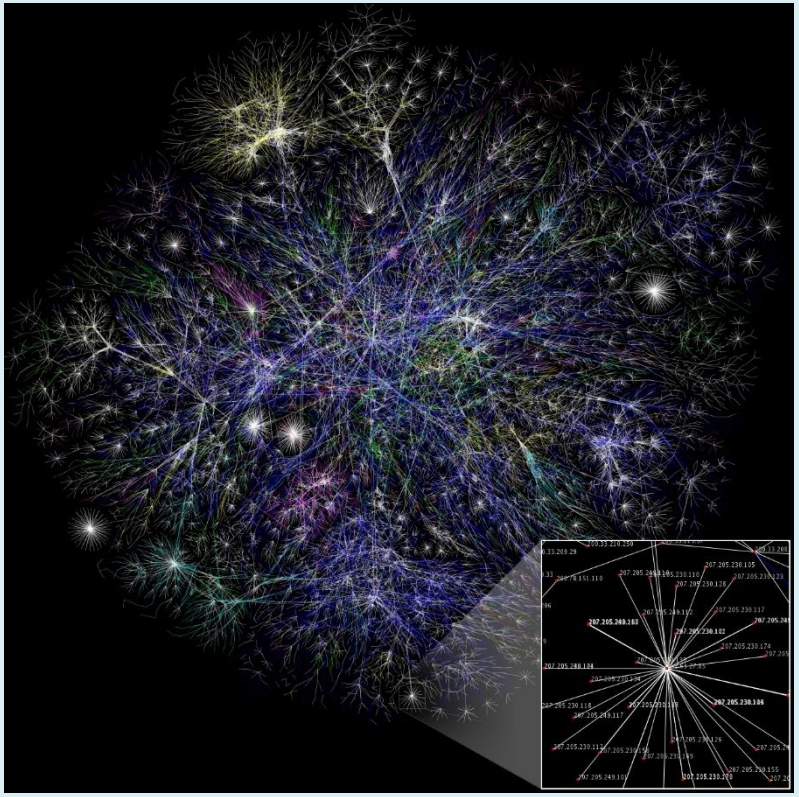
Introduction to Bioinformatics

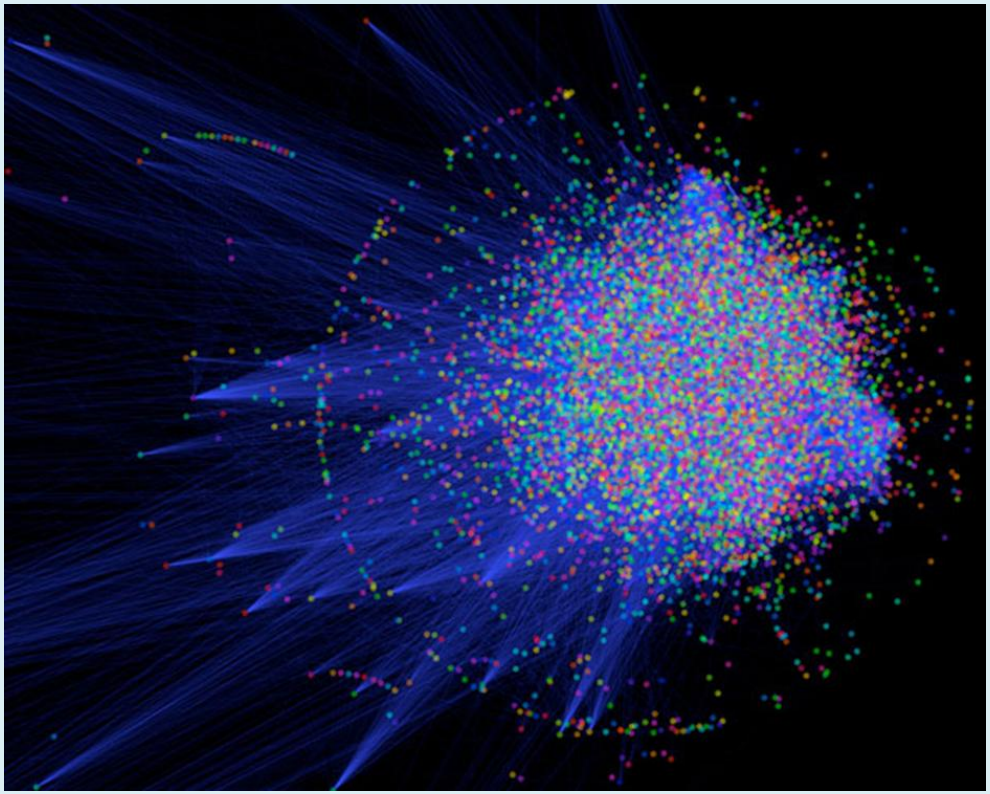
Introduction

- Learning objectives:
  - After studying these materials you should be able to do the following:
    - define the terms bioinformatics;
    - explain the scope of bioinformatics;

Introduction



- The connectivity of the internet (from the Wikipedia entry for "internet")



- A map of human protein interactions (from the Wikipedia entry for "Protein–protein interaction").

- We seek to understand biological principles on a genome-wide scale using the tools of bioinformatics.

… What is Bioinformatics?...

- Bioinformatics
  - the study of how information is represented and transmitted in biological systems, starting at the molecular level

  is a discipline that need a computer.
  - An ink pen and a supply of traditional laboratory notebooks could be used to record results of experiments.
  - However, to do so would be like foregoing the use of a computer and word-processing program in favor of pen and paper to write a novel.

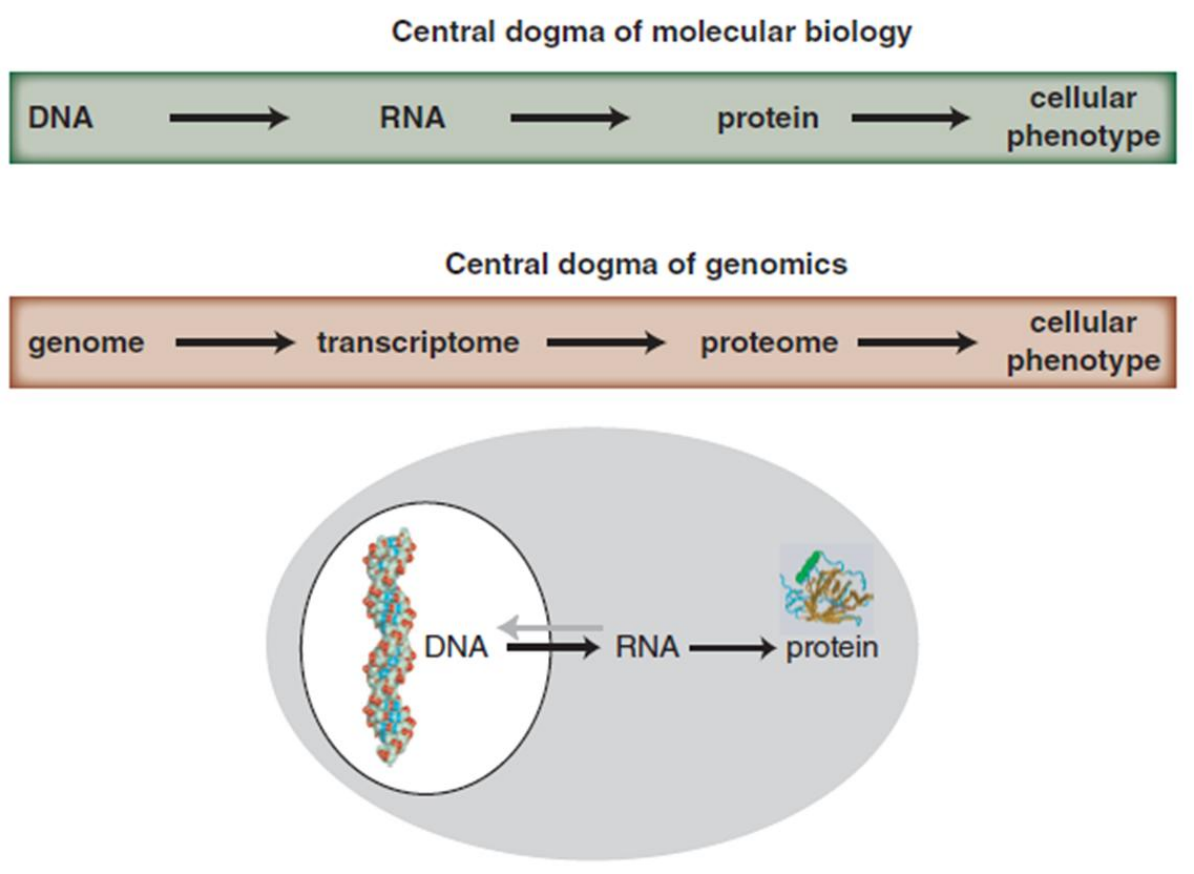
… What is Bioinformatics?...

