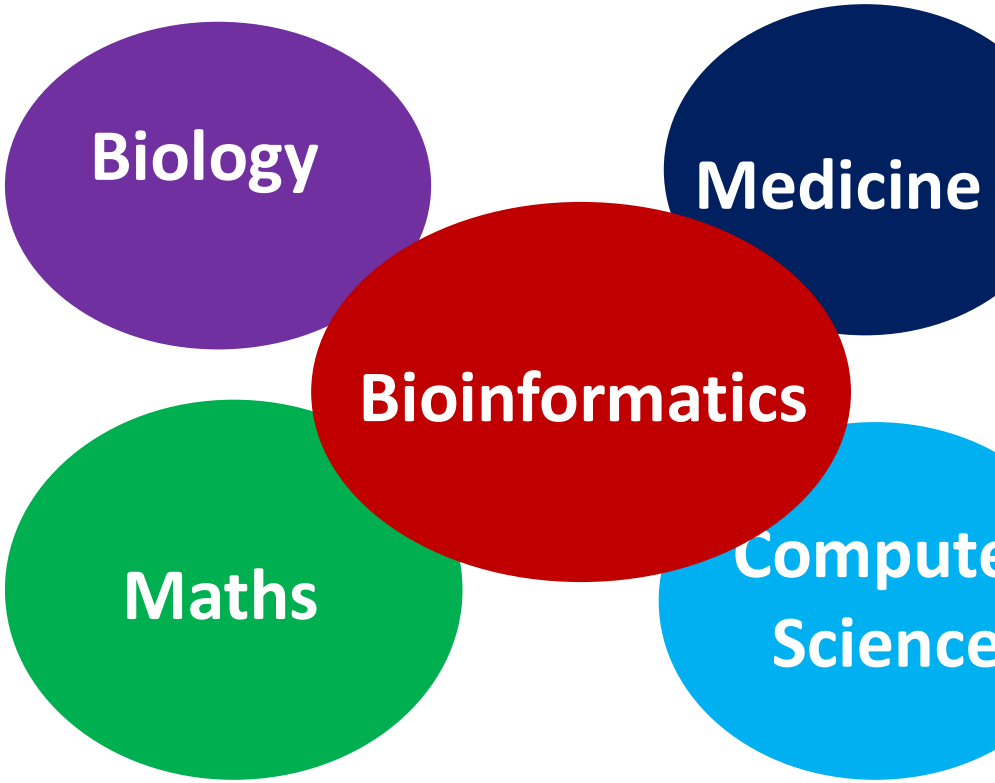
Introduction to Bioinformatics

Bioinformatics is an interdisciplinary scientific field that develops methods and software tools for storing, retrieving, organizing and analysing biological data.

As an interdisciplinary field, bioinformatics combines computer Science, statistics, mathematics and engineering to study biological data and processes.



Bioinformatics deals with

- Design and implementation of new algorithms and statistics which
- assess relationship among members of large data sets.
- Analysis and interpretation of various data types, which includes
- nucleotide and amino acid sequences and structure of protein.
- To develop computational tools and databases that enables efficient analysis access and management of biologically significant information.

Aims of Bioinformatics

- Improve content and utility of databases.
- Develop better tools for data generation, capture, and annotation.
- Develop and improve tools and databases for comprehensive functional studies.
- Develop and improve tools for representing and analysing sequence similarity and variation.
- Create mechanisms to support effective approaches for producing robust, exportable software that can be widely shared.

Goals of Bioinformatics

- Development and implementation of computer programs that enable
- efficient access to , use and management of various type of information
- Development of new algorithms and statistical measures with which to assess relationships among members of large data.
- Understanding the biological process
- Seq. alignment
- Gene finding
- Assembly
- Drug designing
- Protein structure alignment
- Gene expression
- Genome annotation

History of Bioinformatics

1859 – The "On the Origin of Species", published by Charles Darwin that introduced theory of genetic evolution – allows adaptation over time to produce organisms best suited to the environment.

1869 - The DNA from nuclei of white blood cells was first isolated by Friedrich Meischer.

1951 – Linus Pauling and Corey propose the structure for the alphahelix and beta-sheet.

1953 - Watson and Crick propose the double helix model for DNA based on x-ray data obtained by Franklin and Wilkins.

1955 - The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger.

1958 - The Advanced Research Projects Agency (ARPA) is formed in the US.

