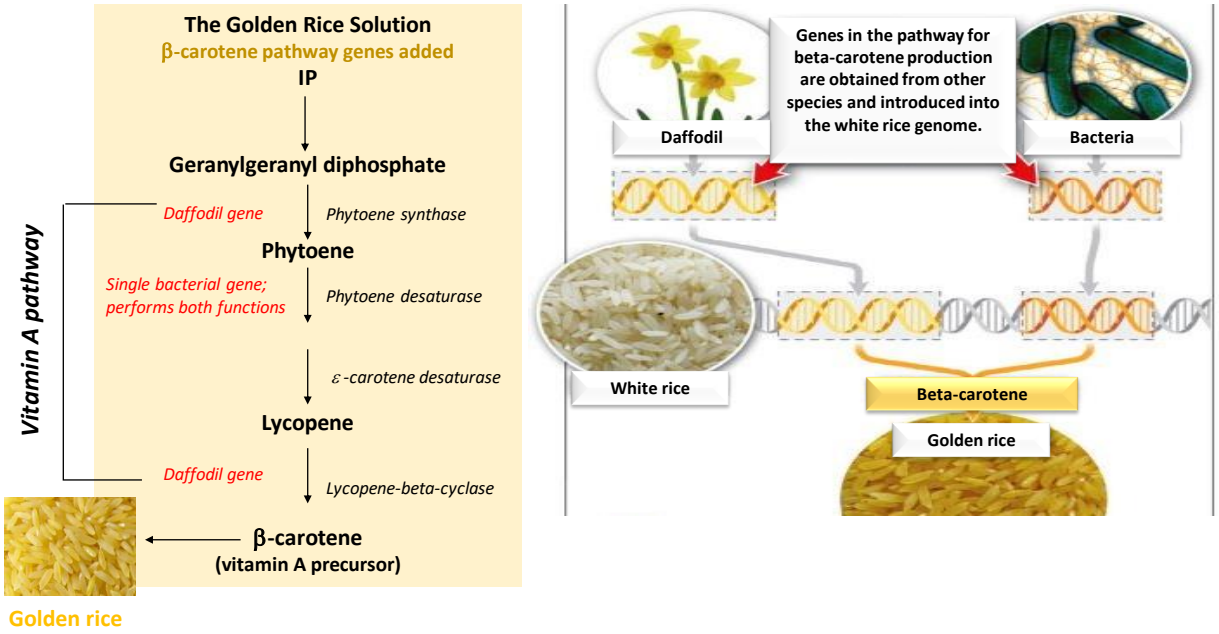


Fig. 7. Transfer of genes into rice to enlarge the levels of Vitamin A

Another example is the achieved progress in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities. These varieties will let agriculture succeed in poorer soil areas, therefore adding much more land to the global production base.

In this regard, the purpose of plant genomics is to understand the genetic and molecular basis of all biological processes in plants which are corresponding to the species. This understanding is fundamental because it will allow efficient exploitation of plants as biological resources in the evolution of new cultivars with improved quality and reduced economic and environmental costs. Traits of

primary interest are, pathogen and abiotic stress resistance, quality characteristics for plant, and reproductive characteristics determining yield.

Agriculturally important biological database

At the beginning of the "genomic revolution", the fundamental task of bioinformatics was to establish and maintain databases to store biological information like nucleotide and amino acid sequences.

A biological database is a big, organized form of constant data, which is generally related to computerized software projected to update, query, and retrieve components of the information stored within the system. For example, a record related to a nucleotide sequence database normally contains data like contact name; the input sequence with a description of the type of molecule; the scientific name of the source organism from which it was isolated; and, frequently, literature citations related to the sequence.

The development of the database include not only design and store information but also the elaboration of user friendly GUI (graphical user interface) so investigators could both access existing data and submit new or revised data e.g., NCBI, Ensembl.

There are many helpful databases where we can obtain the corresponding information about specific plant species.

For example PlantTribes 2.0 database is a plant gene family database based on the inferred proteomes of five sequenced plant species: *Arabidopsis thaliana*, *Carica papaya*, *Medicago truncatula*, *Oryza sativa* and *Populus trichocarpa*. It uses the graph-based clustering algorithm MCL to categorize all of these species' protein-coding genes into supposed gene families, also called tribes, using three clustering stringencies (low, medium and high). For all tribes, it generates protein and DNA alignments and maximum-probability phylogenetic trees (Fig. 8).

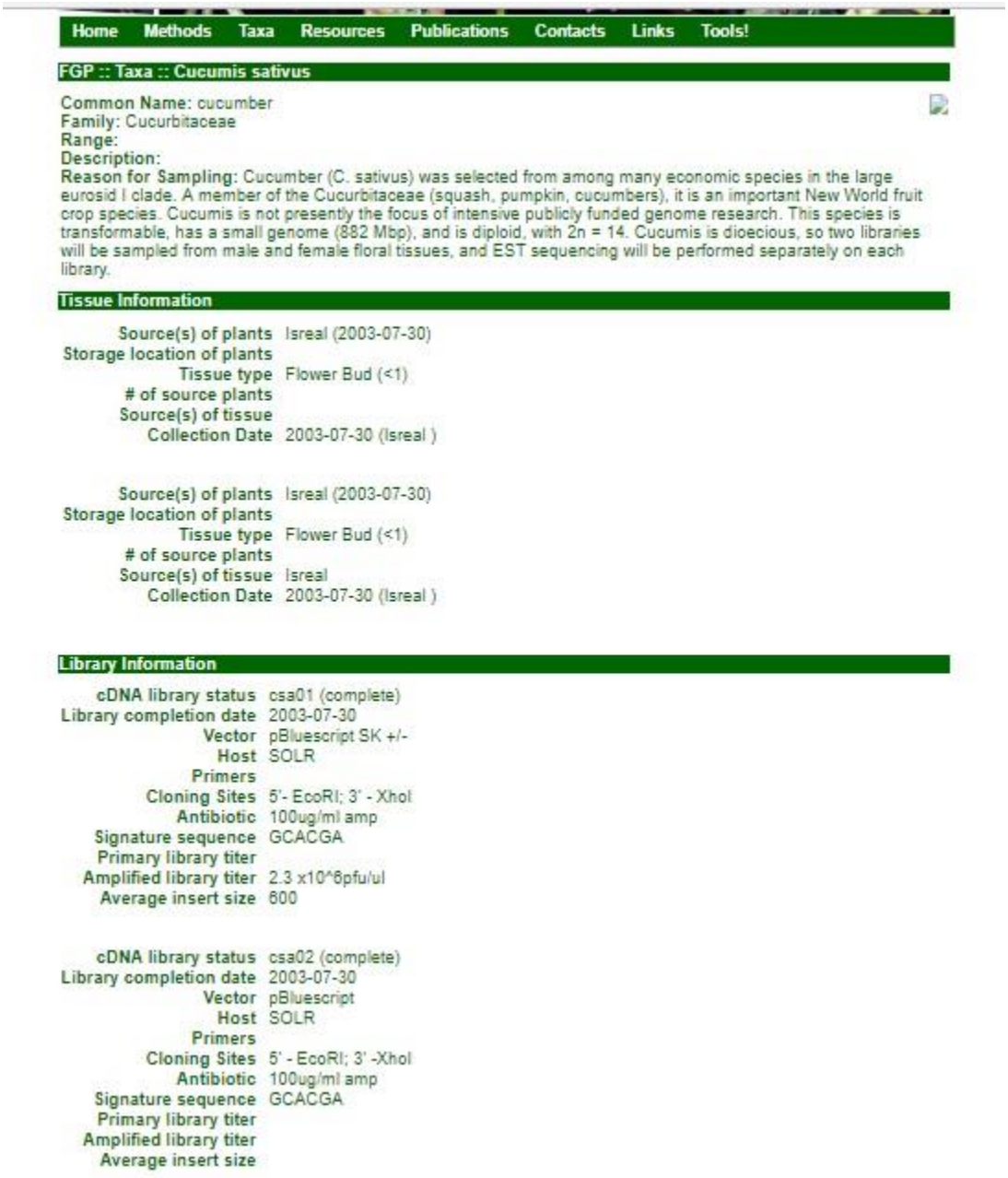


Fig. 8. PlantTribes 2.0 database

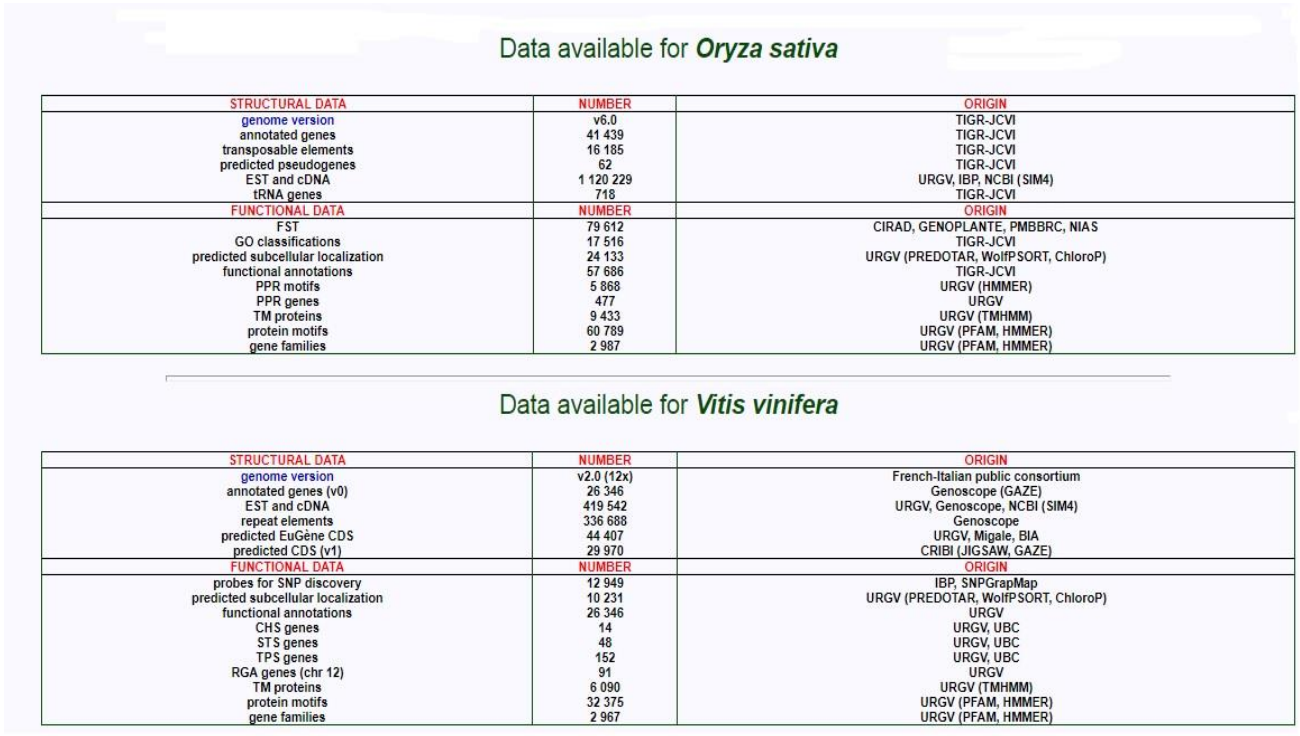
There is also a parallel database of microarray experimental results related to the genes, which allows explorers to identify groups of associated genes and their expression patterns.

SuperTribes, built via second iteration of MCL clustering, connect distant, but potentially related gene clusters. All information and analyses are available by a flexible interface allowing users to explore the classification, to place query sequences within the classification, and to download results for further study.

In his latest version, they have import additional another fine scale classification for identifying orthologous genes based on OrthoMCL algorithm.

Another database, the FlagDB database, characterizes a big integrative collection of the structural and functional annotations, and ESTs from six different plant species. Additionally, there are also information about novel gene predictions, mutant tags, gene families, protein motifs, transcriptome data, repeat sequences, primers and tags for genomic approaches, subcellular targeting, secondary

structures, tertiary models, curated annotations and mutant phenotypes, which are accessible in this database (Fig. 9).



Data available for ***Oryza sativa***

STRUCTURAL DATA	NUMBER	ORIGIN
genome version	v6.0	TIGR-JCVI
annotated genes	41 439	TIGR-JCVI
transposable elements	16 185	TIGR-JCVI
predicted pseudogenes	62	TIGR-JCVI
EST and cDNA	1 120 229	URGV, IBP, NCBI (SIM4)
tRNA genes	718	TIGR-JCVI
FUNCTIONAL DATA	**NUMBER**	**ORIGIN**
FST	79 612	CIRAD, GENOPLANTE, PMBBRC, NIAS
GO classifications	17 516	TIGR-JCVI
predicted subcellular localization	24 133	URGV (PREDOTAR, WolfPSORT, ChloroP)
functional annotations	57 686	TIGR-JCVI
PPR motifs	5 868	URGV (HMMER)
PPR genes	477	URGV
TM proteins	9 433	URGV (TMHMM)
protein motifs	60 789	URGV (PFAM, HMMER)
gene families	2 987	URGV (PFAM, HMMER)

Data available for ***Vitis vinifera***

STRUCTURAL DATA	NUMBER	ORIGIN
genome version	v2.0 (12x)	French-Italian public consortium
annotated genes (v0)	26 346	Genoscope (GAZE)
EST and cDNA	419 542	URGV, Genoscope, NCBI (SIM4)
repeat elements	336 688	Genoscope
predicted EuGène CDS	44 407	URGV, Migale, BIA
predicted CDS (v1)	29 970	CRIBI (JIGSAW, GAZE)
FUNCTIONAL DATA	**NUMBER**	**ORIGIN**
probes for SNP discovery	12 949	IBP, SNPGrapMap
predicted subcellular localization	10 231	URGV (PREDOTAR, WolfPSORT, ChloroP)
functional annotations	26 346	URGV
CHS genes	14	URGV, UBC
STS genes	48	URGV, UBC
TPS genes	152	URGV, UBC
RGA genes (chr 12)	91	URGV
TM proteins	6 090	URGV (TMHMM)
protein motifs	32 375	URGV (PFAM, HMMER)
gene families	2 967	URGV (PFAM, HMMER)

Fig. 9. Data available in FlagDB database (FLAGdb++ v6.2)

Another important example is the Plant genome database: PlantGDB is a catalogue of genomic sequences of all the plant species, created for the purpose to perform comparative genomics. This database also classifies EST sequences into contigs which could characterize and distinguish unique genes (Fig. 10).

Fig. 10. The Plant genome database: PlantGDB

Other agriculturally important databases along with description and URL are given at Health Science Library System.