- According to a National Institutes of Health (NIH) definition, bioinformatics is
  - "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, analyze, or visualize such data."
    - The related discipline of computational biology is "the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems."

… What is Bioinformatics?...

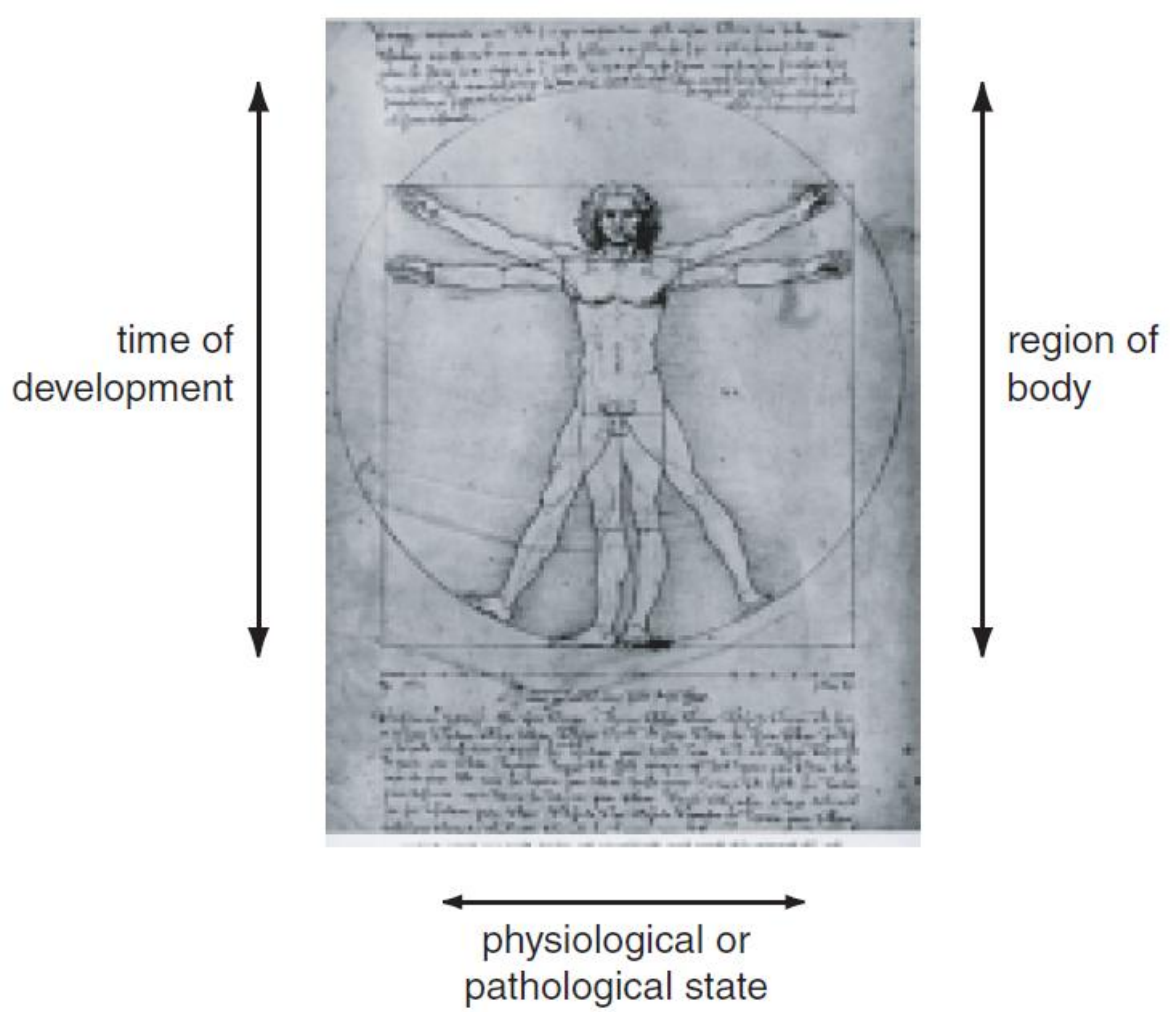
- Another definition from the National Human Genome Research Institute (NHGRI) is that
  - "Bioinformatics is the branch of biology that is concerned with the acquisition, storage, display, and analysis of the information found in nucleic acid and protein sequence data."
- Russ Altman (1998) and Altman and Dugan (2003) offer two definitions of bioinformatics.
  - The first involves information flow following the central dogma of molecular biology (next slide)
  - The second definition involves information flow that is transferred based on scientific methods. This definition includes problems such as
    - designing, validating, and sharing software;
    - storing and sharing data;
    - performing reproducible research workflows;
    - interpreting experiments.

… What is Bioinformatics?...



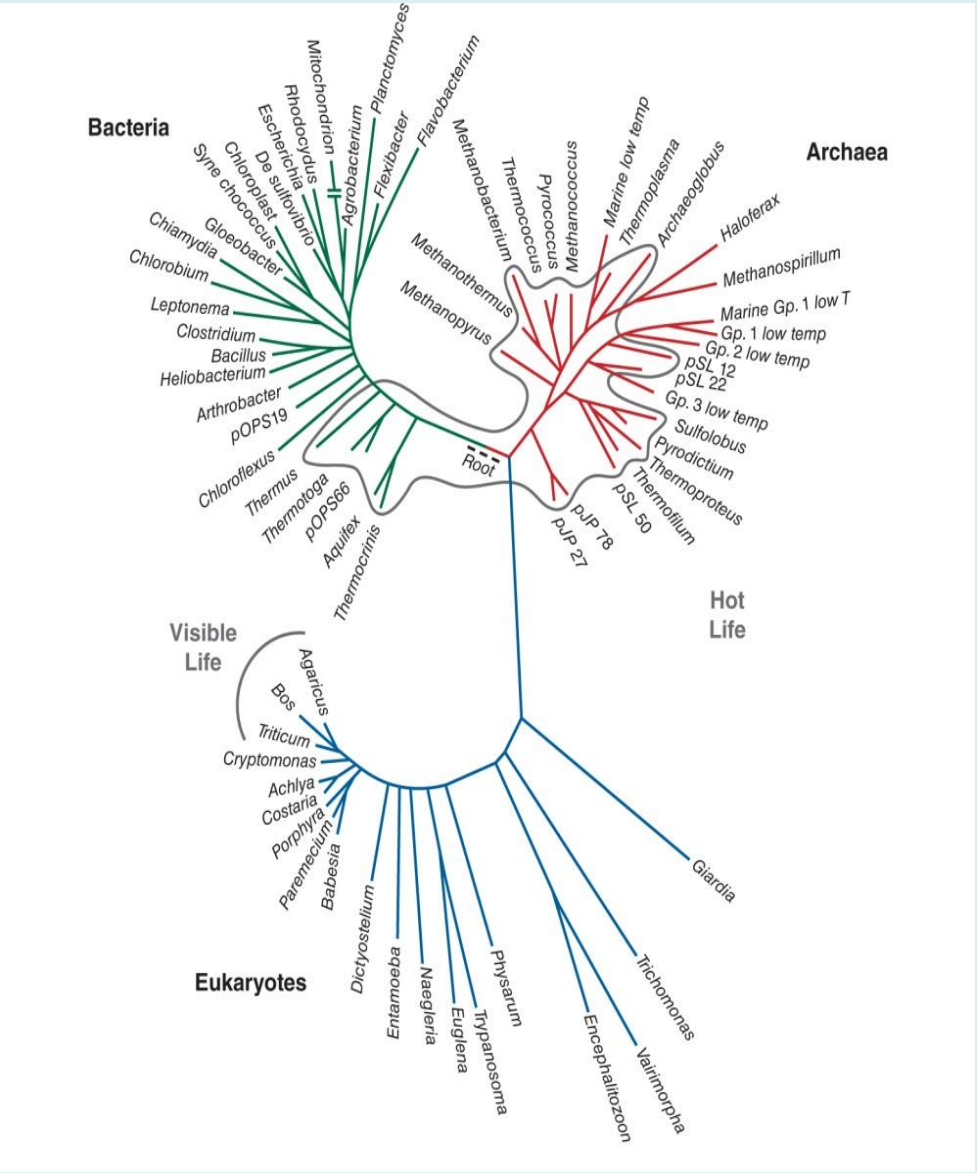
- A 1st perspective of the field of bioinformatics is the cell.
- Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data

… What is Bioinformatics?...