History of Bioinformatics

- 1973 - The Brookhaven Protein Data Bank(PDB) is announced.
- 1987 - Perl (Practical Extraction Report Language) is released by Larry Wall.
- 1988 - National Centre for Biotechnology Information (NCBI) founded at NIH/NLM.
- 1990 - Human Genome Project launched, BLAST program introduced by S. Karlin and S.F. Altshul, Tim Berners-Lee, a British scientist invented the World Wide Web in 1990.
- 1992 - The Institute for Genome Research (TIGR), associated with plans to exploit sequencing commercially through gene identification and drug discovery, was formed.
- 2001 - The human genome (3,000 Mbp) is published.
- 2010 :Completion of the 2010 Project: to understand the function of all genes within their cellular, organism and evolutionary context of Arabidopsis thaliana.
- 2050: To complete of the first computational model of a complete cell, or maybe even alread a complete organism.

The IBM 7090 computer

Margaret Oakley Dayhoff, who created the first large scale database of protein sequences, and developed many of the algorithms for analysis of relatedness of proteins, early 1970s

What are the biological problems

- What is the role of a particular gene?
- Does a particular gene help cause a disease?
- How does a drug affect a cell?
- Can we insert a gene into corn to protect it against diseases or pests?
- Can we design a drug to accomplish a particular purpose?
- Can we build a cell that eats pollution?

Why do we need bioinformatics?

- Rapid increase in data due to genomics.
- Too much data to characterize genes/proteins individually.
- Bioinformatics is “smart use” of information.
- Ideally computational and experimental biology are partners.

>100000 species are represented in GenBank (NCBI)

All species	128,941
Viruses	6,137
Bacteria	31,262
Archaea	2,100
Eukaryota	87,147

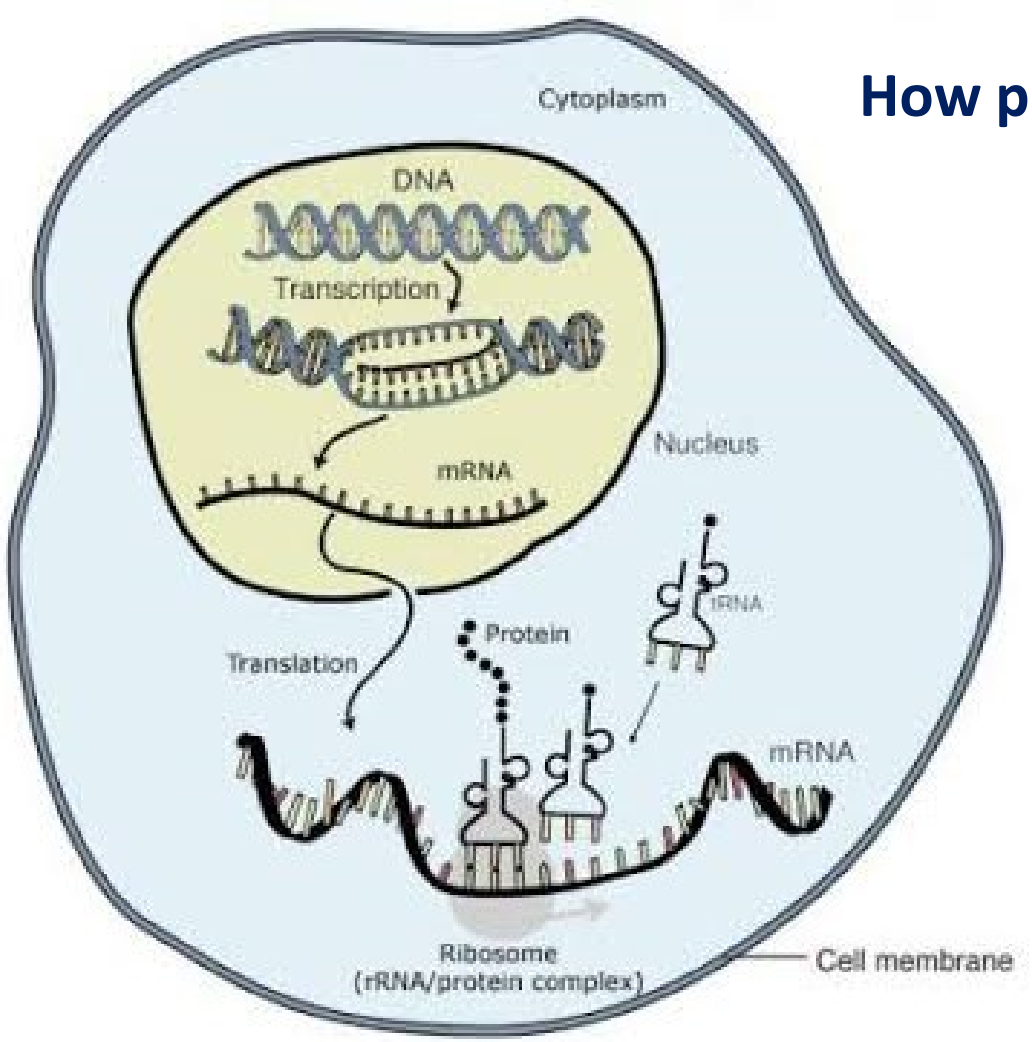
How much information is there?