Plant genomics

The role of model organism

Over the final century, the investigation and research on a few number of life forms has played an essential role in our understanding of various biological cycles and processes. This is because numerous aspects of science, especially biological processes, are comparable in most even in all living organisms. However, often it is much easier to explore a specific aspect or process in one organism than in others. In this case, these organisms are commonly suggested as model organisms, because their characteristics make them appropriate for laboratory study.

In 1980s, much more people started to think that major investments in studies of numerous different plants like corn, oilseed rape or soybean will dilute efforts to fully understand the main properties of all plants. Moreover, scientists started to realize that their purpose of fully understanding the plant physiology and development is so ambitious that the best decision is to use a model plant species that many scientists can solely explore.

The most well known model organisms have to possess solid preferences for experimental research, such as fast development with short life cycle, small adult measures, ready availability, and tractability. Due to the exstensive study of their characteristics these model organisms become even more valuable. In this point a huge amount of data can be determined from these organisms, giving important information for the analysis of normal human or crop development; gene control, genetic infections and deseases, and evolutionary forms.

For example, *Medicago* (alfalfa) is a real brilliant diploid which has a significant role in fixing soil nitrogen and has a major part of forage diets. Other grasses and legumes are being also used for extensive EST sequencing and for genetic maps construction. Luckily, the total sequencing of all the genes of one representative plant species will give much more knowledge and information for all higher plants. Also, using model species will further expand the knowledge about all higher plants, especially in revealing the role of proteins and discovery of their functions. For example, the comparison of genome sequences of rice and *Arabidopsis* revealed planty of useful information for plant genomics because of their extensive but complex designs of synthesis.

Arabidopsis thaliana has become a well-known model plant for most of the researchers. In spite of the fact that it is a non-commercial plant, it is preferred because of its reproduction, development and reaction to stress and disease in the same way as many crop plants. *Arabidopsis thaliana* has a small genome which does not have the repeated, less-informative DNA sequences that hinder genome analysis performance. Its advantages are that it has large genetic and physical maps of all 5 chromosomes (MapViewer); a fast life cycle (around 6 weeks from seed germination to grown seed); productive seed manufacture and simple cultivation in limited space; a huge number of mutant lines and genomic resources (Stock Centers) and multinational research society of academic, government and industry laboratories.

The whole genome of *Arabidopsis* has duplicated once throughout its evolution and this event is followed by subsequent gene loss and extensive local gene duplications. The genome has 25,498 genes encoding proteins from 11,000 families (Fig. 11).

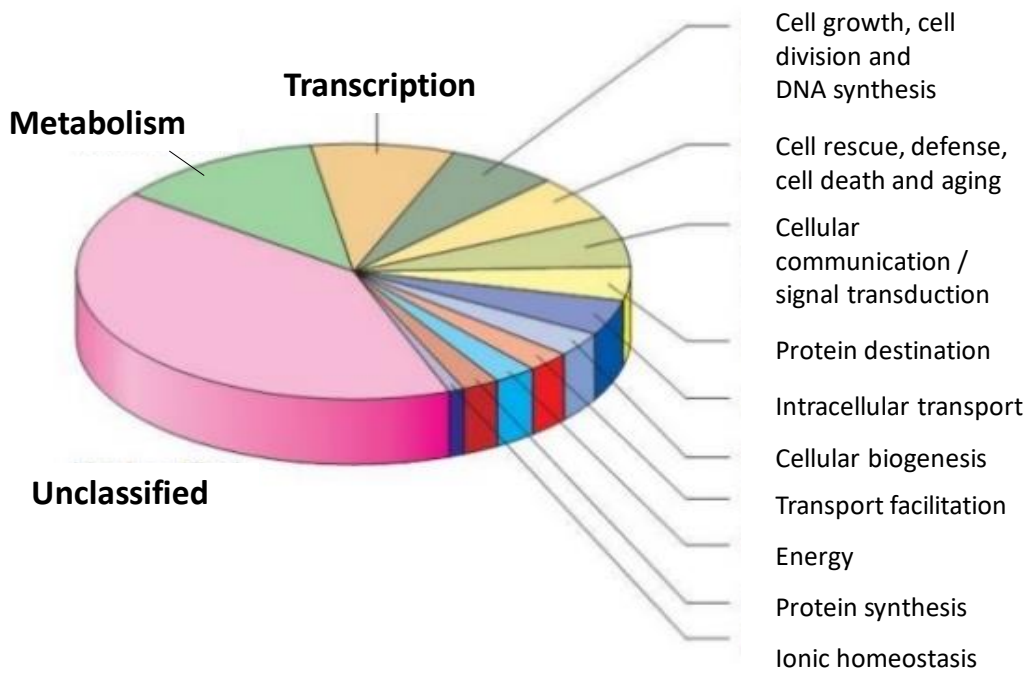


Fig. 11. Analysis of *Arabidopsis thaliana*

Like other model organisms, there is much more information for *Arabidopsis* genome than the complete genome sequence. The website for the *Arabidopsis* Information Resource, TAIR, allows explorers to integrate the genome sequence with a large EST database and with the genetic and physical maps, offers links to functional and molecular genetic information and the literature for specific genes and indicates an ever expanding list of mutant stocks.

Alternative plants that are used as model organisms for research are tomatoes, rice, maize and wheat, because of their significant characteristics.

All the available research and genetic data for different model plants are uploaded on corresponding websites. Generally, they are made by particular research groups who integrate the research efforts from all over the world. A few valuable websites include the UK CropNet, the U.S. Agricultural Research Service and organism-specific resources like MaizeDB. These sites aim to link seed stock and actual genetic resources to virtual information on linkage mapping information. That is why various search engines and complex relational databases are under development.

Managing and distributing plant genome data

Genome science has profited significantly from the progress in computing capabilities and bioinformatics, as with numerous areas of science and technology. The growth of the Internet has been vital for genome researchers as well as the improved computational speed.

In conjunction with the development of modern database technology, the World Wide Web has become the native medium for managing and disseminating genomic resources and this led to the creation of shared public resources, which were used for searching and analyzing the contents of genomic databases. Some of the Websites like NCBI and EMBL give quick access to colossal amounts of information and analysis tools, free of charge, from anyplace of the globe. In addition, the advantages of networking have been important for the management of laboratory data with little or no human intervention.

LIMS or laboratory information management systems, let users at different workstations or geographic locations to browse, edit, analyze and comment the data. The main part of the genomic data is a database system and most databases can be classified as either relational databases (RDB) or object-oriented databases (OODB) (Fig. 12).

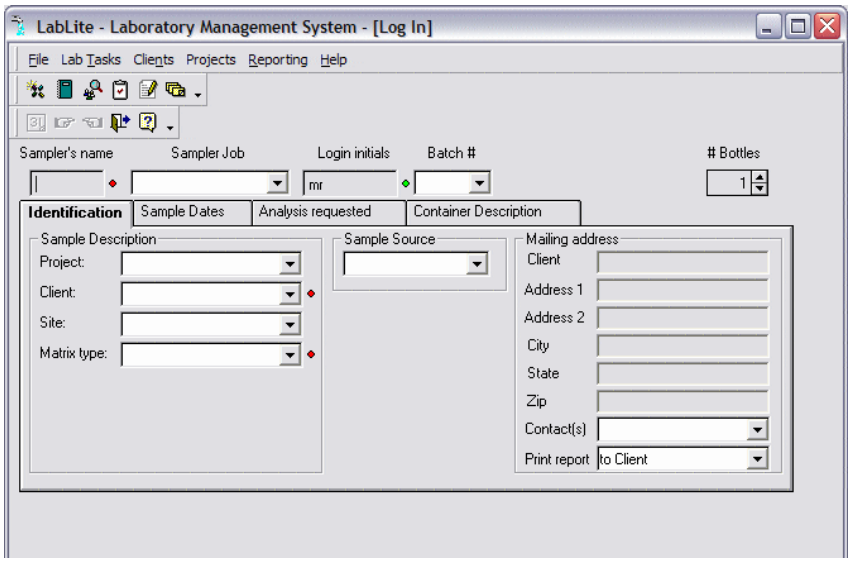


Fig. 12. Laboratory information management systems (LIMS)

There are three essential sequence databases: GenBank (NCBI), the Nucleotide Sequence Database (EMBL) and the DNA Databank of Japan (DDBJ) which are repositories for plant raw sequence information. So also, SWISS-PROT and TrEMBL are the major essential databases for the storage of plant protein sequences. There are also secondary databases such as PROSITE, PRINTS and BLOCKS and the sequences they contain are not raw data, but are derived from the data in the primary databases.

The early bioinformatics databases emphasized primary on data capture. To the early part of this decade the emphasis moved from information capture to information aggregation and integration. Model Organism Databases (MODs), integrated depositories of all the electronic data resources relating to a specific experimental plant or animal species, became the first choice of the bioinformatics world. Integrating numerous types of biological information over several species, these resources enable analysts to make disclosures that wouldn't be possible by analyzing a single species alone. These systems integrate information on numerous organisms and use comparative analysis to find patterns in genome that might otherwise be missed.

The maize genome, for example, is around the same length as the human genome, and won't be fully sequenced for another few years, but the rice which is one tenth the size of human's, is already sequenced. Because the two grains are closely related in evolutionary aspect, specific maps have been successfully created that relate maize's genetic map to the rice genome sequence. This lets analysts to follow a genetically mapped characteristic in maize, such as tolerance to high salt levels in the soil, and move into the relevant region in the rice genome, thereby recognising candidate genes for salt tolerance.

Currently, different bioinformatics approache are applied when studing plant genome data. Some of the most popular are:

Sequence alignment methods and applications for comparing genome sequences: The progress of technologies for the large scale quantification and identification of biological molecules combined with the progress of computing technologies and the internet has contributed to facilitate the delivery of major volumes of biological data to the analysts. The increased productivity was gained through automation, miniaturization, and integration of technologies and applying this approach to the assays

of other biological molecules including mRNA, proteins, and metabolites has effected in a large increase in the generation of biological information.

Very often the main essence of the bioinformatics strategies for sequence alignment is the comparison of cDNA/EST and genomic sequences and annotation. In addition to whole genome sequencing, plant sequence information have been collecting from three main sources: sample sequencing of bacterial artifcial chromosomes (BACs), genome survey sequencing (GSS) and sequencing of expressed sequence tags (ESTs).

Sequence alignment: This is the arrangement of two or more amino acid or nucleotide sequences from an organism or organisms in such a way as to adjust areas of the sequences sharing common properties. Well known versions for pairwise alignment are the Smith-Waterman algorithm for local alignment and the Needelman-Wunsch algorithm for global alignment.

Multiple sequence alignment: Multiple alignment demonstrates relationships between two or more sequences. When the involved sequences are different, the conserved residues are often key residues related to maintenance of structural stability or biological function. Multiple alignments can divulge a lot of clues about protein structure and function. The most commonly used alignment software is the ClustalW package.

Sequence Similarity Searching Algorithms: Possibly the most used of these are FASTA and BLAST. Both tools BLAST and FASTA provide very fast searches of sequence databases (Fig 13, 14).

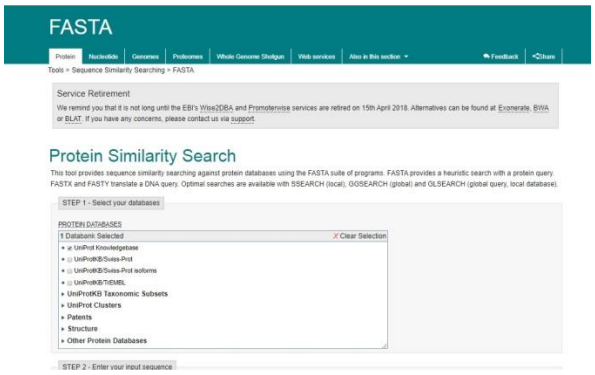


Fig. 13. FASTA

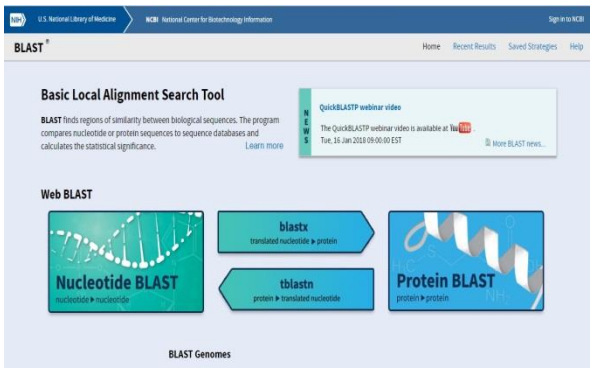


Fig. 14 BLAST

Genome Comparison Tools: MegaBlast is NCBI BLAST based algorithm for large sequence similarity search. MegaBlast is used to liken the raw genomic sequences to a database of contaminant sequences.

Expressed sequence tags (ESTs): ESTs are fractional, gene sequences which have been produced or are in the process of being produced in several laboratories using different species and cultivars as well as diversed tissues and developmental stages. ESTs are now widely used throughout the genomics and molecular biology society for gene discovery, mapping, polymorphism assay, expression studies, and gene prediction.

Molecular plant breeding

Because the resolution of genetic maps in the important crops expands, and because the molecular basis for particular characteristics or physiological responses becomes better clarified, it will be much more possible to associate candidate genes, found in model species, with relevant loci in crop plants. Appropriate relational data will make it possible to freely connect through genomes with regard to gene sequence, supposed function, or genetic map position.

Once this kind of tools have been realized and implemented, the difference between breeding and molecular genetics will disappear. Breeders will use computer models to formulate predictive hypotheses to establish phenotypes of interest from difficult complex allele combinations, and then make those combinations by scoring major populations for a lot of numbers of genetic markers (Fig. 15).