- ✓ A 2nd perspective of bioinformatics is the organism.
- ✓ Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products.
- ✓ For an individual organism, bioinformatics tools can therefore be applied to describe changes through developmental time, changes across body regions, and changes in a variety of physiological or pathological states.

… What is Bioinformatics?...



- ✓ A third perspective of the field of bioinformatics is represented by the tree of life.
- ✓ The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes.
- ✓ Viruses, which exist on the borderline of the definition of life, are not depicted here.
- ✓ For all species, the collection and analysis of molecular sequence data allow us to describe the complete collection of DNA that comprises each organism (the genome).
- ✓ We can further learn the variations that occur between species and among members of a species, and we can deduce the evolutionary history of life on Earth.

…What is Bioinformatics?...

- From a practical sense, bioinformatics is a science that involves
  - collecting,
  - manipulating,
  - analyzing,
  - transmitting

  huge quantities of data,
- uses computers
- bioinformatics refers to computational bioinformatics.

Bioinformatics

- an interdisciplinary field that develops
  - methods and software tools for understanding biological data
- combines
  - computer science,
  - statistics,
  - mathematics,
  - engineering

to analyze and interpret biological data

…What is Bioinformatics?...

- has been used for in silico analyses of biological queries using mathematical and statistical techniques.
    - [In silico (Latin for "in silicon") is an expression used to mean "performed on computer or via computer simulation.]
- primary goal is to increase the understanding of biological processes.
- focuses on developing and applying computationally intensive techniques to achieve this goal.

…What is Bioinformatics?...

- Techniques used include
  - pattern recognition, data mining, machine learning algorithms, and visualization
- Analyzing biological data to produce meaningful information involves writing and running software programs that use algorithms from
  - graph theory, artificial intelligence, soft computing, data mining, signal processing, image processing, and computer simulation.

…What is Bioinformatics?...

- The algorithms in turn depend on theoretical foundations such as
  - discrete mathematics
  - control theory
  - system theory
  - information theory
  - statistics

...What is Bioinformatics?...

- Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information
  - stored in the genetic code,
  - experimental results from various sources,
  - patient statistics,
  - and scientific literature.
- Research in bioinformatics includes method development for
  - storage,
  - retrieval,
  - analysis

  of the data.

...What is Bioinformatics?...

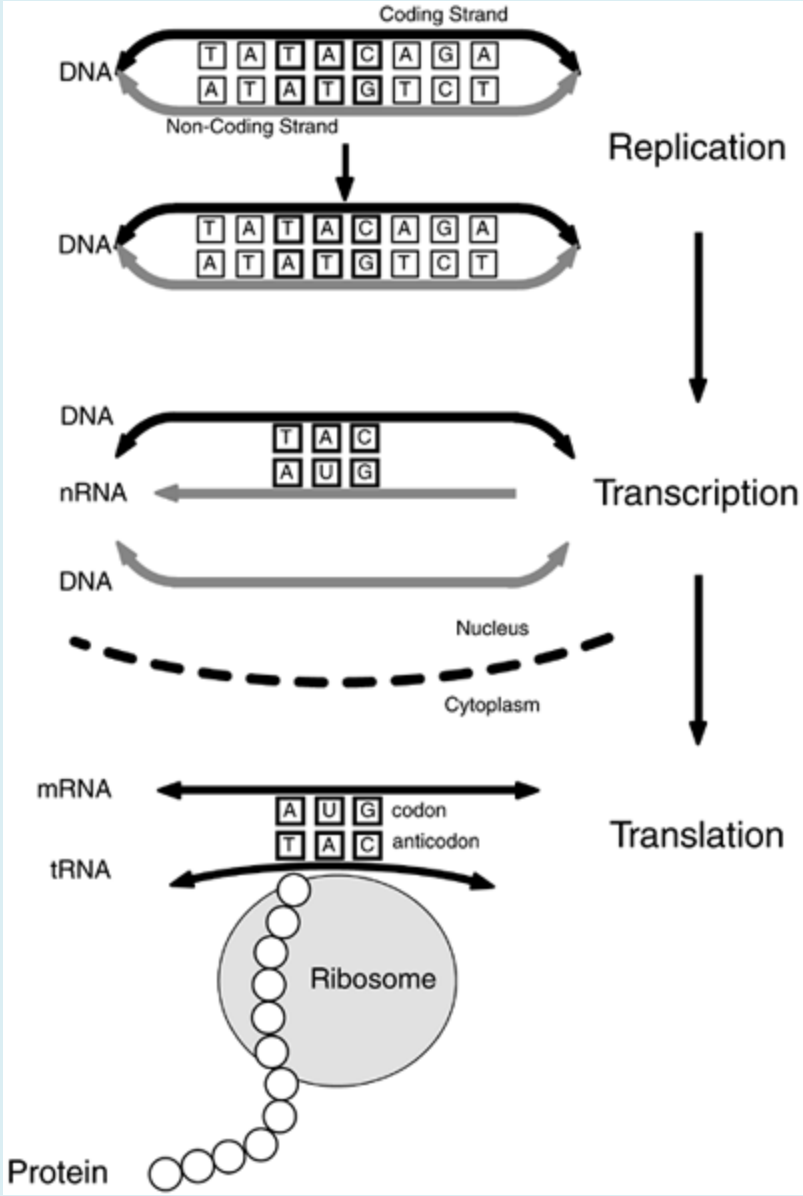
- Bioinformatics
  - a rapidly developing branch of biology
  - highly interdisciplinary,
  - using techniques and concepts from
    - informatics,
    - statistics,
    - mathematics,
    - chemistry,
    - biochemistry,
    - physics,
    - linguistics.

…What is Bioinformatics?...

- The relationship between computer science and biology is a natural one for several reasons.
  - 1st,
    - the phenomenal rate of biological data being produced provides challenges:
      - massive amounts of data have to be stored, analysed, and made accessible.
  - 2nd,
    - the nature of the data is often such that a statistical method, and hence computation, is necessary.
      - This applies in particular to the information on the building plans of proteins and of the temporal and spatial organisation of their expression in the cell encoded by the DNA.
  - 3rd,
    - there is a strong analogy between the DNA sequence and a computer program
      - it can be shown that the DNA represents a Turing Machine.