- Nucleotide records
  9,102,634
- Nucleotides
  10,335,692,655
- Protein sequences
  1,183,833
- 3D structures
  12,863
- Expression data points
  >20,000,000
- Human Unigene clusters
  84,130
- Maps and complete genomes
  11,166
- Different taxonomy nodes
  162,025
- dbSNP
  1,463,178
- Human Refgene records
  14,133
- Human contigs >500 kb (28,525 MB)
  257
- PubMed records
  10,965,353
- OMIM records
  11,950

What do you need to do bioinformatics?

- Knowledge in Molecular Biology
- Statistics
- Mathematics (algorithm development)
- Communicate biological problems to computer scientists
- Working knowledge in bioinformatics tools
- Computer proficiency (windows/command line)
- Programming Skills
- Data administration

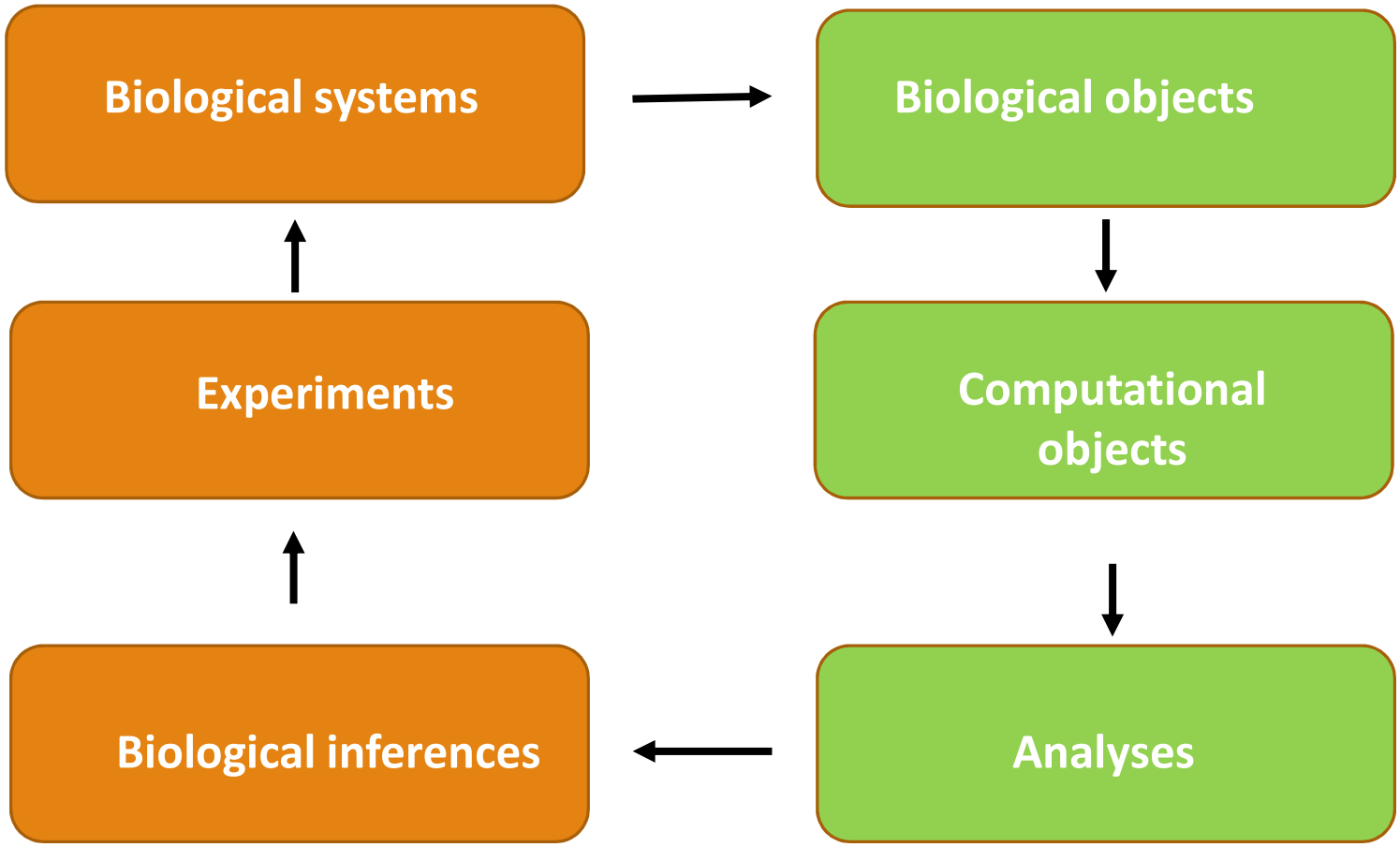
Knowledge in Molecular Biology



How protein is synthesized??