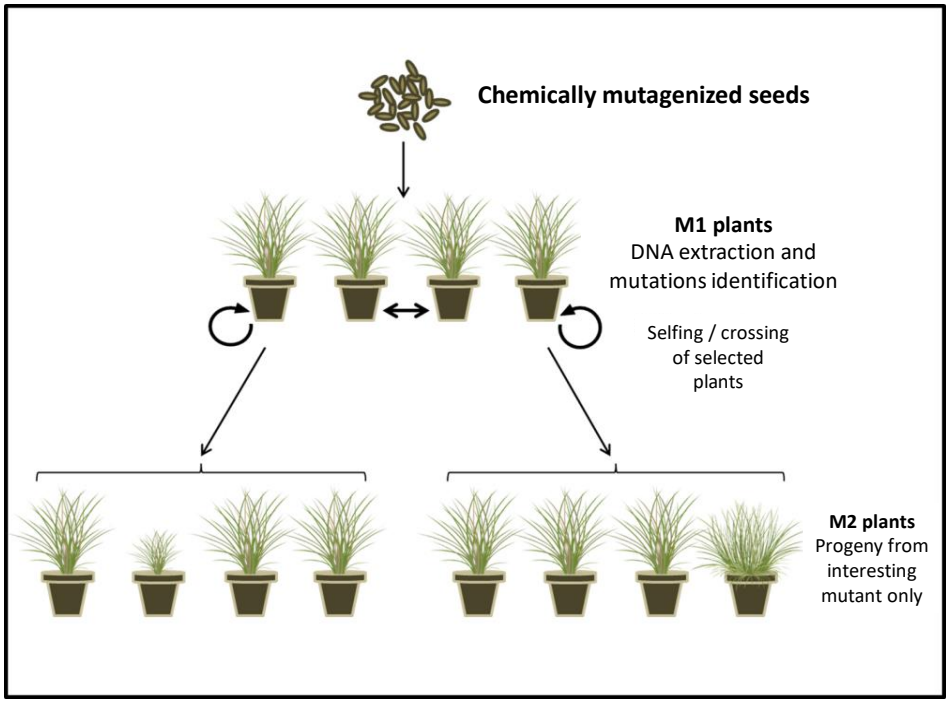


Fig. 15. Reverse genetics in perennial ryegrass

The tremendous resource including breeding knowledge collected over the last decades will become straight linked to basic plant biology, and will increase the ability to clarify gene function in model organisms. For example, characteristics which are badly determined at the biochemical level but well established as a visible phenotype can be related to high resolution mapping with candidate genes.

Orthologous genes in a model species, such as *Arabidopsis* or rice, may not have a well known connection with a quantitative characteristic like that seen in the crop, but might have been involved in a specific pathway or signaling chain by genetic or biochemical tests. This kind of cross-genome referencing will guide to a convergence of economically corresponding breeding information with main molecular genetic data.

The particular phenotypes of commercial interest which are expected to be spectaculary improved by this progress include both the improvement of factors which frequantly limit agronomic performance (input traits) and the change of the amount and type of materials that crops produce (output traits). Examples include:

- abiotic stress tolerance (cold, drought, immersion, salt);
- biotic stress tolerance (fungal, bacterial, viral);
- nutrient use efficiency;
- management of plant architecture and progress (size, shape, number, and position, timing of evolution, senescence);
- metabolite division (redirecting of carbon flow through existing pathways, or moving into new pathways).

Rational plant improvement

The implications of genomics with relation to food, feed and fibre production can be visualized on a lot of fronts. At the most essential level, the progress in genomics will considerably speed up the acquisition of knowledge and that, in turn, will directly effect on many aspects of the processes associated with plant improvement. Knowledge of the function of all plant genes, according to the further elaboration of tools for modifying and examining genomes, will lead to the evolution of an original genetic engineering paradigm in which rational changes can be intented and modelled from first principles.

The goal of plant genomics is to understand the genetic and molecular basis of all biological processes in plants which are related to the species. This understanding is essential to allow efficient maintenance of plants as biological resources in the development of new cultivars with improved quality and reduced economic and environmental costs (Fig. 16).

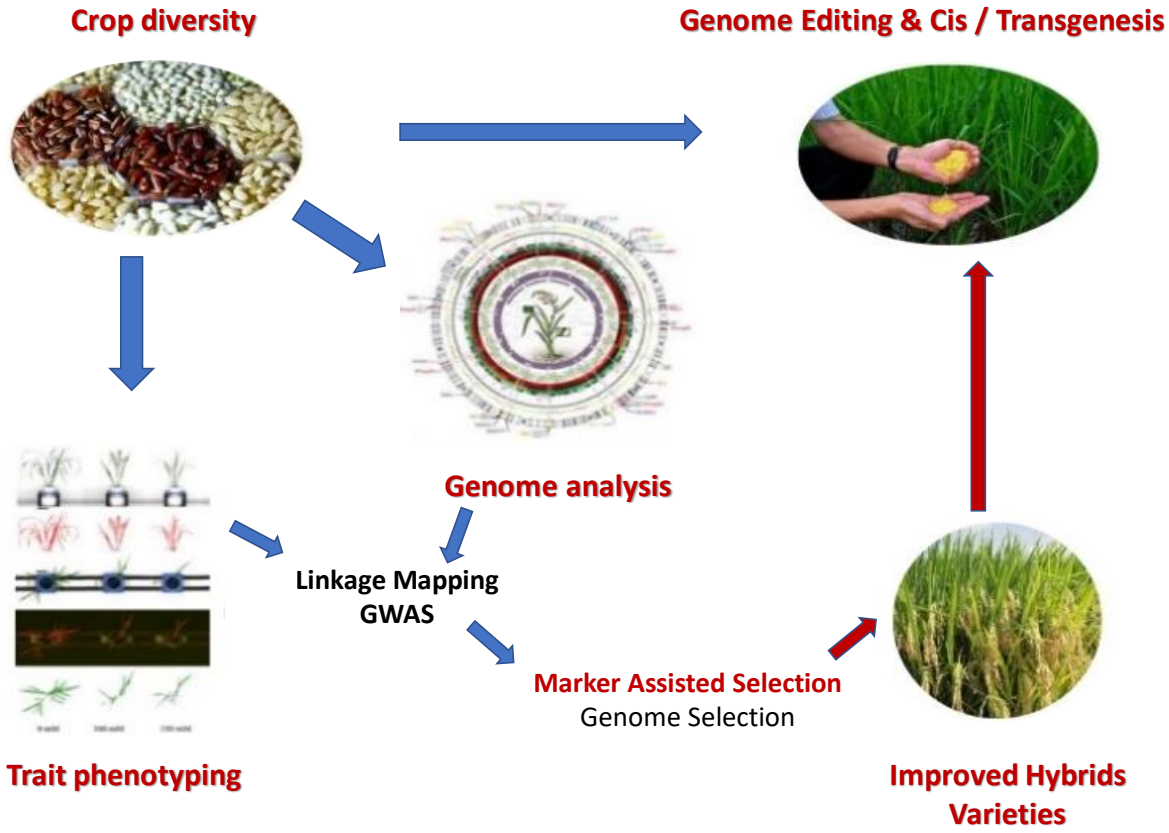


Fig. 16. Plant improvement

This knowledge is also fundamental for the progress of new plant diagnostic tools. Characteristics which are considered of primary importance are, pathogen and abiotic stress resistance, quality traits for plant, and reproductive traits defining output. A genome program can now be envisioned as an extremely important tool for plant improvement.

Such an approach to determine key genes and understand their function will result in a "quantum leap" in plant improvement. Additionally, the capability to explore gene expression will let us realize how plants react to and interact with the physical environment and management practices.

This information, together with suitable technology, may provide predictive measures of plant health and quality and become an essential part of future plant breeding solution management systems.

Current genome programs produce a large amount of information which will require processing, storage and alignment to the multinational research society. The data incorporate not only sequence information, but information on mutations, markers, maps, functional discoveries, etc. Key objectives for plant bioinformatics include: to favor the submission of all sequence data into the public domain, by repositories, to supply rational annotation of genes, proteins and phenotypes, and to make relationships both within the plants' data and between plants and other organisms.

Genotype building experiments

In the last few years an increasing amount of data for the DNA polymorphism and sequencing was collected in different plant varieties and cultivars. Most of this data was used for the goal of recognition of various cultivars as well as for their comparison of distances and analogy. This kind of distances are measured by the polymorphism on a part of the chromosome with unknown function.

This kind of polymorphism is widely used in the genomic learning through the species. The information for the polymorphism are analyzed for a potential link with a quantitative characteristic of interest of the particular phenotypes. As such a link is discovered it is called indirect marker. Indirect markers are closely linked, occasionally they may overlap, with a locus which identify this quantitative characteristic, QTL.

QTLs are determined as genes or regions of chromosomes which affect a particular trait. QTLs by themselves are very difficult to be recognized. In both cases this data, or as it is called, markers, can be used in further selection goals. This selection process is named as MAS.

QTLs (Quantitative Trait Locus) analysis and mapping

QTLs and mapping: The main problem is to determine which populations are appropriate for QTL-analyses, unstructured and F2 crosses and in plant - large scale populations in order to screen for potential QTLs. Because selection is based most on markers, higher density of mapping is extremely important. The interval between marker and QTL of about 5 centi Morgans (cM) seemed enough for effective selection. However, the simulation studies indicate that selection precision dropped down to 81% and 74% with 2 cM and 4 cM distance compared to 1cM (Fig. 17).

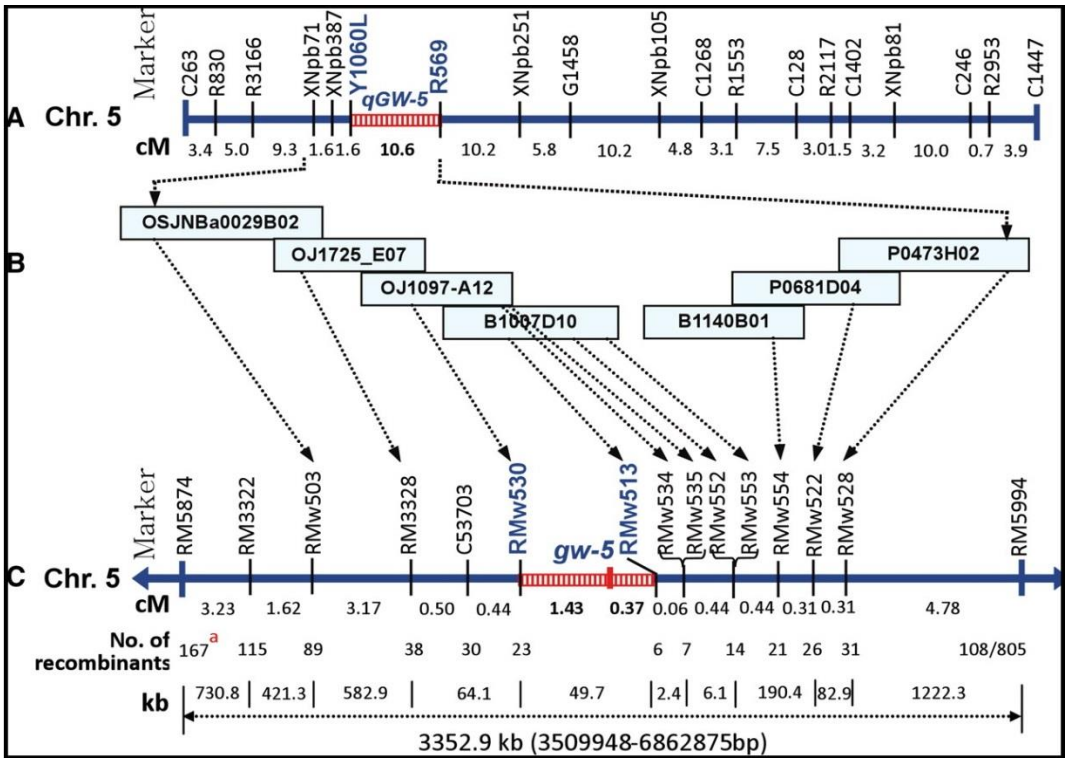


Fig. 17. QTL mapping of the *qGW-5* locus

Use of QTL information: It is supposed to be that some but not all loci are determined, so selection should be based on the combination of phenotypic and molecular data; in the process of selection, the link of markers and traits could reduce so this link should be observed over the generations; in the process of selection, QTLs demonstrate contemporaneous existence of the wanted genes in a line; in crossbred programs, QTLs could predict the efficiency of untested crosses, including their non-additive effect on the data of the parent lines and restricted number of crosses.

Future prospective: With cumulation of molecular data genotype building programs will be elaborated which will define homozygous desirable markers; in introgression programs for combining the intended traits from two lines in one; finally, the real world of agriculture is on the stage of accumulation of molecular information.

Analytical approaches: One of the statistical tools for making the QTL analyses such is the meta-analysis, which synthesize solid QTL data and improve the QTL position. A program of this class is the French BioMercator. Also PlaNet, the European plant genome database network, which is available at is an environment with complex research opportunities.

Further progress and detailed discussion on QTLs involves the statistical aspects of MAS, setting up the threshold of importance of marker effects, overestimation or deviation in estimation of QTL effects, optimization of selection programs for various generations with concomitant using of MAS and phenotypic data. A particular feature is that discovery should be made on plant specific parts, leaves, roots, fruits etc., as it was proved for the grapes.

Experimental results not all the time verify the efficiency of MAS as regards to the genotype building. The major reason is insufficient accuracy of the primary assessment of a QTL, its place and effect. Also some QTLs could be lost in the genotype building process. For complex productivity traits the epistatic waste would be a reason for changes in the value of QTL effect in the parent and progeny generation. Then it is recommended that election is based on the allelic combinations rather on the separate QTLs. It is in accordance to the numerous GxE interactions and with the selection within the environment of interest in the case of disease/drought resistance. Therefore, efficiency of MAS will

depend on the complexity of species/trait genetic architecture, on the progress of the characteristic in the environment and on their interaction.

For complex traits the assessment of QTLs should be in different environments. Also phenotypic evaluation/check over the consistent generations is absolutely necessary. For example: drought resistance seemed to be more complex trait vs. disease resistance.

From the economics point of view the use of markers will cost collection of DNA, genotyping, analyses, and discovery of QTLs etc. This high value is paid for the genotype building for characteristics which are expensive for evaluation, disease resistance, or characteristics with low heritability.

References

Angellotti M.C., Bhuiyan S.B., Chen G. And Wan Xiu-Feng (2007) Nucleic Acids Research, 35, W132-W136.

Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. Science. 2011; 333: 601-607.

Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002; 324: 1018-1022.

Betz FS, Hammond BG, Fuchs RL. Safety and advantages of Bacillus thuringiensis-protected plants to control insect pests. Regul Toxicol Pharmacol. 2000; 32: 156-173.

Bevivino A, Paganin P, Bacci G, Florio A, Pellicer MS, Papaleo MC, Mengoni A, Ledda L, Fani R, Benedetti A. Soil Bacterial community response to differences in agricultural management along with seasonal changes in a mediterranean region. 2014.

Blanchfield J. Genetically modified food crops and their contribution to human nutrition and food quality. J Food Science. 2004, 69(1):CRH28-CRH30.

Blum A. Plant breeding for stress environments1988: CRC Press, Inc.

Boserup E. The conditions of agricultural growth: The economics of agrarian change under population pressure 2005: Transaction Publishers.

Boyle G. Renewable energy2004: OXFORD university press.

Carbonetto B, Rascovan N, Álvarez R, Mentaberry A, Vázquez MP. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage systems in Argentine Pampas. 2014.

Conway GR, Barbier EB. After the green revolution: sustainable agriculture for development. Routledge 2013.

Cusick ME, Klitgord N, Vidal M, Hill DE. Interactome: gateway into systems biology. Hum Mol Genet. 2005; 14 Spec No.

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, et al. PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res. 2008; 36: D959-965.