The Central Dogma of Molecular Biology

- DNA is transcribed to messenger RNA in the cell nucleus, which is in turn translated to protein in the cytoplasm.
- The Central Dogma, shown here from a structural perspective, can also be depicted from an information flow perspective



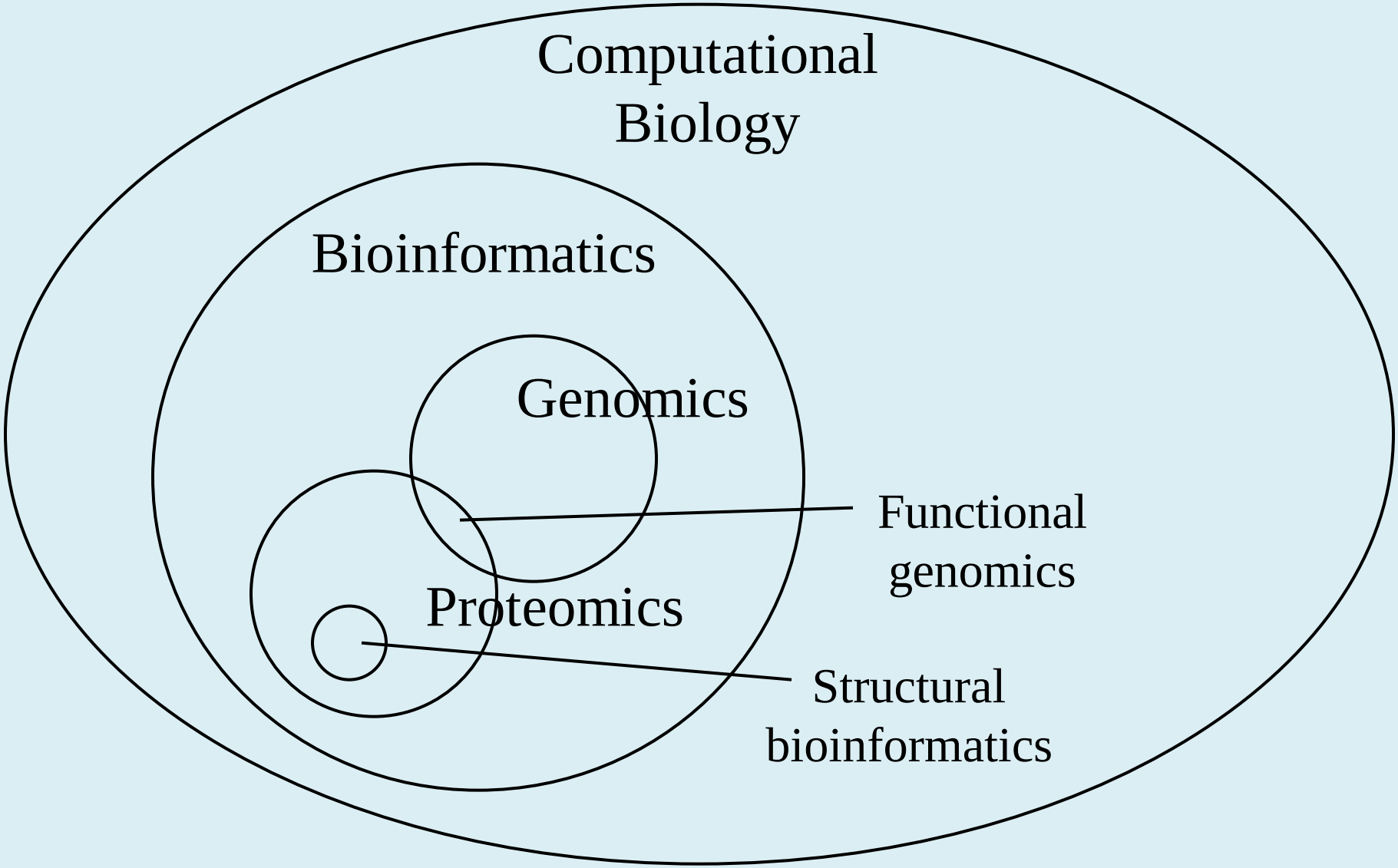
Path to the Bioinformatics

- 1st,
  - Learn Biology.
- 2nd,
  - Decide and pick a problem that interests you for experiment.
- 3rd,
  - Find and learn about the Bioinformatics tools.
- 4th,
  - Learn the Computer Programming Languages.
    - Perl, Pyton, R, Java, etc.
- 5th,
  - Experiment on your computer and learn different programming techniques.

Why is Bioinformatics Important?

- Applications areas include
  - Medicine
  - Pharmaceutical drug design
  - Toxicology
  - Molecular evolution
  - Biosensors
  - Biomaterials
  - Biological computing models
  - DNA computing

Scope of Computational Biology



Genomics

- The study of the genome,
  - which is the complete set of the genetic material or DNA present in an organism.
- studies all genes and their inter relationships in an organism, so as to identify their combined influence on its growth and development.
- The field of genomics attracted worldwide attention in the late 1990s with the race to map the human genome.
  - The Human Genome Project (HGP), completed in April 2003, made available for the first time the complete genetic blueprint of a human being.

Proteomics

- large-scale study of proteomes,
  - which is a set of proteins produced in an organism, system, or biological context.
    - We may refer to, for instance, the proteome of a species (eg, Homo sapiens) or an organ (eg, the liver).
  - The proteome is not constant;
    - it differs from cell to cell and changes over time.
  - To some degree, the proteome reflects the underlying transcriptome.
    - However, protein activity (often assessed by the reaction rate of the processes in which the protein is involved) is also modulated by many factors in addition to the expression level of the relevant gene.

Proteomics

- is used to investigate:
  - when and where proteins are expressed;
  - rates of protein production, degradation, and steady-state abundance;
  - how proteins are modified (for example, post-translational modifications (PTMs) such as phosphorylation);
  - the movement of proteins between subcellular compartments;
  - the involvement of proteins in metabolic pathways;
  - how proteins interact with one another.
- can provide significant biological information for many biological problems, such as:
  - Which proteins interact with a particular protein of interest (for example, the tumor suppressor protein p53)?
  - Which proteins are localized to a subcellular compartment (for example, the mitochondrion)?
  - Which proteins are involved in a biological process (for example, circadian rhythm)?

Structural bioinformatics/genomics

- is the branch of bioinformatics
  - which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA.
- deals with generalizations about macromolecular 3D structure such as comparisons of overall folds and local motifs, principles of molecular folding, evolution, and binding interactions, and structure/function relationships, working both from experimentally solved structures and from computational models.

Functional genomics

- is a field of molecular biology,
  - which attempts to make use of the vast wealth of data given by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing) to describe gene (and protein) functions and interactions.
    - Unlike structural genomics, it focuses on the dynamic aspects such as gene transcription, translation, regulation of gene expression and protein–protein interactions, as opposed to the static aspects of the genomic information such as DNA sequence or structures.
- attempts to answer questions about the function of DNA at the levels of genes, RNA transcripts, and protein products.

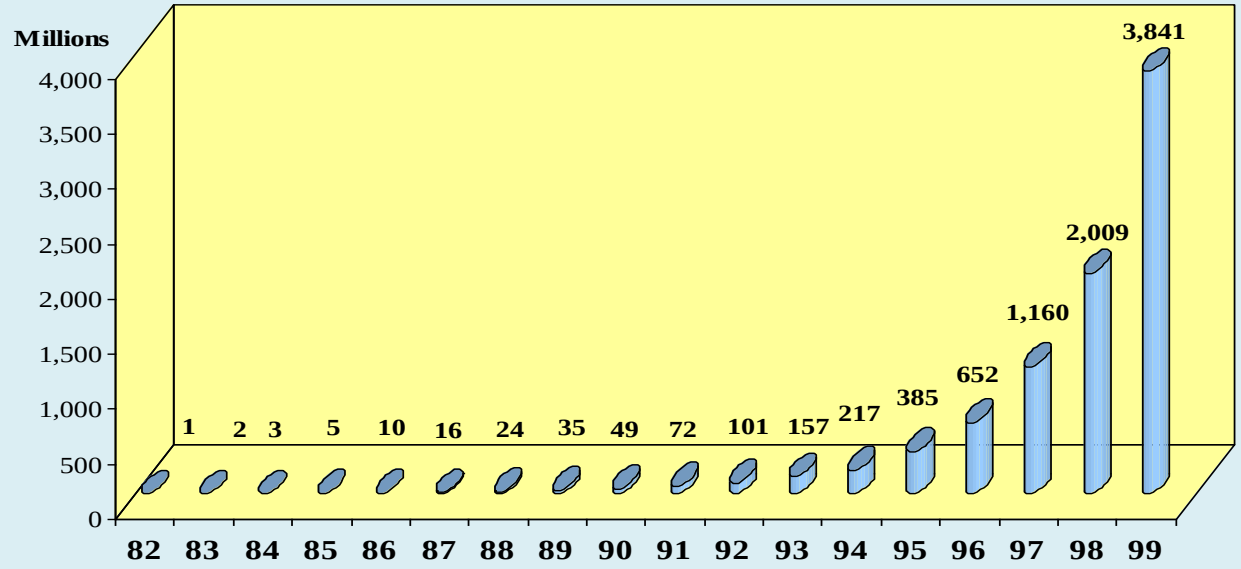
Why is bioinformatics hot?