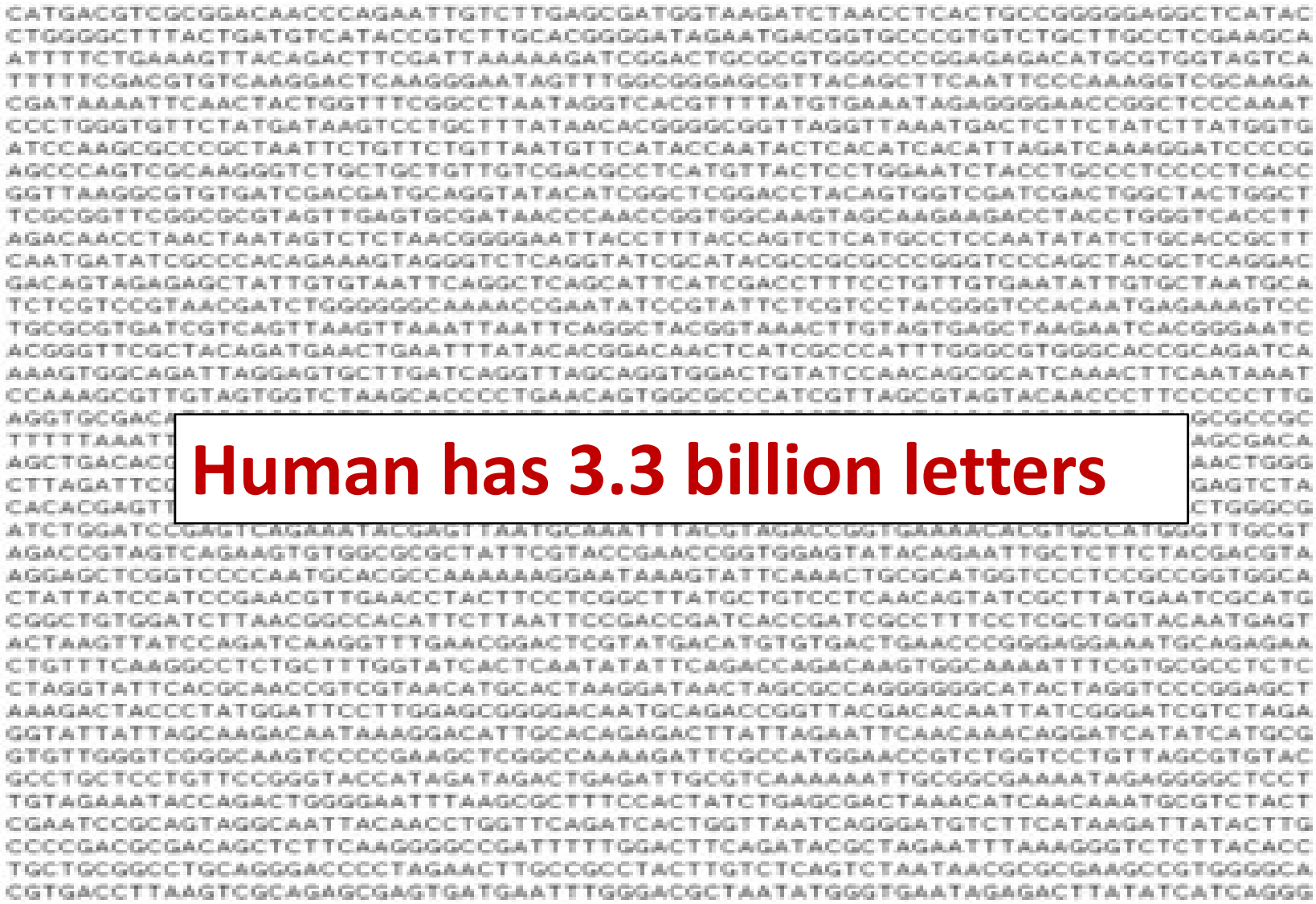
Computational-experimental cycle

Helping biologist solve problems