Edwards D, Batley J. Plant genome sequencing: applications for crop improvement. Plant Biotechnol J. 2010; 8: 2-9.

Ellegren H. Genome sequencing and population genomics in nonmodel organisms. Trends Ecol Evol. 2014;29(1):51–63.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30: 1575-1584.

Evenson RE, Gollin D. Assessing the impact of the green revolution, 1960 to 2000. Science. 2003; 300: 758-762.

Fouts DE, Szpakowski S, Purushe J, Torralba M, Waterman RC, MacNeil MD, Alexander LJ, Nelson KE. Next generation sequencing to define prokaryotic and fungal diversity in the bovine rumen. 2012.

German JB, Hammock BD, Watkins SM. Metabolomics: building on a century of biochemistry to guide human health. Metabolomics. 2005; 1: 3-9.

Graham RD, Welch RM. Breeding for staple food crops with high micronutrient density 1996: Intl Food Policy Res Inst.

Grattapaglia D, Plomion C, Kirst M, Sederoff RR. Genomics of growth traits in forest trees. Curr Opin Plant Biol. 2009; 12: 148-156.

Hack C, Kendall G. Bioinformatics: Current practice and future challenges for life science education. Biochem Mol Biol Educ. 2005; 33: 82-85.

Iovene M, Barone A, Frusciante L, Monti L, Carputo D. Selection for aneuploid potato hybrids combining a low wild genome content and resistance traits from Solanum commersonii. Theor Appl Genet. 2004;109(6):1139–46.

Kale U.K., Bhosle S.G., Manjari G.S., Joshi M., Bansode S. and Kolaskar A.S. (2006) BMC Bioinformatics, S12-S27.

Kearsey MJ (1998) The principles of QTL analysis (a minimal mathematics approach). Journal of Experimental Botany, 49(327): 1619-1623.

Lewis W.A. Theory of economic growth. Vol. 7. 2013: Routledge.

Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13: 2178-2189.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alfoldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476–82.

Ma JKC, Drake PMW, Christou P. The production of recombinant pharmaceutical proteins in plants. Nat Rev Genet. 2003;4(10):794–805.

Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011; 470: 198-203.

Mendes LW, Kuramae EE, Navarrete AA, van Veen JA, Tsai SM. Taxonomical and functional microbial community selection in soybean rhizosphere. The ISME journal. 2014;8(8):1577–87.

Mohammadi SA and Prasanna BM (2003) Analysis of Genetic Diversity in Crop Plants—Salient Statistical Tools and Considerations. Crop Science, 43: 1235-1248.

Morgante M and Salamini F. (2003) From plant genomics to breeding practice. Current Opinion in Biotechnology, 14: 214-219.

Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nat Rev Genet. 2012;13(2):85–96.

Morsy M, Gouthu S, Orchard S, Thorneycroft D, Harper JF, Mittler R, et al. Charting plant interactomes: possibilities and challenges. Trends Plant Sci. 2008; 13: 183-191.

Nestel P, Bouis HE, Meenakshi JV, Pfeiffer W. Biofortification of staple food crops. J Nutr. 2006; 136: 1064-1067.

Organization EPS. European plant science: a field of opportunities. J Exp Bot. 2005;56(417):1699–709.

Orr HA. (2005) The genetic theory of adaptation: a brief history. Nature Review Genetics, 6: 119-127.

Ouzounis CA. Rise and demise of bioinformatics? Promise and progress. PLoS Comput Biol. 2012; 8: e1002487.

Paine JA, Shipton CA, Chaggar S, Howells RM, Kennedy MJ, Vernon G, et al. Improving the nutritional value of Golden Rice through increased pro-vitamin A content. Nat Biotechnol. 2005; 23: 482-487.

Pan Y, Cassman N, de Hollander M, Mendes LW, Korevaar H, Geerts RH, van Veen JA, Kuramae EE. Impact of long-term N, P, K, and NPK fertilization on the composition and potential functions of the bacterial community in grassland soil. FEMS Microbiol Ecol. 2014;90(1):195–205.

Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011; 52: 413-435.

Pingali PL. Green revolution: impacts, limits, and the path ahead. Proc Natl Acad Sci U S A. 2012; 109: 12302-12308.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, et al. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell. 2009; 21: 3718-3731.

Randhawa MS. Green Revolution: John Wiley and Sons. 1974

Rastogi G, Coaker GL, Leveau JH. New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches. FEMS Microbiol Lett. 2013;348(1):1–10.

Reif JC, Melchinger AE and Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Science, 45: 1-7.

Rhee SY, Dickerson J, Xu D. Bioinformatics and its applications in plant biology. Annu Rev Plant Biol. 2006; 57: 335-360.

Samson F, Brunaud V, Balzergue S, Dubreucq B, Lepiniec L, Pelletier G, et al. FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of Arabidopsis thaliana T-DNA transformants. Nucleic Acids Res. 2002; 30: 94-97.

Samson F, Brunaud V, Duchêne S, De Oliveira Y, Caboche M, Lecharny A, et al. FLAGdb++: a database for the functional analysis of the Arabidopsis genome. Nucleic Acids Res. 2004; 32: D347-350.

Sen S and Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics, 159, 371-387.

Souza RC, Hungria M, Cantão ME, Vasconcelos ATR, Nogueira MA, Vicente VA. Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. Appl Soil Ecol. 2015;86:106–12.

Svizzero S, Tisdell C. The Neolithic Revolution and human societies: diverse origins and development paths. School of Economics. University of Queensland. 2014.

Taiz L. Agriculture, plant physiology, and human population growth: past, present, and future. Theoretical and Experimental Plant Physiology. 2013; 25: 167-181.

Thompson GA, Goggin FL. Transcriptomics and functional genomics of plant defence induction by phloem-feeding insects. J Exp Bot. 2006; 57: 755-766.

Tsuru T. and Kobayashi I. (2008 Molecular Biology Evolution, 25, 2457-2473.

Tuberosa R, Salvi S. Genomics-based approaches to improve drought tolerance of crops. Trends Plant Sci. 2006; 11: 405-412.

Turner JA. A realizable renewable energy future Science. 1999; 285: 687-689.

Van Borm S, Belák S, Freimanis G, Fusaro A, Granberg F, Höper D, King DP, Monne I, Orton R, Rosseel T. Next-generation sequencing in veterinary medicine: how can the massive amount of information arising from high-throughput technologies improve diagnosis, control, and management of infectious diseases? In: Veterinary infection biology: molecular diagnostics and high-throughput strategies. Berlin: Springer; 2015. p. 415–36.

van der Vlugt R, Minafra A, Olmos A, Ravnikar M, Wetzel T, Varveri C, Massart S. Application of next generation sequencing for study and diagnosis of plant viral diseases in agriculture. 2015.

Wall PK, Leebens-Mack J, Müller KF, Field D, Altman NS, dePamphilis CW. PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Res. 2008; 36: D970-976.

Walsh B (2001) Quantitative genetics in the age of genomics. Theoretical Population Biology, 59: 175-184.

Weigel D, Mott R. The 1001 genomes project for Arabidopsis thaliana. Genome Biol. 2009;10(5):107.

Wilson SA, Roberts SC. Metabolic engineering approaches for production of biochemicals in food and medicinal plants. Curr Opin Biotechnol. 2014;26:174–82.

Wishart DS. Current progress in computational metabolomics. Brief Bioinform. 2007; 8: 279-293.

Xu Y. Molecular plant breeding2010: CABI.

Yang DT, X. Zhu. Modernization of agriculture and long-term growth. Journal of Monetary Economics, 2013; 60: 367-382.

Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN Jr. Plants to power: bioenergy to fuel the future. Trends Plant Sci. 2008;13(8):421–9.

Zeder MA. 13 Agricultural origins in the ancient world. Anthropology Explored: The Best of Smithsonian AnthroNotes, 2013.

Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, Li J, Simianer H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One. 2014;9(3):e93017.

Application of system biology in bioremediation

Aysel Çağlan GÜNAL
GAZI University
Ankara, Turkey
http://gazi.edu.tr

Aysel Gamze YÜCEL IŞILDAR
GAZI University
Ankara, Turkey
http://gazi.edu.tr

Rabia Sarikaya
GAZI University
Ankara, Turkey
http://gazi.edu.tr

Contents

Introduction

Environmental pollutants have become a major global concern, given their undesirable recalcitrant and xenobiotic compounds. A variety of polycyclic aromatic hydrocarbons (PAHs), xenobiotics, chlorinated and nitro-aromatic compounds were depicted to be highly toxic, mutagenic and carcinogenic for living organisms.

Some of the sources of these contaminants are; chemical (dying, agriculture, pharmaceuticals, etc) petrochemical (oil rafineries, fuel spills), metal (iron and steel industry, shipbuilding, etc.) energy (power plants), mining industries and water supply and sewage works. These contaminants have impacts on nature. While various physico-chemical processes have been developed for treating these

pollutants; these approaches are often prohibitively expensive, non-specific, or have the potential for introducing secondary contamination. However, microbial population may also degrade the pollutant and considered as one of the environment friendly and cost-effective method for restorationof ecological niches contaminated with chemical pollutants. As a result, there has been an increased interest in eco-friendly bio-based treatments commonly known as bioremediation. Though bioremediation has been used to varying degrees for more than 60 years, for example petroleum land farming, it historically has been implemented as a very 'black box' engineering solution whereamendments are added and the pollutants are degraded. This approach is often successful but all to often the results are less than desirable, that is, no degradation of the contaminant or even production of more toxic daughter products. The key to successful bioremediation is to harness the naturally occurring catabolic capability of microbes to catalyze transformations of environmental pollutants.

Bioremediation

Bioremediation is the exploitation of biological activities for mitigation (and wherever possible complete elimination) of the noxious effects caused by environmental pollutants in given sites. If the process occurs in the same place afflicted by pollution then an in situ bioremediation scenario occurs. In contrast, deliberate relocation of the contaminated material (soil and water) into a different place to intensify biocatalysis originates an ex situ case. In bioremediation, microorganisms with biological activity, including algae, bacteria, fungi, and yeast, can be used in their naturally occurring forms.

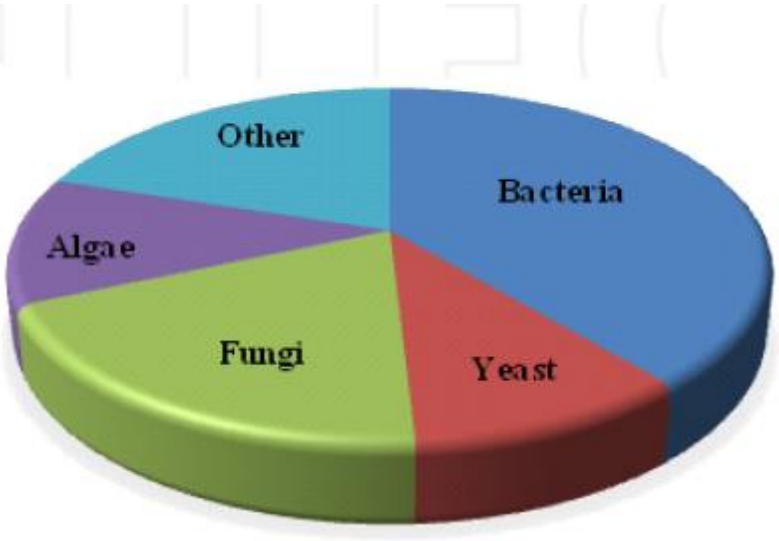


Figure 1. Types of microorganisms used in bioremediation processes (Coelho et al. 2015).

Figure 1 shows the main types of microorganisms used in these processes, based on a search for papers reporting microorganisms and bioremediation studies the microorganisms that have been most commonly used are bacteria and fungi, although yeast and algae are also frequently applied

Types of organisms used in bioremediation

Typically, bioremediation is based on the cometabolism action of one organism or a consortium of microorganisms. In this process, the transformation of contaminants presents a little efficiency or no benefit to the cell, and therefore this process is described as nonbeneficial biotransformation. Several studies have shown that many organisms (prokaryotes and eukaryotes) have a natural capacity to biosorb toxic heavy metal ions. Examples of microorganisms studied and strategically used in bioremediation treatments for heavy metals include the following: (1) bacteria: *Arthrobacter* spp., *Pseudomonas veronii* (Vullo et al. 2008), *Burkholderia* spp., *Kocuria flava*, *Bacillus cereus* and *Sporosarcina*; (2) fungi: *Penicillium canescens*, *Aspergillus versicolor*, and *Aspergillus fumigatus*; (3) algae: *Cladophora fascicularis*, *Spirogyra* spp. and *Cladophora* spp. and *Spirogyra* spp. and *Spirullina* spp and (4) yeast: *Saccharomyces cerevisiae* and *Candida utilis*. Prokaryotes (bacteria and archaeans) are distinguished from eukaryotes (protists, plants, fungi, and animals). The cellular structure of eukaryotes is characterized by the presence of a nucleus and other membrane-enclosed organelles. Also, the ribosomes in prokaryotes are smaller (70S) than in eukaryotes (80S). The way in which microorganisms interact with heavy metal ions is partially dependent on whether they are eukaryotes or prokaryotes, wherein eukaryotes are more sensitive to metal toxicity than prokaryotes. The possible modes of interaction are (a) active extrusion of metal, (b) intracellular chelation (in eukaryotes) by various metal-binding peptides, and (c) transformation into other chemical species with reduced toxicity. For bioremediation to be effective, microorganisms must enzymatically attack the pollutants and convert them to harmless products. Bacteria and higher organisms have developed mechanisms associated with resistance to toxic metals and rendering them innocuous. Several microbes, including aerobes, anaerobes, and fungi, are involved in the enzymatic degradation process. Most of bioremediation systems are run under aerobic conditions, but anaerobic conditions make it possible microbial organisms to degrade otherwise recalcitrant molecules. Because several different types of pollutants can be present at a contaminated site, various types of microorganisms are required for effective remediation. Some types of microorganism are able to degrade petroleum hydrocarbons and use them as a source of carbon and energy. However, the choice of the organisms employed is variable, depending on the chemical nature of the polluting agents, and needs to be selected carefully as they only survive in the presence of a limited range of chemical contaminants. The efficiency of the degradation process is related to the potential of the particular microorganism to introduce molecular oxygen into the hydrocarbon and to generate the intermediates that subsequently enter the general energy yielding metabolic pathway of the cell. Some bacteria search the contaminant and move toward it because they flexibly exhibit the potential as a chemotactic response. Numerous microorganisms can utilize oil as a source of food, and many of them produce potent surface-active compounds that can emulsify oil in water and facilitate its removal. Bacteria that can degrade petroleum products include species of *Pseudomonas*, *Aeromonas*, *Moraxella*, *Beijerinckia*, *Flavobacteria*, *Chrobacteria*, *Nocardia*, *Corynebacteria*, *Modococci*, *Streptomyces*, *Bacilli*, *Arthrobacter*, *Aeromonas*, and cyanobacteria and some yeasts. For example, *Pseudomonas putida* MHF 7109 can be isolated from cow dung microbial consortia for the biodegradation of selected petroleum hydrocarbon compounds, such as benzene, toluene, and o-xylene (BTX).