- Supply/demand: few people adequately trained in both biology and computer science
- Genome sequencing, microarrays, etc lead to large amounts of data to be analyzed
- Leads to important discoveries
- Saves time and money

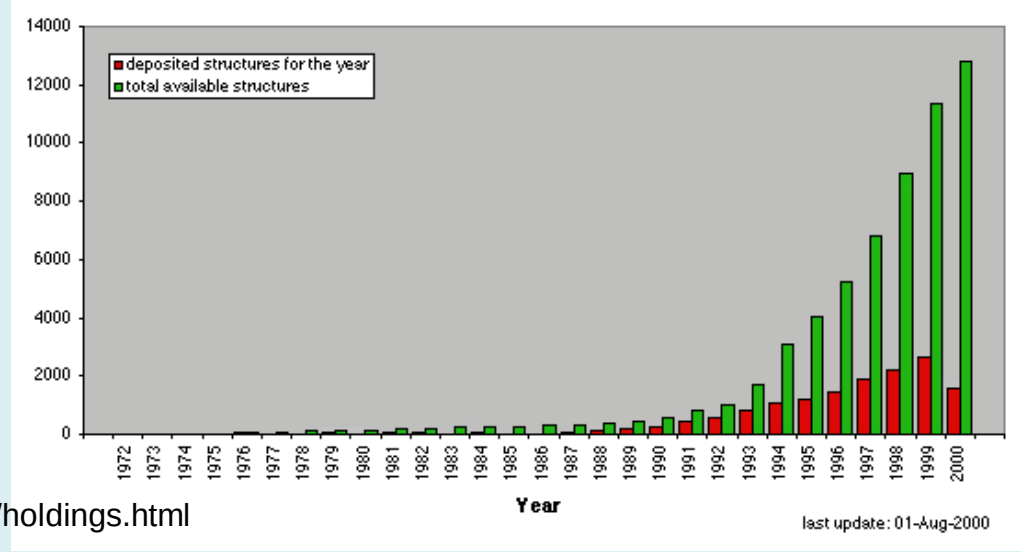
The Role of *Computational* Biology

GenBank BASEPAIR GROWTH

Source: GenBank



3D Structures Growth:

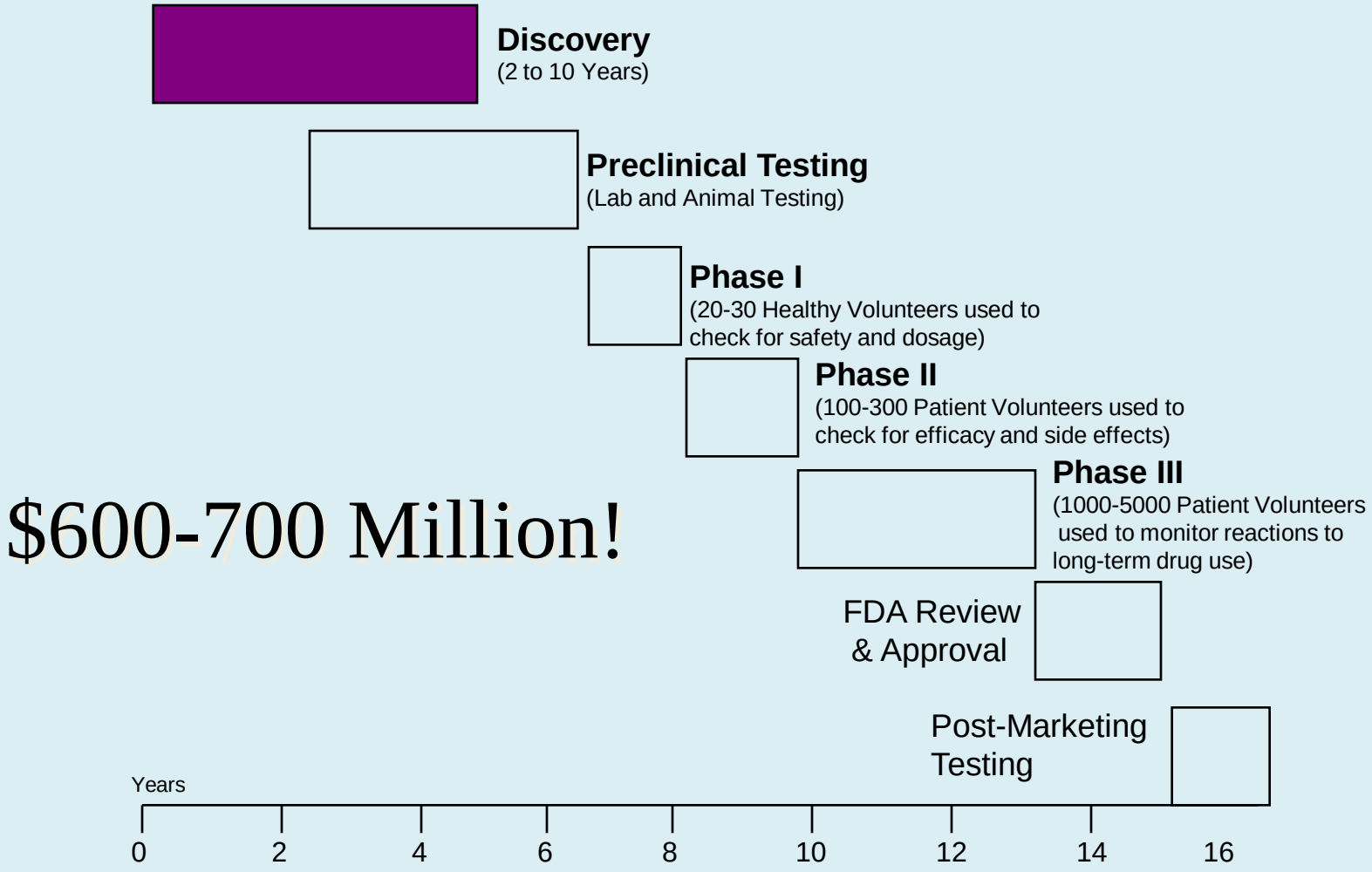


Source:
http://www.rcsb.org/pdb/holdings.html

Fighting Human Disease

- Genetic / Inherited
  - Diabetes
- Viral
  - Flu, common cold
- Bacterial
  - Meningitis, Strep throat