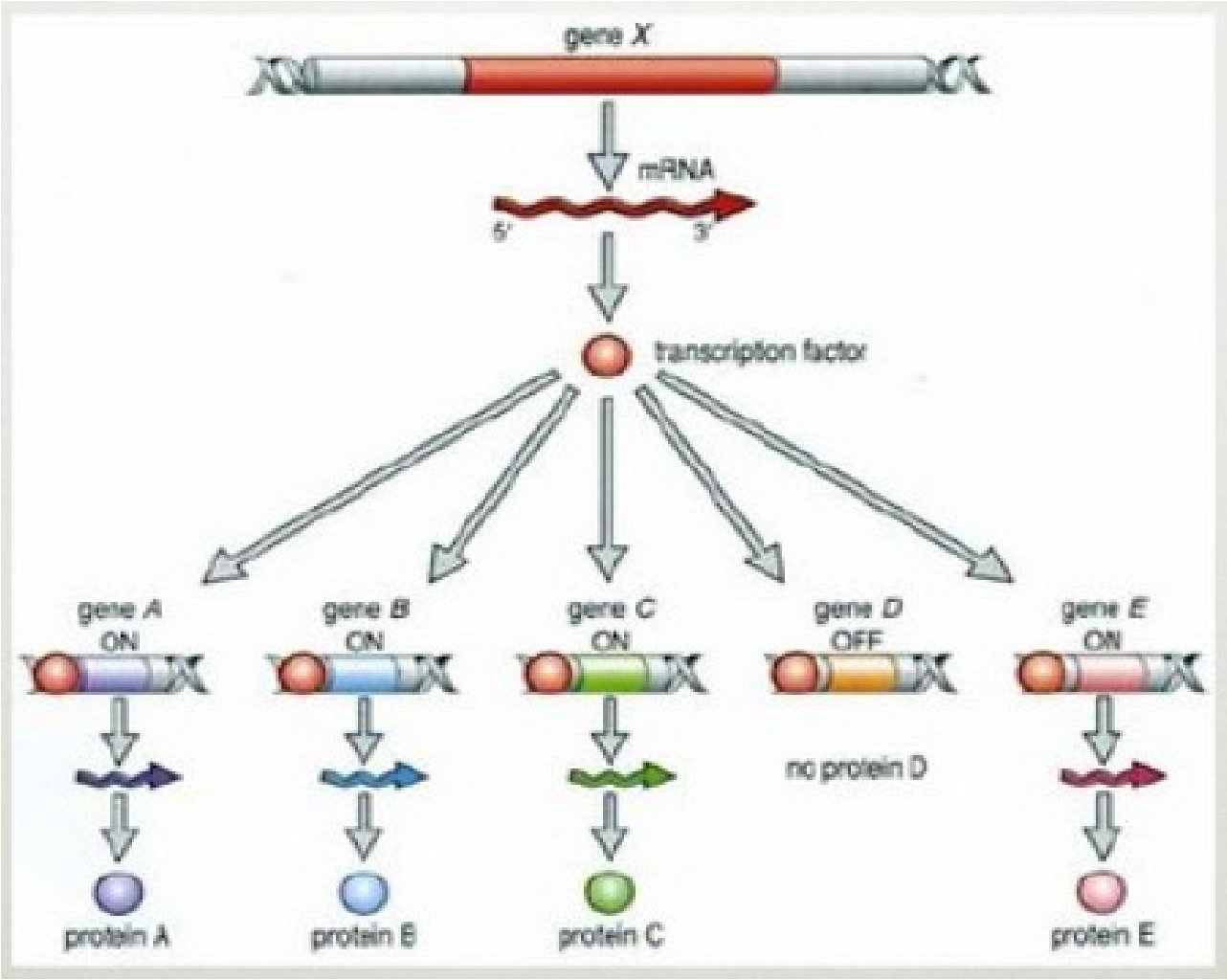


We have the sequence. What does it mean?