Bioremediation strategies

In many cases the clean-up contaminated sites have been carried out using physical and chemical methods such as immobilization, removal (dig and dump), thermal, and solvent treatments.

However, advances in biotechnology have seen the development of biological methods of contaminant degradation and removal, a process known as bioremediation. Potentially bioremediation is cheaper than the chemical and physical options, can deal with lower concentrations of contaminants more effectively, although the process may take longer.

The strategies for bioremediation in both soil and water can be as follows:

- Use the indigenous microbial population
- Encourage the indigenous population
- Bioaugmentation; the addition of adapted or designed inoculants
- Addition of genetically modified micro-organisms
- Phytoremediation

If the process occurs in the same place afflicted by pollution then an in situ bioremediation scenario occurs. In contrast, deliberate relocation of the contaminated material (soil and water) into a different place to intensify biocatalysis originates an ex situ case.

In situ and ex situ methods

Bioremediation technologies can be broadly classified as ex situ and in situ. Ex situ technologies are those treatments which involve the physical removal of the contaminated material for treatment process.

If the process occurs in the same place afflicted by pollution, then an in situ bioremediation scenario occurs. These techniques are generally the most desirable options due to lower cost and less disturbance since they provide the treatment in place avoiding excavation and transport of contaminants. *In situ* treatment is limited by the depth of the soil that can be effectively treated. In many soils, effective oxygen diffusion for desirable rates of bioremediation extend to a range of only a few centimeters to about 30 cm into the soil, although depths of 60 cm and greater have been effectively treated in some cases. The most important land treatments are:

Bioventing is the most common *in situ* treatment and involves supplying air and nutrients through wells to contaminated soil to stimulate the indigenous bacteria. Bioventing employs low air flow rates and provides only the amount of oxygen necessary for the biodegradation while minimizing volatilization and release of contaminants to the atmosphere. It works for simple hydrocarbons and can be used where the contamination is deep under the surface.

In situ biodegradation involves supplying oxygen and nutrients by circulating aqueous solutions through contaminated soils to stimulate naturally occurring bacteria to degrade organic contaminants. It can be used for soil and groundwater. Generally, this technique includes conditions such as the infiltration of water-containing nutrients and oxygen or other electron acceptors for groundwater treatment.

Biosparging involves the injection of air under pressure below the water table to increase groundwater oxygen concentrations and enhance the rate of biological degradation of contaminants by naturally occurring bacteria. Biosparging increases the mixing in the saturated zone and there- by increases the contact between soil and groundwater. The ease and low cost of installing small-diameter air injection points allows considerable flexibility in the design and construction of the system.

Bioaugmentation. Bioremediation frequently involves the addition of microorganisms indigenous or exogenous to the contaminated sites. Two factors limit the use of added microbial

cultures in a land treatment unit: 1) nonindigenous cultures rarely compete well enough with an indigenous population to develop and sustain useful population levels and 2) most soils with long-term exposure to biodegradable waste have indigenous microorganisms that are effective degrades if the land treatment unit is well managed.

***Ex situ* bioremediation** deliberate relocation of the contaminated material (soil and water) into a different place to intensify biocatalysis originates an ex situ case. These techniques involve the excavation or removal of contaminated soil from ground.

Landfarming is a simple technique in which contaminated soil is excavated and spread over a pre- pared bed and periodically tilled until pollutants are degraded. The goal is to stimulate indigenous biodegradative microorganisms and facilitate their aerobic degradation of contaminants. In general, the practice is limited to the treatment of superficial 10–35 cm of soil. Since landfarming has the potential to reduce monitoring and maintenance costs, as well as clean-up liabilities, it has received much attention as a disposal alternative.

Composting is a technique that involves combining contaminated soil with nonhazardous organic amendants such as manure or agricultural wastes. The presence of these organic materials supports the development of a rich microbial population and elevated temperature characteristic of composting.

Biopiles are a hybrid of landfarming and composting. Essentially, engineered cells are con- structed as aerated composted piles. Typically used for treatment of surface contamination with petroleum hydrocarbons they are a refined version of landfarming that tend to control physical losses of the contaminants by leaching and volatilization. Biopiles provide a favorable environment for indigenous aerobic and anaerobic microorganisms.

Bioreactors. Slurry reactors or aqueous reactors are used for *ex situ* treatment of contaminated soil and water pumped up from a contaminated plume. Bioremediation in reactors involves the processing of contaminated solid material (soil, sediment, sludge) or water through an engineered containment system. A slurry bioreactor may be defined as a containment vessel and apparatus used to create a three-phase (solid, liquid, and gas) mixing condition to increase the bioremediation rate of soil-bound and water-soluble pollutants as a water slurry of the contaminated soil and biomass (usually indigenous microorganisms) capable of degrading target contaminants. In general, the rate and extent of biodegradation are greater in a bioreactor system than *in situ* or in solid-phase systems because the contained environment is more manageable and hence more controllable and predictable. Despite the advantages of reactor systems, there are some disadvantages. The contaminated soil requires pretreatment (e.g., excavation) or alternatively the contaminant can be stripped from the soil via soil washing or physical extraction (e.g., vacuum extraction) before being placed in a bioreactor. Table 1 summarizes the bioremediation strategies.

Table 1. Summary of bioremediation strategies

Technology	Examples	Benefits	Limitations	Factors to consider
In situ	*In situ* bioremediation Biosparging Bioventing Bioaugmentation	Most cost efficient Noninvasive Relatively passive Natural attenuation processes Treats soil and water	Environmental constraints Extended treatment time Monitoring difficulties	Biodegradative abilities of indigenous microorganisms Presence of metals and other inorganics Environmental parameters Biodegradability of pollutants Chemical solubility Geological factors Distribution of pollutants
Ex situ	Landfarming Composting Biopiles	Cost efficient Low cost Can be done on site	Space requirements Extended treatment time Need to control abiotic loss Mass transfer problem Bioavailability limitation	See above
Bioreactors	Slurry reactors Aqueous reactors	Rapid degradation kinetic Optimized environmental parameters Enhances mass transfer Effective use of inoculants and surfactants	Soil requires excavation Relatively high cost capital Relatively high operating cost	See above Bioaugmentation Toxicity of amendments Toxic concentrations of contaminants

Advantages and disadvantages of bioremediation

Advantages

- Bioremediation is a natural process and is therefore perceived by the public as an acceptable waste treatment process for contaminated material such as soil. Microbes able to degrade the contaminant increase in numbers when the contaminant is present; when the contaminant is degraded, the biodegradative population declines. The residues for the treatment are usually harmless products and include carbon dioxide, water, and cell biomass.
- Theoretically, bioremediation is useful for the complete destruction of a wide variety of contaminants. Many compounds that are legally considered to be hazardous can be transformed to harmless products. This eliminates the chance of future liability associated with treatment and disposal of contaminated material.
- Instead of transferring contaminants from one environmental medium to another, for example, from land to water or air, the complete destruction of target pollutants is possible.
- Bioremediation can often be carried out on site, often without causing a major disruption of normal activities. This also eliminates the need to transport quantities of waste off site and the potential threats to human health and the environment that can arise during transportation.
- Bioremediation can prove less expensive than other technologies that are used for clean-up of hazardous waste.

Disadvantages

- Bioremediation is limited to those compounds that are biodegradable. Not all compounds are susceptible to rapid and complete degradation.
- There are some concerns that the products of biodegradation may be more persistent or toxic than the parent compound.
- Biological processes are often highly specific. Important site factors required for success include the presence of metabolically capable microbial populations, suitable environmental growth conditions, and appropriate levels of nutrients and contaminants.
- It is difficult to extrapolate from bench and pilot-scale studies to full-scale field operations.
- Research is needed to develop and engineer bioremediation technologies that are appropriate for sites with complex mixtures of contaminants that are not evenly dispersed in the environment.

  Contaminants may be present as solids, liquids, and gases.
- Bioremediation often takes longer than other treatment options, such as excavation and removal of soil or incineration.
- There is no accepted definition of “clean”, evaluating performance of bioremediation is difficult, and there are no acceptable endpoints for bioremediation treatments (Vidali, 2001).

Environmental factors for bioremediation

Nutrients

Although the microorganisms are present in contaminated soil, they cannot necessarily be there in the numbers required for bioremediation of the site. Their growth and activity must be stimulated. Biostimulation usually involves the addition of nutrients and oxygen to help indigenous microorganisms. These nutrients are the basic building blocks of life and allow microbes to create the necessary enzymes to break down the contaminants. All of them will need nitrogen, phosphorous, and carbon. Carbon is the most basic element of living forms and is needed in greater quantities than other elements. In addition to hydrogen, oxygen, and nitrogen it constitutes about 95% of the weight of cells.

Phosphorous and sulfur contribute with 70% of the remainders. The nutritional requirement of carbon to nitrogen ratio is 10:1, and carbon to phosphorous is 30:1.

Environmental requirements

Microbial growth and activity are readily affected by pH, temperature, and moisture. Although microorganisms have been also isolated in extreme conditions, most of them grow optimally over a narrow range, so that it is important to achieve optimal conditions. If the soil has too much acid it is possible to rinse the pH by adding lime. Temperature affects biochemical reactions rates, and the rates of many of them double for each 10 °C rise in temperature. Above a certain temperature, however, the cells die. Plastic covering can be used to enhance solar warming in late spring, summer, and autumn. Available water is essential for all the living organisms, and irrigation is needed to achieve the optimal moisture level. The amount of available oxygen will determine whether the system is aerobic or anaerobic. Hydrocarbons are readily degraded under aerobic conditions, whereas chlorurate compounds are degraded only in anaerobic ones. To increase the oxygen amount in the soil it is possible to till or sparge air. In some cases, hydrogen peroxide or magnesium peroxide can be introduced in the environment. Soil structure controls the effective delivery of air, water, and nutrients. To improve soil structure, materials such as gypsum or organic matter can be applied. Low soil permeability can impede movement of water, nutrients, and oxygen; hence, soils with low permeability may not be appropriate for *in situ* clean-up techniques.

Influence of environmental factors on biodegradation

Earlier studies of bioremediation trials were not performed under natural environmental conditions. Therefore, the impact of environmental factors on the bioremediation process was never expected. However, after the investigation of in situ bioremediation approaches now it is feasible to understand the bioremediation process is influenced significantly by environmental factors such as the physiological and chemical ambience of the contaminated environment, bioavailability of nutrients, concentration and properties of co-contaminants, level of contamination, community organization of the indigenous microbial communities. Various abiotic and biotic factors play important role in

bioremediation. Their dynamic interactions occur in concrete abiotic conditions which are defined by physico-chemical conditions like O_2 supply, electron transport, water, temperature, pH, salt concentration, many of which. The above environmental factors determine the dynamic of endogenous microbial community structures along with the availability of given chemical and energy source.

The factors at play in bioremediation scenarios include more elements than just the biological catalysts and the contaminants discussed above. Their dynamic interactions occur in concrete abiotic settings which are defined by a whole of physico-chemical conditions: O_2 tension, electron acceptors, water, temperature, granulation, and others, many of which change over time and the course of the catalysis. Such abiotic conditions determine the species composition of the endogenous microbial communities as much as (or more than) the availability of given chemical species as C and energy source. Bioremediation is a case of multiscale complexity which is not amenable to the typically reductionist approaches (e.g. one compound, one strain, and one pathway) that have dominated many studies on biodegradation. How to overcome this impasse?

Since microbes are the drivers of bioremediation, shifts in the composition and activity of a microbial community may impact the fate of a contaminant in the environment Recent studies have employed next-generation sequencing approaches to better understand the microbial communities involved in various bioremediation interventions. These approaches have greatly expanded our understanding of the microbial processes involved in bioremediation as well as the impact of various response strategies for contaminant cleanup. The use of molecular biology and metagenomics has also greatly expanded our understanding of the biological systems found in these contaminated environments and in many cases have greatly enhanced our understanding of the microbial world. Here, we seek to provide a key background on metagenomic approaches and summarize how these tools have been employed to understand contaminated environments in an effort to inform the best practices for environmental cleanup.

Bioremediation requires the integration of huge amounts of data from various sources: chemical structure and reactivity of organic compounds; sequence, structure and function of proteins (enzymes); comparative genomics; environmental microbiology; and so on.

Systems biology

The process of bioremediation employs a microbial community to clean up an environmental contaminant. The rates of contaminant detoxification are dependent on a number of factors including the composition of the native microbial community, the environmental conditions, and the nature of the contaminant. Therefore, optimization of bioremediation requires combining complex variables together to understand and predict the fate of environmental contaminants. stems biology—the study of the systematic properties and dynamic interactions in a biological system has been employed to understand complex biological systems and how they will respond to various perturbations. A systems biology approach to understanding environmental systems and bioremediation can be employed to investigate complex environmental microbial communities and the environmental constraints on contaminant degradation.

There is need to in silico study for predicting the possible degradation pathways by using various computational tools. There are large number of databases and computer programs available to perform the computational analysis for assisting the development and implementation of microbial bioremediation. The huge data from biology mainly in the form of DNA, RNA and protein sequences is putting heavy demand on computers and computational scientists. Systems biology is an integrated

research approach to study complex biological systems, by investigating interactions and networks at the molecular, cellular, community, and ecosystem levels. A systems biology approach is being adopted to unravel key processes to understand, optimize, predict and evaluate microbial function and survival strategies in the ecosystem of interest. To use a systems biology approach to bioremediation projects they must involve the characterization of microbial community composition, cellular and molecular activity and are complicated by the presence of toxic chemicals that alters the normal behavior of the microbial community.

Some important components of systems biology are the use of computational approaches to develop a predictive understanding of the systems response to a perturbation and understanding contaminant remediation as it combines many levels of a system to predict the fate of environmental contaminants.

It is strongly believed that there are three dimensions for the effectiveness of vital bioremediation process; that is, chemical landscape (nutrients-to-be, electron donors/acceptors and stressors) abiotic landscape, and catabolic landscape of which only the catabolic landscape is genuinely biological. The chemical landscape has a dynamic interplay with the biological interventions on the abiotic background of the site at stake. This includes humidity, conductivity, temperature, matrix conditions, redox status, etc.