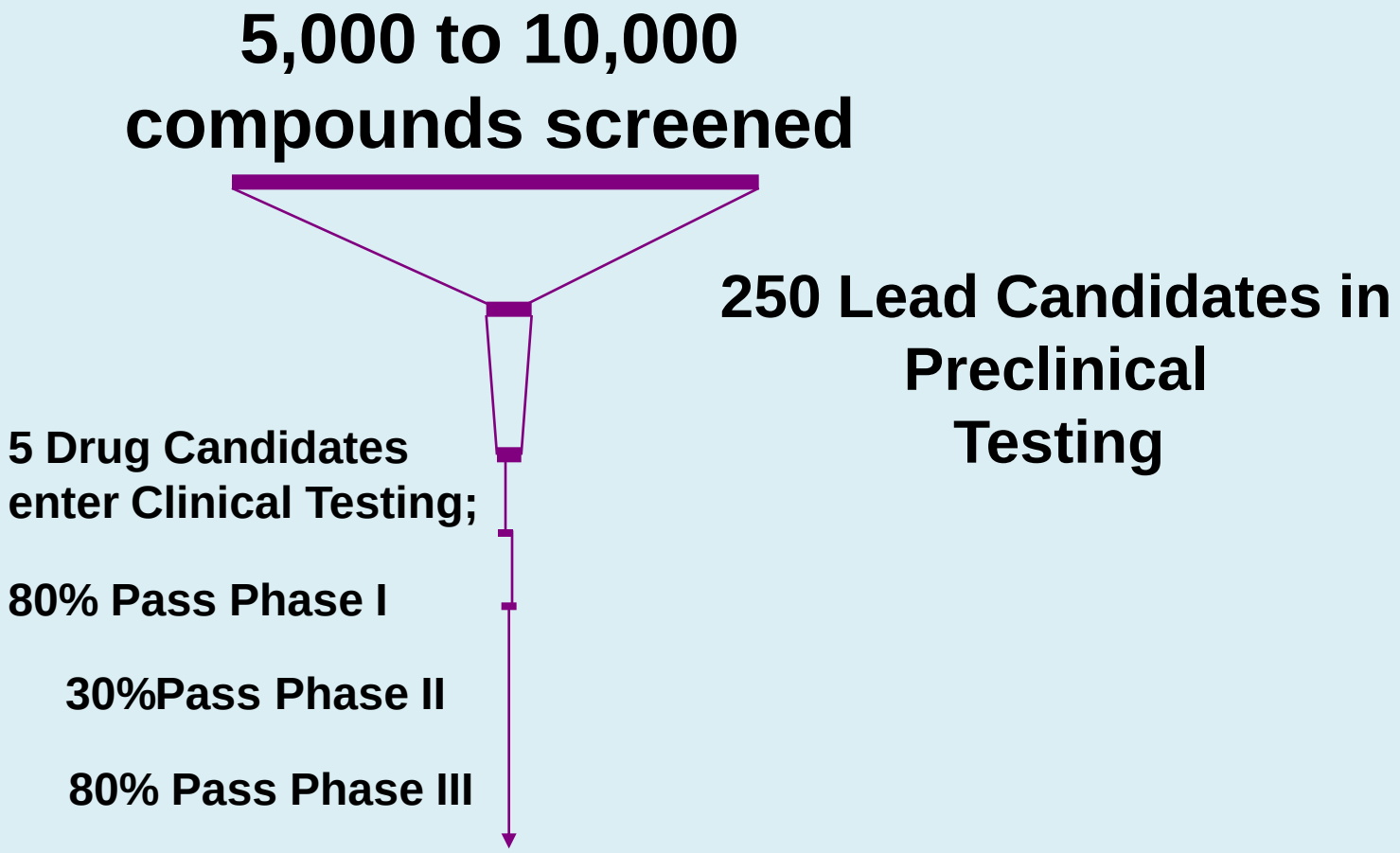
Drug Development Life Cycle



$600-700 Million!

7 – 15 Years!

Drug lead screening



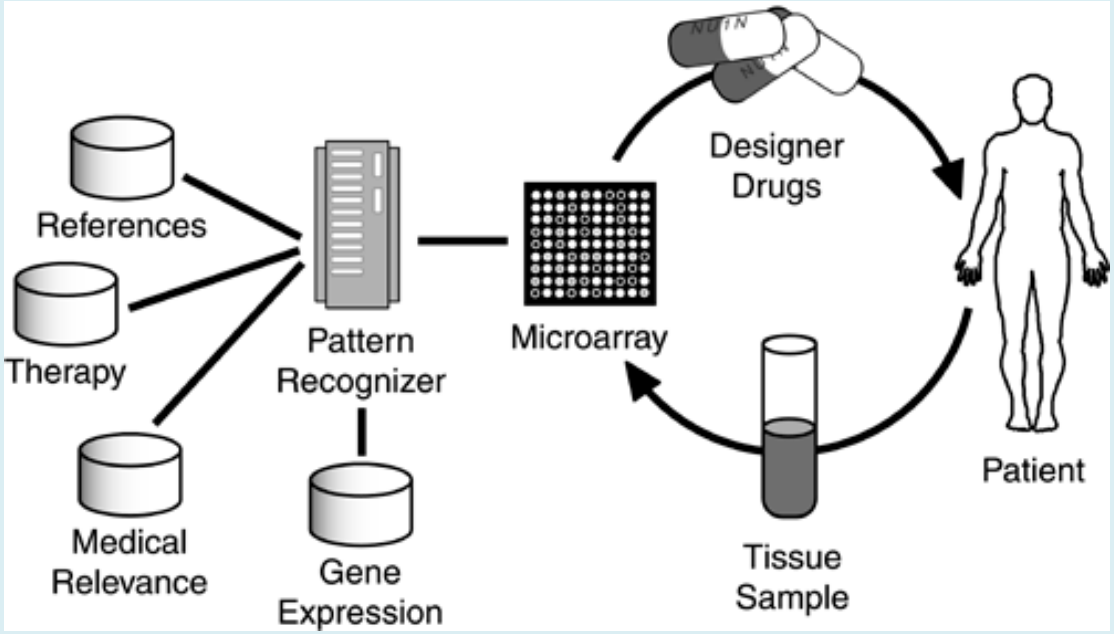
One drug approved by the FDA

Killer application

- In the biotechnology industry, every researcher and entrepreneur hopes to develop or discover the next "killer app"
  - the one application that will bring the world to his or her door and provide funding for R&D, marketing, and production.
    - For example, in general computing, the electronic spreadsheet and the desktop laser printer have been the notable killer apps.
    - The spreadsheet not only transformed the work of accountants, research scientists, and statisticians, but the underlying tools formed the basis for visualization and mathematical modeling.
    - The affordable desktop laser printer created an industry and elevated the standards of scientific communications, replacing rough graphs created on dot-matrix printers with high-resolution images.

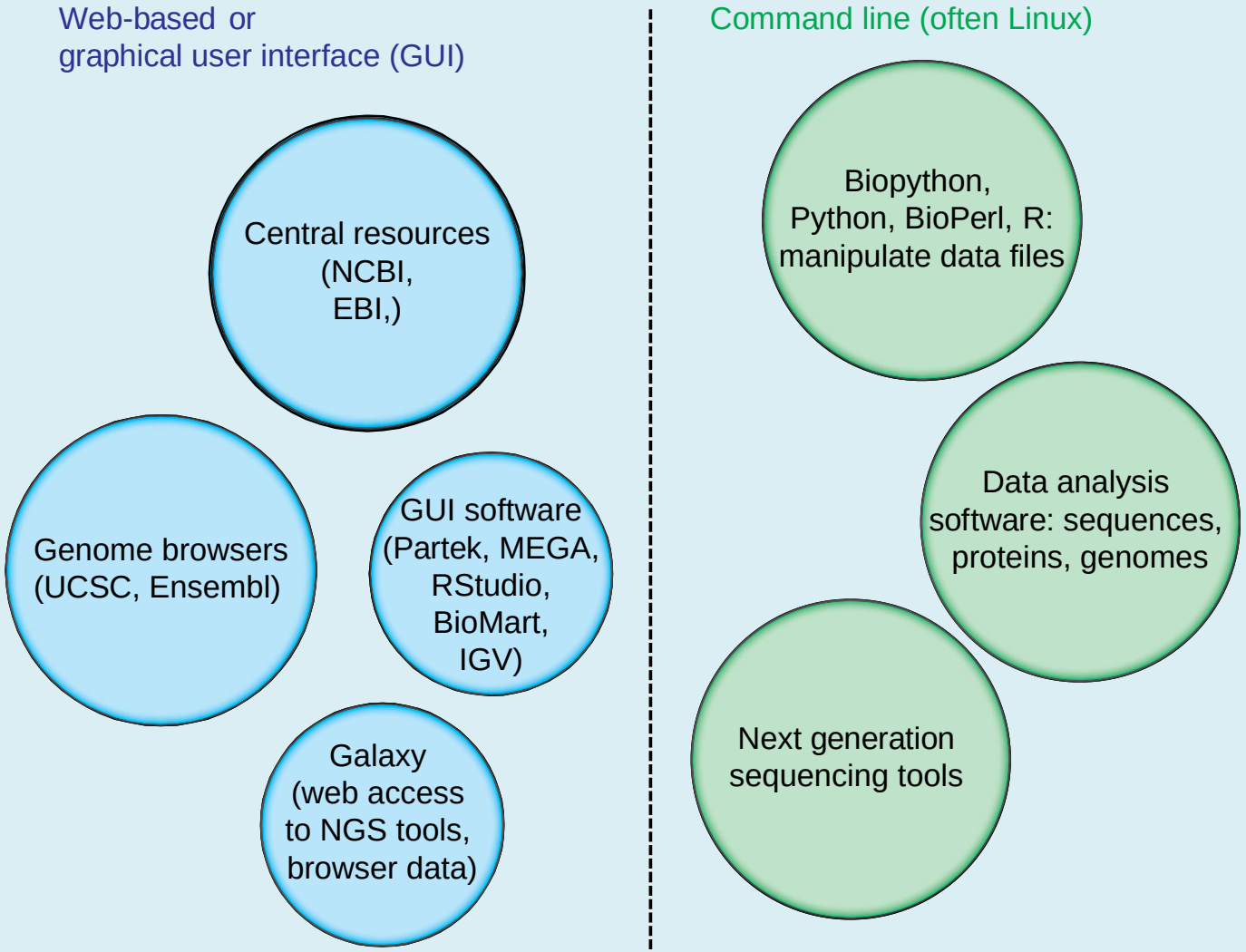
Killer application

- "What might be the computer-enabled 'killer app' in bioinformatics?“
- Although there are numerous military and agricultural opportunities, one of the most commonly cited examples of the killer app is in personalized medicine, as illustrated in Figure



personalized medicine: the custom, just-in-time delivery of medications (popularly called "designer drugs") tailored to the patient‘scondition.