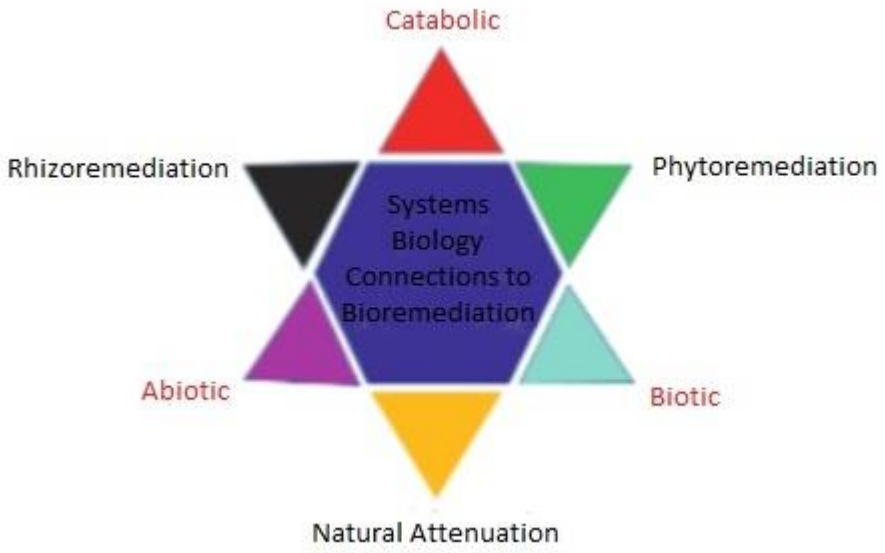


Figure 2. Systems biology connections to bioremediation (Koehmel et al. 2016)

To gain an understanding of complex in situ bioremediation processes, monitoring techniques that inventory and monitor terminal electron acceptors and electron donors, enzyme probes that measure functional activity in the environment, functional genomic microarrays, phylogenetic microarrays, metabolomics, proteomics, and quantitative PCR can provide unprecedented insights into the key microbial reactions employed (Figure 3). In general terms, an ecosystem consists of communities, populations, cells, protein, RNA, and DNA. We can analyze DNA, RNA, and protein at the cellular levels to understand the impacts on the cells, and analyze community and populations to understand effect of bioremediation on structure/function relationships (Figure 3).

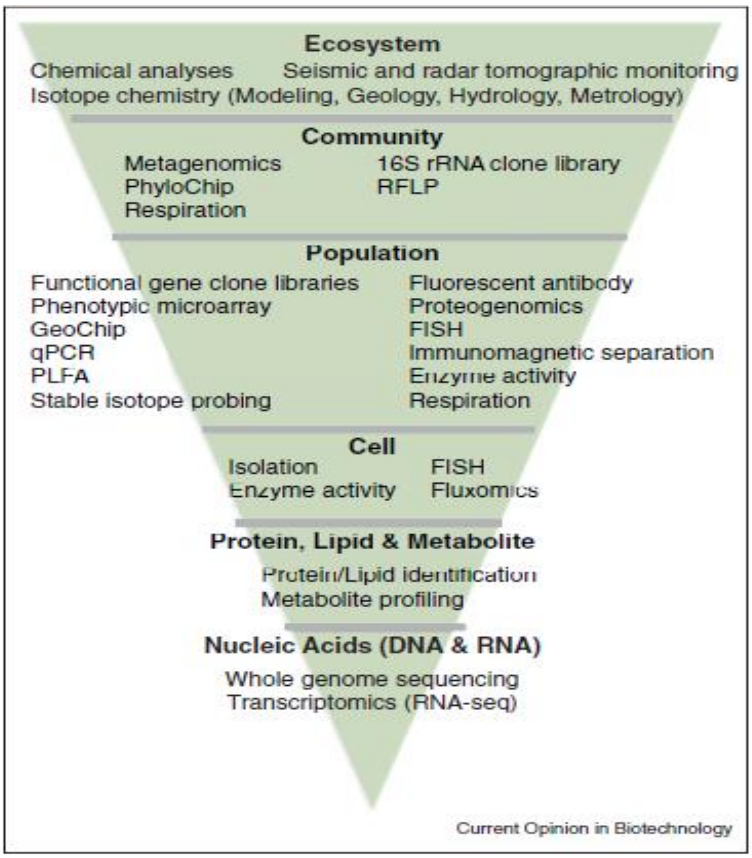


Figure 3. Systems biology from molecules to ecosystems

A system-level understanding of a biological system can be derived from insight into four key properties:

1) System structures. These include the network of gene interactions and biochemical pathways, as well as the mechanisms by which such interactions modulate the physical properties of intracellular and multicellular structures.

2) System dynamics. How a system behaves over time under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as phase portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors. Bifurcation analysis traces time-varying change(s) in the state of the system in a multidimensional space where each dimension represents a particular concentration of the biochemical factor involved.

3) The control method. Mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease.

4) The design method. Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error.

Progress in any of the above areas requires breakthroughs in our understanding of computational sciences, genomics, and measurement technologies, and integration of such discoveries with existing knowledge.

Omics approaches are central to systems biology. Metagenomics—the analysis of the total genomic content of a microbial community—has been widely applied to understanding microbial communities in environmental systems (Figure 4). Other 'omics techniques, including metatranscriptomics (community RNA analysis) and metaproteomics (community protein analysis), have been more recently applied to environmental microbial communities.

Multiple approaches can be applied to understanding different levels of a microbial community. Each of these techniques investigates a particular biological molecule (DNA, RNA, or Protein) thorough analysis of each of these molecules extracted from an environmental community yields key insights into the taxonomic composition a community, the functional potential of a community, or the genes and proteins currently being expressed Techtman and Hazen, 2016.

Metagenomics

Genomic is a powerful computer technology used to understand the structure and function of all genes in an organism based on knowing the organism's entire DNA sequence. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts. Metagenomics is the study of the genomes in a microbial community and constitutes the first step to studying the microbiome. Metagenomics allows us to investigate the composition of a microbial community. Genomic studies consider the genetic material of a specific organism, while metagenomics (meta meaning beyond) refers to studies of genetic material of entire communities of organisms. This process usually involves nextgeneration sequencing (NGS) after the DNA is extracted from the samples. NGS produces a large volume of data in the form of short reads, from which a microbial community profile or other information can be pieced together just like gathering information from the pieces of a puzzle. Although whole-metagenome sequencing (WMS) provides a partial glimpse into the functional profile of a microbial community, it is better inferred using metatranscriptomics, which involves sequencing the complete (meta)transcriptome of the microbial community. Metagenomics provides access to the functional gene composition of microbial communities and thus gives a much broader description than phylogenetic surveys, which are often based only on the diversity of one gene, for instance the 16S rRNA gene. On its own, metagenomics gives genetic information on potentially novel biocatalysts or enzymes, genomic linkages between function and phylogeny for uncultured organisms, and evolutionary profiles of community function and structure. It can also be complemented with metatranscriptomic or metaproteomic approaches to describe expressed activities. Metagenomics is also a powerful tool for generating novel hypotheses of microbial function; the remarkable discoveries of proteorhodopsin-based photoheterotrophy or ammonia-oxidizing Archaea attest to this fact. The rapid and substantial cost reduction in next-generation sequencing has dramatically accelerated the development of sequence-based metagenomics. In fact, the number of metagenome shotgun sequence datasets has exploded in the past few years. In the future, metagenomics will be used in the same manner as 16S rRNA gene fingerprinting methods to describe microbial community profiles. It will therefore become a standard tool for many laboratories and scientists working in the field of microbial ecology.

Metagenomic approaches often take two forms—targeted metagenomics or shotgun metagenomics (Figure 4). In targeted metagenomics—or microbiomics—the diversity of a single gene is probed to identify the full complement of sequences of a particular gene in an environment. Targeted metagenomics is most often employed to investigate both the phylogenetic diversity and relative abundance of a particular gene in a sample. This approach is regularly used to investigate the diversity of small subunit rRNA sequences (16S/18S rRNA) in a sample. Microbial ecologists routinely use

small subunit rRNA sequencing to understand the taxonomic diversity of an environment. It can also be applied as a tool to investigate the impact of environmental contaminants in altering microbial community structure. To perform targeted metagenomics, environmental DNA is extracted, and the gene of interest is PCR amplified using primers designed to amplify the greatest diversity of sequences for that gene of interest. The strength of targeted metagenomics is that it provides a fairly comprehensive catalog of the microbial taxa present in a set of samples and allows for in-depth comparison of shifts in microbial diversity before and after a perturbation.

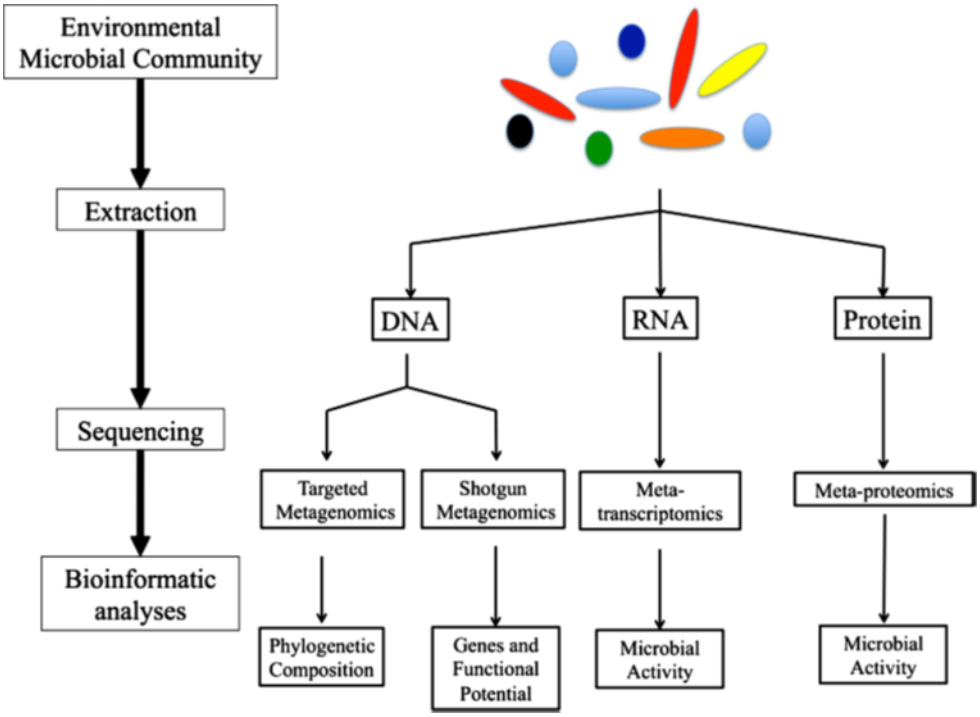


Figure 4. Metagenomic approaches to understanding microbial communities.

In shotgun metagenomics, the total genomic complement of an environmental community is probed through genomic sequencing (Figure 4). In this approach, environmental DNA is extracted and then fragmented to prepare sequencing libraries. These libraries are then sequenced to determine the total genomic content of that sample. Shotgun metagenomics is a powerful technique where the functional potential of a microbial community can be identified.

Shotgun metagenomics is often most limited by the depth of sequencing. Microarray-based techniques have been developed. PhyloChip and GeoChip are the two most commonly used microarray technologies. PhyloChip is a 16S rRNA-based microarray able to probe the diversity of 10,993 sub-families in 147 phyla (Hazen et al. 2010). GeoChip is a functional gene microarray able to probe the diversity of 152,414 genes from 410 gene categories. Microarray techniques are not dependent on the depth of sequencing to provide comprehensive insights into the microbial community. They also have the advantage of providing rigorous annotation for the various taxa/genes present on the chip alleviatingthe limitation of the need for good homologs in the database to achieve accurate classification. Microarray-based approaches are, however, limited in that only the genes on the chip can be detected, thus limiting the potential for discovery of new genes or pathways in a sample.

Microarray- based approaches are often a helpful complement to sequencing-based approaches as an additional line of evidence.

Metatranscriptomics-metaproteomics-metabolomics

Using a proteomics approach, the physiological changes in an organism during bioremediation provide further insight into bioremediation-related genes and their regulation.. Metatranscriptomics and metaproteomics are increasingly being applied to environmental systems (Figure 4). These approaches provide key insights into the actively expressed genes in a microbial community and are thus good indicators for the microbial functions being expressed under the conditions at the time of sampling. In metatranscriptomics, RNA is extracted from an environmental sample. The RNA is converted into cDNA and sequenced in a similar fashion to metagenomics (Figure 4). This approach provides an inventory of the actively expressed genes in a sample. Metaproteomics does not involve nucleic acid sequencing, but rather high-resolution mass spectrometry combined with enzymatic digests of proteins and liquid chromatography. Metaproteomics provides insights into the complement of proteins found in an environmental sample including posttranslational modifications in proteins that may impact their activity.

By focusing on what genes are expressed by the entire microbial community, metatranscriptomics sheds light on the active functional profile of a microbial community. The metatranscriptome provides a snapshot of the gene expression in a given sample at a given moment and under specific conditions by capturing the total mRNA. As for metagenomics, it is now possible to perform whole metatranscriptomics shotgun sequencing. This (meta)genome-wide expression provides the expression and functional profile of a microbiome. When processing reads, a typical metatranscriptomics analysis pipeline will either (1) map reads to a reference genome or (2) perform de novo assembly of the reads into transcript contigs and supercontigs. The first strategy, in a manner similar to the alignment-based methods in WMS, maps reads to reference databases, thus gathering information to infer the relative expression of individual genes. The second strategy infers the same but with assembled sequences. The first strategy is limited by the information in the database of reference genomes. The second strategy is limited by the ability of software programs to assemble contigs and supercontigs correctly from short reads data. tools and techniques. The application of metatranscriptomics to the study of the microbiome is far less common relative to other omics reviewed in this article. Most analysis pipelines described in the literature were built ad hoc. The majority of these methods follow the aforementioned first strategy based on read mapping.

Metabolomics is the comprehensive analysis by which all metabolites of a sample (small molecules released by the organism into the immediate environment) are identified and quantified. The metabolome is considered the most direct indicator of the health of an environment or of the alterations in homeostasis (i.e. dysbiosis). Variation in the production of signature metabolites are related to changes in activity of metabolic routes, and therefore, metabolomics represents an applicable approach to pathway analysis. Additionally, the application of metabolomics for drug discovery and pharmacogenomics represents a promising avenue for personalized medicine. The metabolomic profile associated with the microbiome may show a strong dependence on environmental factors (e.g. diet, exposure to xenobiotics, and environmental stressors), providing valuable information not just about the characteristics of the microbiome but also about the interactions of the microbial community with the host environment. Thus, metabolomics aims to improve our understanding of the role of the microbiome in the transformation of nutrients and pollutants as well as other abiotic factors that may affect the homeostasis of the host environment. The analysis pipeline for spectral metabolomic data

involves three steps: (1) preprocessing, (2) statistical analysis, and (3) machine learning techniques for pattern recognition. In the first step, denoising and peak-picking improve the quality of the data to be processed.

Several *in silico* softwares, pipelines, web resources and algorithms have been developed to interpret or correlate molecular and x-omics data. Nonetheless, bioinformatic resources of bioremediation are still scarce. The University of Minnesota Biocatalysis/Biodegradation Database (UMBBD) has enlisted 200 pathways, 1350 reactions, 1195 compounds, >1000 enzymes, 491 microorganism entries and 259 biotransformation rules encompassing microbial bioremediation (http://umbbd.msi.umn.edu/) (Gao et al. 2011). Metarouter is yet another system for maintaining heterogeneous information related to bioremediation and biodegradation in a framework that allows updating query modifications (Desai et al. 2010). The system can be accessed and administrated through a web interface (Pazos et al. 2005). Other software platforms re: Kyoto Encyclopedia of Genes and Genomes (KEGG) at http://www.genome.ad.jp/kegg/kegg.html. (Moriya et al. 2010); Boehringer Mannhein Biochemical Pathways (BMBP) on the ExPASy server, Switzerland (http://www.expasy.org/cgi-bin/search-biochem-index); International Society for the Study of Xenobiotics (http://www.issx.org); PathDB; Methabolic Pathways Database at NCGR (http://www.ncgr.org/Pathdb/) etc.

Existing computational database, software and tools and their collective integration will help to determine the environmental fate of any compounds more precisely and accurately.

Practical Applications

Radionuclide biotransformation

Groundwater and soil at the Area 3 FRC site in Oak Ridge is not only contaminated with Uranium (up to 200 mM), but poses a unique bioremediation problem due to its low pH (3), high nitrate (200 mM), and high calcium concentrations along with presence of chlorinated organic solvents. Research at this site by various investigators exemplifies successful application of systems biology tools to reveal a deeper understanding of the microbiology at play in the subsurface. Previously, 16S clone library-based community analysis during an in situ biostimulation test at this site have identified *Desulfovibrio,Geobacter*, *Anaeromyxobacter*, *Desulfosporosinus*, *Acidovorax*, and *Geothrix* spp. present concomitant with U(VI) reduction (Cardenas et al. 2008). Clone libraries of functional gene markers like dsrAB, nirK, nirS, amoA, and pmoA showed high microbial diversity in functional genes. However, recent metagenomic analysis from well FW106 specifically using a random shotgun sequencing-based strategy revealed a highly enriched community dominated by denitrifying b-Proteobacteria and g-Proteobacteria. Geo-Chip analysis of several groundwater monitoring wells reported widespread diversity of dsrAB genes, which showed that sulfate-reducing bacteria were key players in U(VI) reduction. During the U(VI) reoxidation phase as studied in a sediment column with samples from FRC, observed decrease in biomass, but increase in microbial activity. Using the PhyloChip, the study showed no decline in *Geobacter* or *Geothrix* spp. during the reoxidation phase, but members of *Actinobacteria*, *Firmicutes*, *Acidobacteria*, and *Desulfovibrionaceae* exhibited increased abundance. GeoChip analysis during the reoxidation phase from field samples showed a decline in dsr genes but reoxidation did not appear to effect microbial functional diversity suggesting that the microbial community was able to recover and continue to reduce U(VI) in the post oxidation phase

Metals bioimmobilization

The Hanford 100H area adjacent to the Columbia River in Washington is contaminated with Chromium (Cr) as a result of being a weapons production site. In 2004, Hydrogen Release Compound HRCtm was injected in an effort to mediate sustained bioimmobilization of Cr(VI) in situ by stimulating indigenous microbial flora Hubbard et al. (2008) used time-lapse seismic and radar tomographic geophysical monitoring to determine spatiotemporal distribution of the injected HRC and biogeochemical transformations associated with Cr(VI) bioremediation post injection of HRC. Direct cell counts revealed that while cell numbers reached 108 cells/ml, Cr(VI) levels decreased from 100 ppb to below background levels within a year. PhyloChip analysis showed enrichment of sulfate reducers along with nitrate reducers, iron reducers, and methanogenic populations during this time. Targeted enrichments resulted in isolation of sulfate-reducing *Desulfovibrio vulgaris* like strain RCH1, nitrate reducing strain *Pseudomonas stutzeri* strain RCH2, and iron-reducing strain *Geobacter metallireducens* strain RCH3, all capable of Cr(VI) reduction. mFlowFISH (integrated fluorescence in situ hybridization and flow cytometry) analysis was able to detect and sort *Pseudomonads* similar to strain RCH2 directly from Hanford 100H field water samples collected in 2009 and 2010.

Hydrocarbon bioremediation

The dependence of petroleum-based energy source has fueled industrial growth and prosperity. However, it also brought dispersal of hydrocarbons into different environments. Fortunately, the organic nature of hydrocarbons enables microbes to metabolize these petroleum compounds as substrates. Notable reviews on a systems biology approach to bioremediation are Atlas and Hazen (2011), Harayama et al. (2004), Zhou et al. (2011), Fredrickson et al. (2008), de Lorenzo (2008) and Chakraborty et al. (2012). The MC252 oil spill in the Gulf of Mexico in 2010 was the largest in US history. Many environmental factors distinguished this spill from previous ones, including hydrocarbon composition, environmental variables, depth of the spill, and the availability of systems biology tools. Information on chemical analyses is crucial in support of a system's biology approach for oil bioremediation in the MC252 spill. While Camilli et al. (2010) concluded that microbial respiration rates within the deep plume were extremely low based on dissolved oxygen concentration, measurement of microbial respiration rates, enzyme activity, phosphate concentration, and polar membrane lipid concentration in surface water affected by the oil spill. Edwards et al. (2011) concluded that enzyme activities and respiration rates were found to be higher inside the oil slick. Valentine et al. (2010) investigated the fate of methane, propane, and ethane gases of the deep hydrocarbon plume at depth greater than 799 m, and found that propane and ethane were degraded faster than methane.13C-labled substrates, as well as 13C and 3H tracers, were used to measure d13C-DIC. In another study, methane was found to be the most abundant hydrocarbon released during the MC252 spill, and that there was a rapid response of methanotrophic bacteria rapidly respiring the released methane. PhyloChip, clone library, GeoChip, phospholipid fatty acid (PLFA), and isotope chemistry were used to compare microbial communities inside and outside the deep plume (Hazen et al. 2010). The results identified Oceanospirillales, which were found to degrade hydrocarbons at 58°C inside the plume. The GeoChip demonstrated genes that were significantly correlated to concentration of oil contaminants, such as phdC1 (naphthalene degradation), and alkB (oxidation of alkanes), as well as a shift in C, N, P, S cycling processes in the deep plume samples. The involvement of federal agencies and pending lawsuits is the impetus for a concerted effort in collating all data collected resulting in a comprehensive database useful for researchers. By integrating chemical analyses with studies utilizing a systems

biology approach, there was an unprecedented near real-time understanding of chemical and biological reactions involved in the hydrocarbon degradation. In order to gain a more comprehensive understanding of the microbiological processes, data from transcriptomics studies will provide information on whether the cultivatable dominant microbes are the in situ active ones, and proteomics studies will identify enzymes central to hydrocarbon degradation.

Chlorinated solvents bioremediation

Chlorinated solvents, such as TCE and dichloroethene (DCE), are recalcitrant carcinogenic compounds that persist in the environment once released. Microbes, such as *Dehalococcoides*, are capable of using the chlorinated solvents as electron acceptors anaerobically and dechlorinating the compounds to ethene. Another biodegradation pathway is the aerobic co-metabolism of the chlorinated compounds to carbon dioxide and chloride by microbes such as methane-oxidizers with methane monooxygenases (MMOs) (. Descriptions of techniques that monitor mass loss, geochemical fingerprints, isotope fractionation associated with biodegradation, microbial communities in biostimulation and natural attenuation studies, quantitative real-time PCR methods targeting reductive dehalogenase genes are included in several reviews. Between 1955 and 1972, low-level radioactive isotopes, sewage and chlorinated solvents were injected into the aquifer through a 95 m deep well at Test Area North (TAN) in Idaho National Laboratory. The plume contained TCE concentrations ranging from 5 ppb to 300 ppm extending for more than 2 km. An enhanced in situ bioremediation pilot study started in 1999 to treat the chlorinated solvents contaminated groundwater by injecting the electron donor Lactate to stimulate in situ reductive dechlorination. A comparison of microbial communities in the core and groundwater samples was assessed by characterizing total biomass, PLFA analysis, culturing and community-level physiological profiling (CLPP) using Biolog GN microplates (Lehman et al. 2004). DGGE analysis indicated that wells with high concentrations of chlorinated solvents had different microbial communities from wells with minimal concentrations of the contaminants, and that attached, and the free-living microbes had different functional and composition profile Additionally, qPCR of the *Dehalococcoides* sp. 16S rRNA genes provided the most convincing result in quantifying dechlorinating potential of a community compared to community analysis by terminal restriction fragment length polymorphism (T-RFLP), and RFLP analysis with clone sequencing. Erwin et al. (2005) demonstrated the presence of bacteria harboring MMOs and potential of TCE co-metabolism at TAN from a pristine area using PCR amplification to generate a function gene fragment library and sequencing. Stable carbon isotope ratios of groundwater samples taken in 2000 confirmed the complete conversion of TCE to ethene, and minimal biodegradation of t-DCE (Song et al. 2002). Using the PhyloChip for bacterial composition characterization, a decrease in reductive dechlorinating organisms and an increase in methane-oxidizing microbes capable of aerobic co-metabolism of TCE was observed. Further studies that would complement the investigation at the TAN site would be to employ a shotgun proteomics approach as reported by Werner et al. (2009) Their method allowed for detection of peptides, such as FdhA, TceA, PceA, and HupL that could potentially be used as bioindicators of chlorinated ethene dehalorespiration.

References

Achal V., Pan X., Fu Q., Zhang D. Biomineralization based remediation of As (III) contaminated soil by Sporosarcina ginsengisoli. Journal of Hazardous Materials 2012; 201–202, 178–184.

Achal V., Pan X., Zhang D. Remediation of copper-contaminated soil by Kocuria flava CR1, based on microbially induced calcite precipitation. Ecological Engineering 2011; 37 (10) 1601–1605.

Alivisatos AP, Blaser MJ, Brodie EL, Chun M, Dangl JL, Donohue TJ, Dorrestein PC, Gilbert JA, Green JL, Jansson JK, Knight R, Maxon ME, McFall-Ngai MJ, Miller JF, Pollard KS, Ruby EG, Taha SA (2015) A unified initiative to harness Earth's microbiomes. Science 350:507–508. doi:10.1126/science.aac8480

Atlas RM, Hazen TC: Oil biodegradation and bioremediation: a tale of the two worst spills in US history. Environ Sci Technol 2011, 45:6709-6715.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF: Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 2000, 289(5486):1902-1906.

Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, Hazen TC, Richardson PM, Herman DJ, Tokunaga TK et al.: Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl Environ Microbiol 2006, 72:6288-6298

Camilli R, Reddy CM, Yoerger DR, Van Mooy BAS, Jakuba MV, Kinsey JC, McIntyre CP, Sylva SP, Maloney JV: Tracking hydrocarbon plume transport and biodegradation at Deepwater Horizon. Science 2010, 330:201-204.

Cardenas E, Wu W-M, Leigh MB, Carley J, Carroll S, Gentry T, Luo J, Watson D, Gu B, Ginder-Vogel M et al.: Microbial communities in contaminated sediments, associated with bioremediation of uranium to submicromolar levels. Appl Environ Microbiol 2008, 74:3718-3729.

Chakraborty R, Wu CH, Hazen TC (2012) Systems biology approach to bioremediation. Curr Opin Biotechnol 23:1–8.

Conrad ME, Brodie EL, Radtke CW, Bill M, Delwiche ME, Lee MH, Swift DL, Colwell FS: Field evidence for co-metabolism of trichloroethene stimulated by addition of electron donor to groundwater. Environ Sci Technol 2010, 44:4697-4704.

Coulon F, McKew BA, Osborn AM, McGenity TJ, Timmis KN (2007) Effects of temperature and biostimulation on oil-degrading microbial communities in temperate estuarine waters. Environ Microbiol 9: 177-186.

Cupples AM: Real-time PCR quantification of Dehalococcoides populations: methods and applications. J Microbiol Methods 2008, 72:1-11.

de Lorenzo V (2008) Systems biology approaches to bioremediation. Curr Opin Biotechnol 19:579–589.

Deng L., Su Y., Su H., Wang X., Zhu X. Sorption and desorption of lead (II) from wastewater by green algae Cladophora fascicularis. Journal of Hazardous Materials 2007; 143 (1–2) 220–225.

Desai C, Pathak H, Madamwar D (2010) Advances in molecular and ‘‘-omics” technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites. Biores Technol 101:1558–1569.

Edwards BR, Reddy CM, Camilli R, Carmichael CA, Longnecker K, Van Mooy BAS: Rapid microbial respiration of oil from the Deepwater Horizon spill in offshore surface waters of the Gulf of Mexico. Environ Res Lett 2011, 6:035301.

Erwin DP, Erickson IK, Delwiche ME, Colwell FS, Strap JL, Crawford RL: Diversity of oxyenase genes from methane- and ammonia-oxidizing bacteria in the Eastern Snake River Plain aquifer. Appl Environ Microbiol 2005, 71:2016-2025.

Eyers L, Smoot JC, Smoot LM, Bugli C, Urakawa H, et al. (2006) Discrimination of shifts in a soil microbial community associated with TNT-contamination using a functional ANOVA of 16S rRNA hybridized to oligonucleotide microarrays. Environ Sci Technol 40: 5867-5873.

F. M. von Fahnestock, G. B. Wickramanayake, K. J. Kratzke, W. R. Major. Biopile Design, Operation, and Maintenance Handbook for Treating Hydrocarbon Contaminated Soil, Battelle Press, Columbus, OH (1998).

Faybishenko B, Hazen TC, Long PE, Brodie EL, Conrad ME, Hubbard SS, Christensen JN, Joyner D, Borglin SE, Chakraborty R et al.: In situ long-term reductive bioimmobilization of Cr(VI) in groundwater using hydrogen release compound. Environ Sci Technol 2008, 42:8478-8485.

Fields MW, Bagwell CE, Carroll SL, Yan T, Liu X, Watson DB, Jardine PM, Criddle CS, Hazen TC, Zhou J: Phylogenetic and functional biomakers as indicators of bacterial community responses to mixed-waste contamination. Environ Sci Technol 2006, 40:2601-2607.

Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL et al.: Towards environmental systems biology of Shewanella. Nat Rev Microbiol 2008, 6:592-603.

Fulekar MH, Geetha M, Sharma J (2009) Bioremediation of Trichlorpyr Butoxyethyl Ester (TBEE) in bioreactor using adapted Pseudomonas aeruginosa in scale up process technique. Biol Med 1(3):1–6

Fulekar MH, Sharma J., (2008) Bioinformatics applied in bioremediation. Innovative Romanian Food Biotechnology. 2(2) 28-36.