Bioinformatics Software: Two Cultures



Bioinformatics Software: Two Cultures

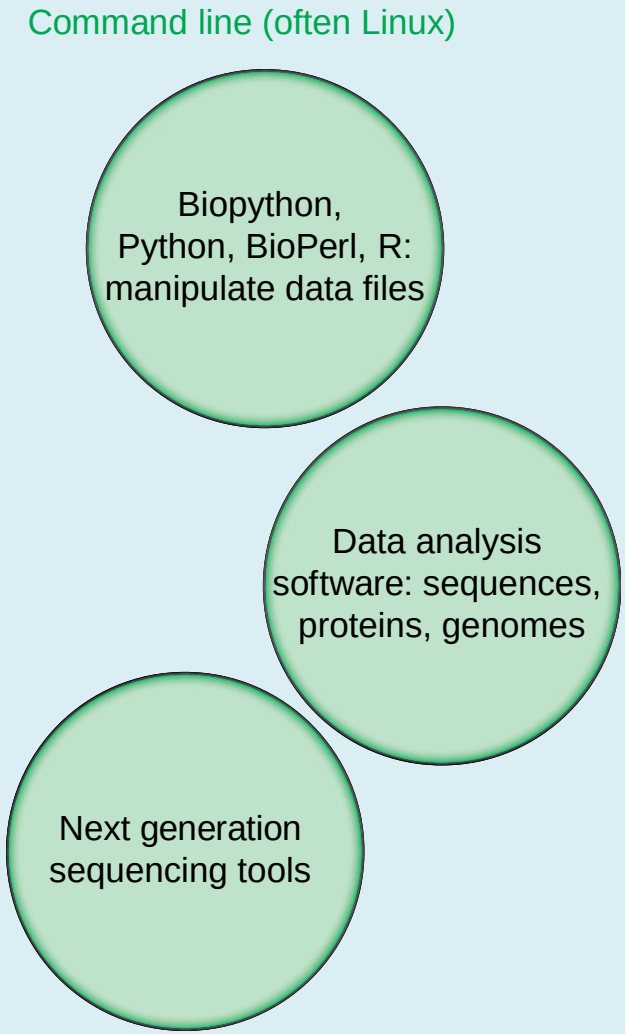
- Many bioinformatics tools and resources are available on the internet, such as major genome browsers and major portals (NCBI, Ensembl, UCSC).
- These are:
  - accessible (requiring no programming expertise)
  - easy to browse to explore their depth and breadth
  - very popular
  - familiar (available on any web browser on any platform)

Bioinformatics Software: Two Cultures

- Many bioinformatics tools and resources are available on the command-line interface (sometimes abbreviated CLI).
  - These are often on the Linux platform (or other Unix-like platforms such as the Mac command line).
  - They are essential for many bioinformatics and genomics applications.
  - Most bioinformatics software is written for the Linux platform.
    - Many bioinformatics datasets are so large (e.g. high throughput technologies generate millions to billions or even trillions of data points) requiring command-line tools to manipulate the data.

CLI

- Should you learn to use the Linux operating system?
  - Yes, if you want to use mainstream bioinformatics tools.
- Should you learn Python or Perl or R or another programming language?
  - It's a good idea if you want to go deeper into bioinformatics, but also, it depends what your goals are.
  - Many software tools can be run in Linux on the command-line without needing to program.
- Think of this figure like a map.
  - Where are you now?
  - Where do you want to go?



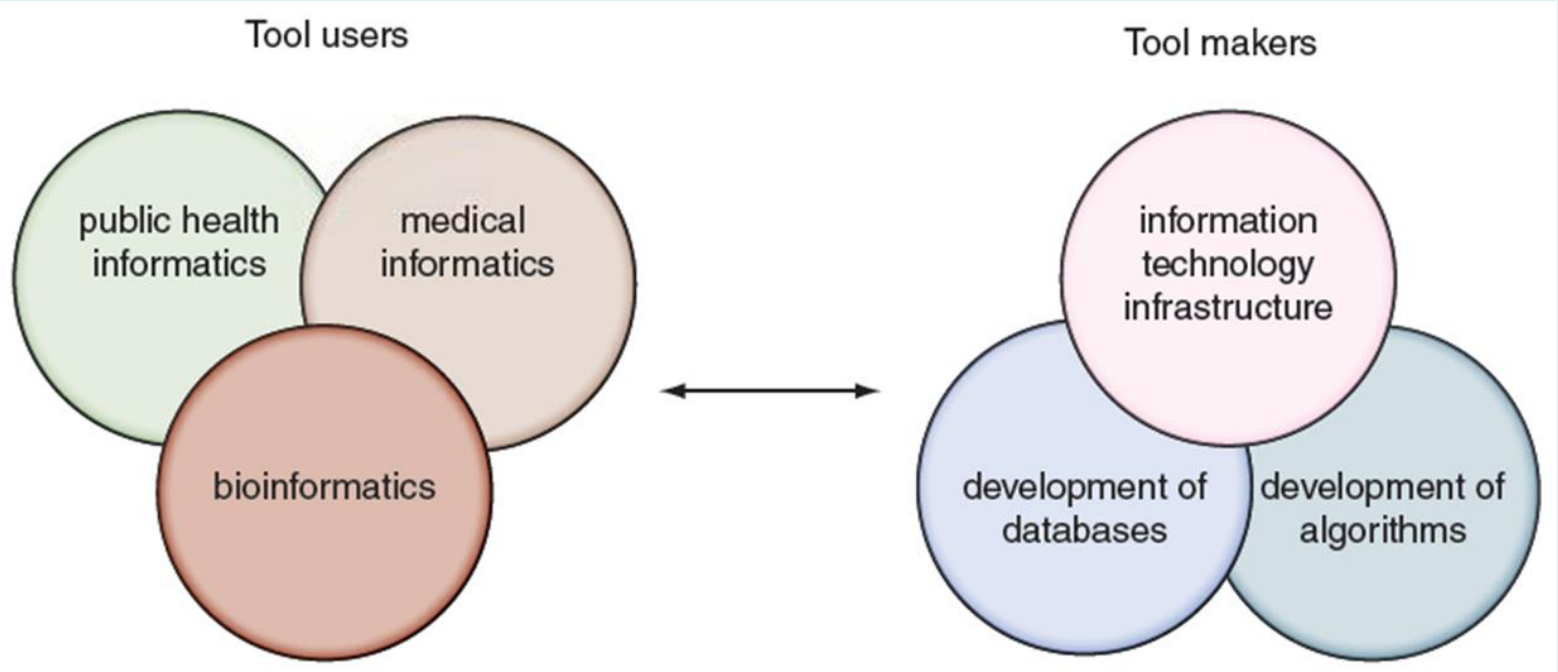
Some web-based (GUI) and command-line (CLI) software

Topic	Web-based or GUI software	Command-line software
Access to information	BioMart Genome Workbench	EDirect
Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
BLAST	BLAST	BLAST+
Database searching	DELTA-BLAST Megablast	HMMER
Multiple alignment	Pfam, MUSCLE	MAFFT
Phylogeny	MEGA	MrBayes
Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
RNA	RNAfam, tRNAscan	

Some web-based (GUI) and command-line (CLI) software

RNAseq	Galaxy	affy (R package), RSEM
Proteomics	ExPASy	pepstats (EMBOSS)
Protein structure	Cn3D, Pymol	psiphi (EMBOSS)
Functional genomics	FLink, Cytoscape	
Tree of life		Velvet (assembly)
Viruses		MUMmer (alignment)
Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
Fungi	YGOB	Ensembl (variants)
Eukaryotic genomes		
Human genome		PLINK
Human disease	OMIM, BioMart	EDirect, MitoSeek

Tool makers and tool users across informatics disciplines



- Many informatics disciplines have emerged in recent years.
- Bioinformatics is distinguished by its particular focus on DNA and proteins (impacting its databases, its tools, and its entire culture).

Reproducible Research in Bioinformatics

- Science by its nature is cumulative and progressive.
- Whether you use web-based or command-line tools, research should be conducted in a way that is reproducible by the investigator and by others.
- This facilitates the cumulative, progressive nature of your work.
- In the realm of bioinformatics this means the following.

Reproducible Research in Bioinformatics

- A workflow should be well documented.
  - This may include keeping text documents on your computer in which you can copy and paste complex commands, URLs, or other forms of data.
- To facilitate your work, information stored on a computer should be well organized.
- Data should be made available to others.
  - Repositories are available to store high-throughput data in particular.
    - Examples are Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) at NCBI and ArrayExpress and European Nucleotide Archive (ENA) at EBI.
- Metadata can be equally as crucial as data.
  - Metadata refers to information about datasets.
    - For a bacterial genome that has been sequenced, the metadata may include the location from which the bacterium was isolated, the culture conditions, and whether it is pathogenic.
- Databases that are used should be documented.
  - Since the contents of databases change over time, it is important to document the version number and the date(s) of access.
- Software should be documented.
  - For established packages, the version number should be provided.
    - Further documenting the specific steps you use allows others to independently repeat your analyses.
      - In an effort to share software, many researchers use repositories such as GitHub.

Bioinformatics Tools

- ✓ The processes of designing a new drug using bioinformatics tools have open a new area of research. However, computational techniques assist one in searching drug target and in designing drug *in silco*, but it takes long time and money. In order to design a new drug one need to follow the following path.
  - ✓ Identify target disease
  - ✓ Study Interesting Compounds
  - ✓ Detection the Molecular Bases for Disease
  - ✓ Rational Drug Design Techniques
  - ✓ Refinement of Compounds
  - ✓ Quantitative Structure Activity Relationships (QSAR)
  - ✓ Solubility of Molecule
  - ✓ Drug Testing

1. Identify Target Disease:-

- ✓ One needs to know all about the disease and existing or traditional remedies. It is also important to look at very similar afflictions and their known treatments.
- ✓ Target identification alone is not sufficient in order to achieve a successful treatment of a disease.
- ✓ A real drug needs to be developed. This drug must influence the target protein in such a way that it does not interfere with normal metabolism.
- ✓ Bioinformatics methods have been developed to virtually screen the target for compounds that bind and inhibit the protein.

2. Study Interesting Compounds:-

- ✓ One needs to identify and study the lead compounds that have some activity against a disease.
- ✓ These may be only marginally useful and may have severe side effects.
- ✓ These compounds provide a starting point for refinement of the chemical structures.

3. Detect the Molecular Bases for Disease:-

- ✓ If it is known that a drug must bind to a particular spot on a particular protein or nucleotide then a drug can be tailor made to bind at that site.
- ✓ This is often modeled computationally using any of several different techniques.
- ✓ Traditionally, the primary way of determining what compounds would be tested computationally was provided by the researchers' understanding of molecular interactions.
- ✓ A second method is the brute force testing of large numbers of compounds from a database of available structures.

4. Rational drug design techniques:-

- ✓ These techniques attempt to reproduce the researchers' understanding of how to choose likely compounds built into a software package that is capable of modeling a very large number of compounds in an automated way.
- ✓ Many different algorithms have been used for this type of testing, many of which were adapted from artificial intelligence applications.
- ✓ The complexity of biological systems makes it very difficult to determine the structures of large biomolecules.
- ✓ Ideally experimentally determined (x-ray or NMR) structure is desired, but biomolecules are very difficult to crystallize .

5. Rational drug design techniques:-

- ✓ Once you got a number of lead compounds have been found, computational and laboratory techniques have been very successful in refining the molecular structures to give a greater drug activity and fewer side effects.
- ✓ Done both in the laboratory and computationally by examining the molecular structures to determine which aspects are responsible for both the drug activity and the side effects.

6. Quantitative Structure Activity Relationships (QSAR):-

- ✓ Computational technique should be used to detect the functional group in your compound in order to refine your drug.
- ✓ QSAR consists of computing every possible number that can describe a molecule then doing an enormous curve fit to find out which aspects of the molecule correlate well with the drug activity or side effect severity.
- ✓ This information can then be used to suggest new chemical modifications for synthesis and testing.

7. Solubility of Molecule:-

- ✓ One need to check whether the target molecule is water soluble or readily soluble in fatty tissue will affect what part of the body it becomes concentrated in.
- ✓ The ability to get a drug to the correct part of the body is an important factor in its potency.
- ✓ Ideally there is a continual exchange of information between the researchers doing QSAR studies, synthesis and testing.
- ✓ These techniques are frequently used and often very successful since they do not rely on knowing the biological basis of the disease which can be very difficult to determine.

Applications of Bioinformatics

` Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. The main uses of bioinformatics include:

- ✓ Bioinformatics plays a vital role in the areas of structural genomics, functional genomics, and nutritional genomics.
- ✓ It covers emerging scientific research and the exploration of proteomes from the overall level of intracellular protein composition (protein profiles), protein structure, protein-protein interaction, and unique activity patterns (e.g. post-translational modifications).

- ✓ Bioinformatics is used to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically.
- ✓ Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modeling, virtual screening, etc.
- ✓ Bioinformatics is playing an increasingly important role in almost all aspects of drug discovery and drug development.
- ✓ Bioinformatics tools are very effective in prediction, analysis and interpretation of clinical and preclinical findings.

Applications in Other Fields

` Its major applications include in the following fields:

Molecular medicine

- ✓ The human genome will have profound effects on the fields of biomedical research and clinical medicine.
- ✓ The completion of the human genome and the use of bioinformatic tools means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- ✓ This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised medicine

- ✓ Clinical medicine will become more personalised with the development of the field of pharmacogenomics.
- ✓ This is the study of how an individual's genetic inheritence affects the body's response to drugs.
- ✓ Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.
- ✓ In the future, doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative medicine

- ✓ With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality.

Gene therapy

- ✓ In the not too distant future with the use of bioinformatics tool, the potential for using genes themselves to treat disease may become a reality.
- ✓ Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a person's genes.

Drug development

- ✓ At present all drugs on the market target only about 500 proteins.
- ✓ With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed.
- ✓ These highly specific drugs promise to have fewer side effects than many of today's medicines.

Microbial genome applications

- ✓ The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications.
- ✓ For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction.
- ✓ By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

Waste cleanup

- ✓ Deinococcus radiodurans is known as the world's toughest bacteria and it is the most radiation resistant organism known.
- ✓ Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

Climate change Studies

- ✓ Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change.
- ✓ Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels.
- ✓ One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

Biotechnology

- ✓ The archaeon Archaeoglobus fulgidus and the bacterium Thermotoga maritima have potential for practical applications in industry and government-funded environmental remediation.
- ✓ These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defense, and private companies with heat-stable enzymes suitable for use in industrial processes
- ✓ Other industrially useful microbes include, Corynebacterium glutamicum which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine.
- ✓ The substance is employed as a source of protein in animal nutrition.

- ✓ Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bonemeal.
- ✓ Lactococcus lactis is one of the most important micro-organisms involved in the dairy industry.
- ✓ Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of lactis to serve as a vehicle for delivering drugs.

Forensic analysis of microbes

- ✓ Scientists used their genomic tools to help distinguish between the strain of Bacillus anthracis that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

The reality of bioweapon creation

- ✓ Scientists have recently built the virus poliomyelitis using entirely artificial means.
- ✓ They did this using genomic data available on the Internet and materials from a mail-order chemical supply.
- ✓ The research was financed by the US Department of Defense as part of a biowarfare response program to prove to the world the reality of bioweapons.
- ✓ The researchers also hope their work will discourage officials from ever relaxing programs of immunization.
- ✓ This project has been met with very mixed feelings.