Gao J, Ellis LBM, Wackett LP (2011) The University of Minnesota pathway prediction system: multi-level prediction and visualization. Nucleic Acids Res 39:W406–W411

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I: Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 2008, 3(8):e3042.

Han RY, Geller JT, Yang L, Brodie EL, Chakraborty R, Larsen JT, Beller HR: Physiological and transcriptional studies of Cr(VI) reduction under aerobic and denitrifying conditions by an aquifer-derived pseudomonad. Environ Sci Technol 2010, 44:7491-7497.

Harayama S, Kasai Y, Hara A: Microbial communities in oilcontaminated seawater. Curr Opin Biotechnol 2004, 15:205-214.

Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortney JL, Stringfellow WT, Bill M, Conrad ME, Tom LM, Chavarria KL, Alusi TR, Lamendella R, Joyner DC, Spier C, Baelum J, Auer M, Zemla ML, Chakraborty R, Sonnenthal EL, D’haeseleer P, Holman HYN, Osman S, Lu ZM, Van Nostrand JD, Deng Y, Zhou JZ, Mason OU

(2010) Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science 330:204–208. doi:10.1126/ Science.1195979

Hazen TC, Rocha AM, Techtmann SM (2013) Advances in monitoring environmental microbes. Curr Opin Biotech 24:526–533. doi:10.1016/J.Copbio.2012.10.020 11.

Hazen TC, Sayler GS (2016) Environmental systems microbiologyof contaminated environments. In: Yates M, Nakatsu C,Miller RSP (eds) Manual of environmental microbiology, vol 4th edn. ASM Press, Washington, DC, pp 5.1.6-1–5.1.6-10

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P et al.: GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J 2007, 1:67-77

Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S, Barry K, Tringe SG, Watson DB, He Z et al.: Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. ISME J 2010, 4:660-672

Hettich RL, Pan CL, Chourey K, Giannone RJ (2013) Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. Anal Chem 85:4203–4214. doi:10.1021/ac303053e

Hubbard SS, Williams K, Conrad ME, Faybishenko B, Peterson J,Chen JS, Long P, Hazen T: Geophysical monitoring of hydrological and biogeochemical transformations associated with Cr(VI) bioremediation. Environ Sci Technol 2008, 42:3757-3765.

Illman WA, Alvarez PJ: Performance assessment of bioremediation and natural attenuation. Crit Rev Environ Sci Technol 2009, 39:209-270.

Jiang C. Y., Sheng X. F., Qian M., Wang Q. Y Isolation and characterization of heavy metal resistant Burkholderia species from heavy metal contaminated paddy field soil and its potential in promoting plant growth and heavy metal accumulation in metal polluted soil. Chemosphere 2008; 72:157–164.

Kanmani P., Aravind J., Preston D. Remediation of chromium contaminants using bacteria. International Journal of Environmental Science ad Technology 2012; 9:183–193.

Katsivela E, Moore ER, Maroukli D, Strömpl C, Pieper D, et al. (2005) Bacterial community dynamics during in-situ bioremediation of petroleum waste sludge in landfarming sites. Biodegradation 16: 169-180.

Ken Killham; Jim I. Prosser. The prokaryotes. In: Paul, E. A. (ed.). Soil Microbiology, Ecology, and Biochemistry. Oxford: Elsevier: 2007. p119–144.

Kessler JD, Valentine DL, Redmond MC, Du MR, Chan EW, Mendes SD, Quiroz EW, Villanueva CJ, Shusta SS, Werra LM et al.: A persistent oxygen anomaly reveals the fate of spilled methane in the deep Gulf of Mexico. Science 2011, 331:312-315.

Khan F, Sajid M, Cameotra SS (2013) In Silico Approach for the Bioremediation of Toxic Pollutants. J Phylogenetics Evol Biol 4:161. doi:10.4172/2157-7463.1000161

Kitoni, H. (2002) Systems Biology: A Brief Overview Science .01 Mar 2002: Vol. 295, Issue 5560, pp. 1662-1664.

Klipp E, Liebermeister W, Wierling C, Kowald A, Herwig R(2016) Systems biology: a textbook. Wiley, New York.

Koehmel, J. Sebastian, A., Prasad, M. N. V. (2016) Advancing Bioremediation through systems biology and synthetic biology. Chapter 26. 677-680. In Bioremediation and Bioeconomy. Ed by M. N. V. Prasad. Elsevier, USA.

Kujan P., Prell A., Safár H., Sobotka M., Rezanka T., Holler P. Use of the industrial yeast Candida utilis for cadmium sorption. Folia Microbiologica. 2006; 51 (4) 257–260.

Kumar A., Bisht B. S., Joshi V. D., Dhewa T. Review on bioremediation of polluted environment: a management tool. International Journal of Environmental Sciences 2011; 1 (6) 1079–1093.

Kundu, D., Hazra, C., Chaudhari, A. Bioremediation of Nitroaromatics (NACs)- Based Explosives: Integrating '-omics' and unmined Microblome Richness (2014) Biological Remediation of Explosive Residues ed by. Singh, S. H. Springer. 179-199.

Leahy JG, Colwell RR (1990) Microbial degradation of hydrocarbons in the environment. Microbiol Rev 54: 305-315.

Lee Y. C., Chang S. P. The biosorption of heavy metals from aqueous solution by Spirogyra and Cladophora filamentous macroalgae. Bioresource Technology 2011; 102 (9) 5297–5304.

Lehman RM, O'Connell SP, Banta A, Fredrickson JK, Reysenbach AL, Kieft TL, Colwell FS: Microbiological comparison of core and groundwater samples collected from a fractured basalt aquifer with that of dialysis chambers incubated in situ. Geomicrobiol J 2004, 21:169-182.

Liu P, Meagher RJ, Light YK, Yilmaz S, Chakraborty R, Arkin AP, Hazen TC, Singh AK: Microfluidic fluorescence in situ hybridization and flow cytometry (mFlowFISH). Lab on a Chip 2011, 11:2673-2679.

Lovley DR (2003) Cleaning up with genomics: applying molecular biology to bioremediation. Nat Rev Microbiol 1:35–44.doi:10.1038/nrmicro731

Lu Z, Deng Y, Van Nostrand JD, He Z, Voordeckers J, Zhou A, Lee Y.-J., Mason OU, Dubinsky EA, Chavarria KL et al.: Microbial gene functions enriched in the Deepwater Horizon deep-sea oil plume. ISME J, doi:10.1038/ismej.2011.91.

Luciene M. Coelho, Helen C. Rezende, Luciana M. Coelho, Priscila A.R. de Sousa, Danielle F.O. Melo and Nívia M.M. Coelho (2015). Bioremediation of Polluted Waters Using Microorganisms, Advances in Bioremediation of Wastewater and Polluted Soil, Prof. Naofumi Shiomi (Ed.), InTech, DOI: 10.5772/60770. Available from: https://www.intechopen.com/books/advances-in-bioremediation-of-wastewater-and-polluted-soil/bioremediation-of-polluted-waters-using-microorganisms

Machado M. D., Soares E. V., Soares H. M. Removal of heavy metals using a brewer's yeast strain of Saccharomyces cerevisiae: chemical speciation as a tool in the prediction and improving of treatment efficiency of real electroplating effluents. Journal of Hazardous Materials 2010; 180(1–3) 347–353.

Mane P. C., Bhosle A. B. Bioremoval of some metals by living Algae spirogyra sp. and Spirullina sp. from aqueous solution. International Journal of Environmental Research 2012; 6(2) 571–576.

Mejáre M., Bülow L. Metal-binding proteins and peptides in bioremediation and phytoremediation of heavy metals. Trends in Biotechnology 2001; 19 (2) 67–73.

Mills DK, Fitzgerald K, Litchfield CD, Gillevet PM (2003) A comparison of DNA profiling techniques for monitoring nutrient impact on microbial community composition during bioremediation of petroleum-contaminated soils. J Microbiol Methods 54: 57-74.

Moreels D, Bastiaens L, Ollevier F, Merckx R, Diels L, et al. (2004) Effect of in situ parameters on the enrichment process of MTBE degrading organisms. Commun Agric Appl Biol Sci 69: 3-6.

Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, Kanehisa M (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res 38:W138–W143

Nicol GW, Schleper C: Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle? Trends Microbiol 2006, 14(5):207-212.

Nicolaou S. A., Gaida S. M., Papoutsakis E. T. A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: from biofuels and chemicals, to biocatalysis and bioremediation. Metabolic Engineering 2010; 12 (4) 307–331.

Palumbo AV, Schryver JC, Fields MW, Bagwell CE, Zhou JZ, Yan T, Liu X, Brandt CC: Coupling of functional gene diversity and geochemical data from environmental samples. Appl Environ Microbiol 2004, 70:6525-6534

Pandey J, Chauhan A, Jain RK (2009) Integrative approaches for assessing the ecological sustainability of in situ bioremediation. FEMS Microbiol Rev 33: 324-375.

Rahm BG, Chauhan S, Holmes VF, Macbeth TW, Sorenson KSJ, Alvarez-Cohen L: Molecular characterization of microbial populations at two sites with differing reductive dechlorination abilities. Biodegradation 2006, 17:523-534.

Ramasamy R. K., Congeevaram S., Thamaraiselvi K. Evaluation of isolated fungal strain from e-waste recycling facility for effective sorption of toxic heavy metal Pb (II) ions and fungal protein molecular characterization-a Mycoremediation approach. Asian Journal of Experimental Biological Sciences 2011; 2(2) 342–347.

Roane T. M., Josephson K. L., Pepper I. L. Dual-bioaugmentation strategy to enhance remediation of cocontaminated soil. Applied and Environmental Microbiology 2001; 67 (7) 3208–3215.

S.R. Gill, M. Pop, R.T. DeBoy, P.B. Eckburg, P.J. Turnbaugh, B.S. Samuel, J.I. Gordon, D.A. Relman, C.M. Fraser-Liggett, K.E. Nelson Metagenomic analysis of the human distal gut microbiome Science, 312 (2006), pp. 1355–1359.

Say R., Yimaz N., Denizli A. Removal of heavy metal ions using the fungus Penicillium canescens. Adsorption Science and Technology 2003; 21 (7) 643–650.

Scow KM, Hicks KA: Natural attenuation and enhanced bioremediation of organic contaminants in groundwater. Curr Opin Biotechnol 2005, 16:246-253.

Scragg, A. (2005) Bioremediation. In Environmental Biotechnology. Oxford. 173-229. USA.

Sharma S. Bioremediation: features, strategies and applications. Asian Journal of Pharmacy and Life Science 2012; 2 (2) 202–213.

Singh R., Singh P., Sharma R. Microorganism as a tool of bioremediation technology for cleaning environment: a review. Proceedings of the International Academy of Ecology and Environmental Sciences, 2014; 4(1) 1–6.

Song DL, Conrad ME, Sorenson KS, Alvarez-Cohen L: Stable carbon isotope fractionation during enhanced in situ bioremediation of trichloroethene. Environ Sci Technol 2002, 36:2262-2268.

Tastan B. E., Ertugrul S., Donmez G. Effective bioremoval of reactive dye and heavy metals by *Aspergillus versicolor*. Bioresource Technology 2010; 101(3) 870–876.

Techtmann, S. M., Hazen, T. C. (2016) Metagenomic applications in environmental monitoring and bioremediation J Ind Microbiol Biotechnol (2016) 43:1345–1354.

Thapa B., Kumar A., Ghimire A. A Review on bioremediation of petroleum hydro- carbon contaminants in soil. Kathmandu University Journal of Science, Engineering and Technology 2012; 8 (1) 164–170.

V. Desjardin, R. Bayard, N. Huck, A. Manceau, R. Gourdon Effect of microbial activity on the mobility of chromium in soils Waste Manag, 22 (2002), pp. 195–200.

Valentine DL, Kessler JD, Redmond MC, Mendes SD, Heintz MB, Farwell C, Hu L, Kinnaman FS, Yvon-Lewis S, Du MR et al.: Propane respiration jump-starts microbial response to a deep oil spill. Science 2010, 330:208-211.

Van Nostrand JD, Wu W-M, Wu L, Deng Y, Carley J, Carroll S, He Z, Gu B, Luo J, Criddle CS et al.: GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. Environ Microbiol 2009, 11:2611-2626.

Vidali M (2001) Bioremediation. An overview. Pure Appl Chem 73: 1163–1172.

Vullo D. L., Ceretti H. M., Daniel M. A., Ramírez S. A., Zalts A. Cadmium, zinc and copper biosorption mediated by Pseudomonas veronii 2E. Bioresource Technology 2008; 99 (13) 5574–5581.

Waldron PJ, Wu L, Nostrand JDV, Schadt CW, He Z, Watson DB, Jardine PM, Palumbo AV, Hazen TC, Zhou J: Functional gene array-based analysis of microbial community structure in groundwaters with a gradient of contaminant levels. Environ Sci Technol 2009, 43:3529-3534.

Wasilkowski D., Swedziol Ż., Mrozik A. The applicability of genetically modified microorganisms in bioremediation of contaminated environments. Chemik 2012; 66 (8) 822–826.

Wenderoth DF, Rosenbrock P, Abraham WR, Pieper DH, Höfle MG (2003) Bacterial community dynamics during biostimulation and bioaugmentation experiments aiming at chlorobenzene degradation in groundwater. Microb Ecol 46: 161-176.

Werner JJ, Ptak AC, Rahm BG, Zhang S, Richardson RE: Absolute quantification of Dehalococcoides proteins: enzyme bioindicators of chlorinated ethene dehalorespiration. Environ Microbiol 2009, 11:2687-2697.

Wilmes P, Bond PL: Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 2006, 14(2):92-97.

Y. Hu, C. Fu, Y. Yin, G. Cheng, F. Lei, X. Yang, J. Li, E. Ashforth, L. Zhang, B. Zhu Construction and preliminary analysis of a deep-sea sediment metagenomic fosmid library from Qiongdongnan Basin, South China Sea Mar Biotechnol, 12 (2010), pp. 719–727.

Zhou AF, He ZL, Qin YJ, Lu ZM, Deng Y, Tu QC, Hemme CL, Van Nostrand JD, Wu LY, Hazen TC, Arkin AP, Zhou JZ (2013) StressChip as a high-throughput tool for assessing microbial community responses to environmental stresses. Environ Sci Technol 47:9841–9849. doi:10.1021/es4018656

Zhou JZ, He Q, Hemme CL, Mukhopadhyay A, Hillesland K, Zhou AF, He ZL, Van Nostrand JD, Hazen TC, Stahl DA et al.: How sulphate-reducing microorganisms cope with stress: lessons from systems biology. Nat Rev Microbiol 2011, 9:452-466.