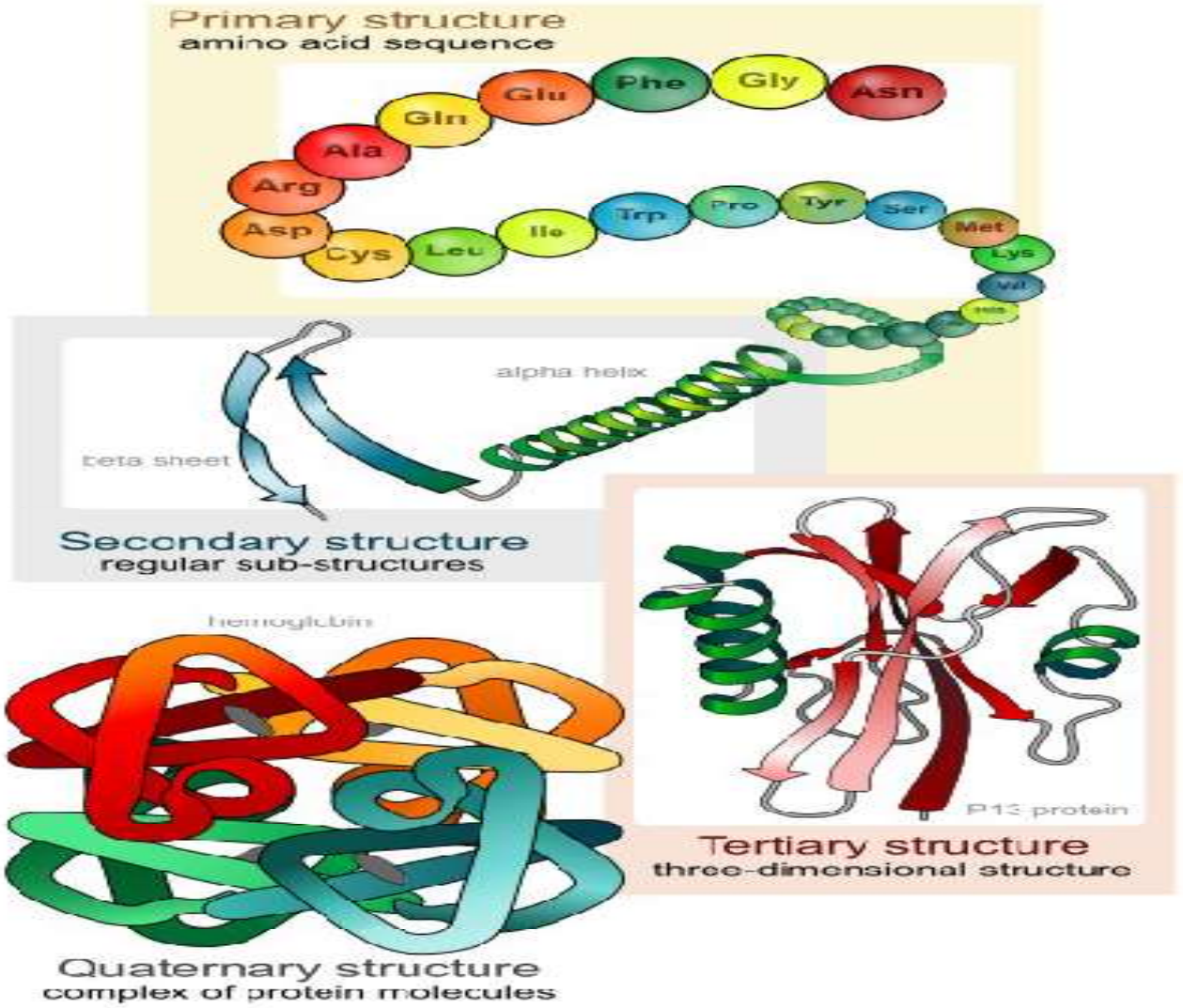
ACGTACCGCATTTAAAGTCACGTAAATCGGGTAA
AACCGATACACGCCATATTGAGAAGTCACGTAAC
TAAATCGGGTAACGATACCGCCATATGTTAAGTC
ACGTAAATCGGGTAACCGATACACGCCATATTGA
GAAGTCACGTAACTAAATCGGGTAAAACCGATAC
AAACCGATACACGCCATATTGAGAAGTCACGTAA
CTAAATCGGGTAAAACCGATACACGCCATATTGT
TAAGTCACGTAAATCGGGTAACCGATACACGCCA
TATTGAGAAGTCACGTAACTAAATCGGGTAAAAC

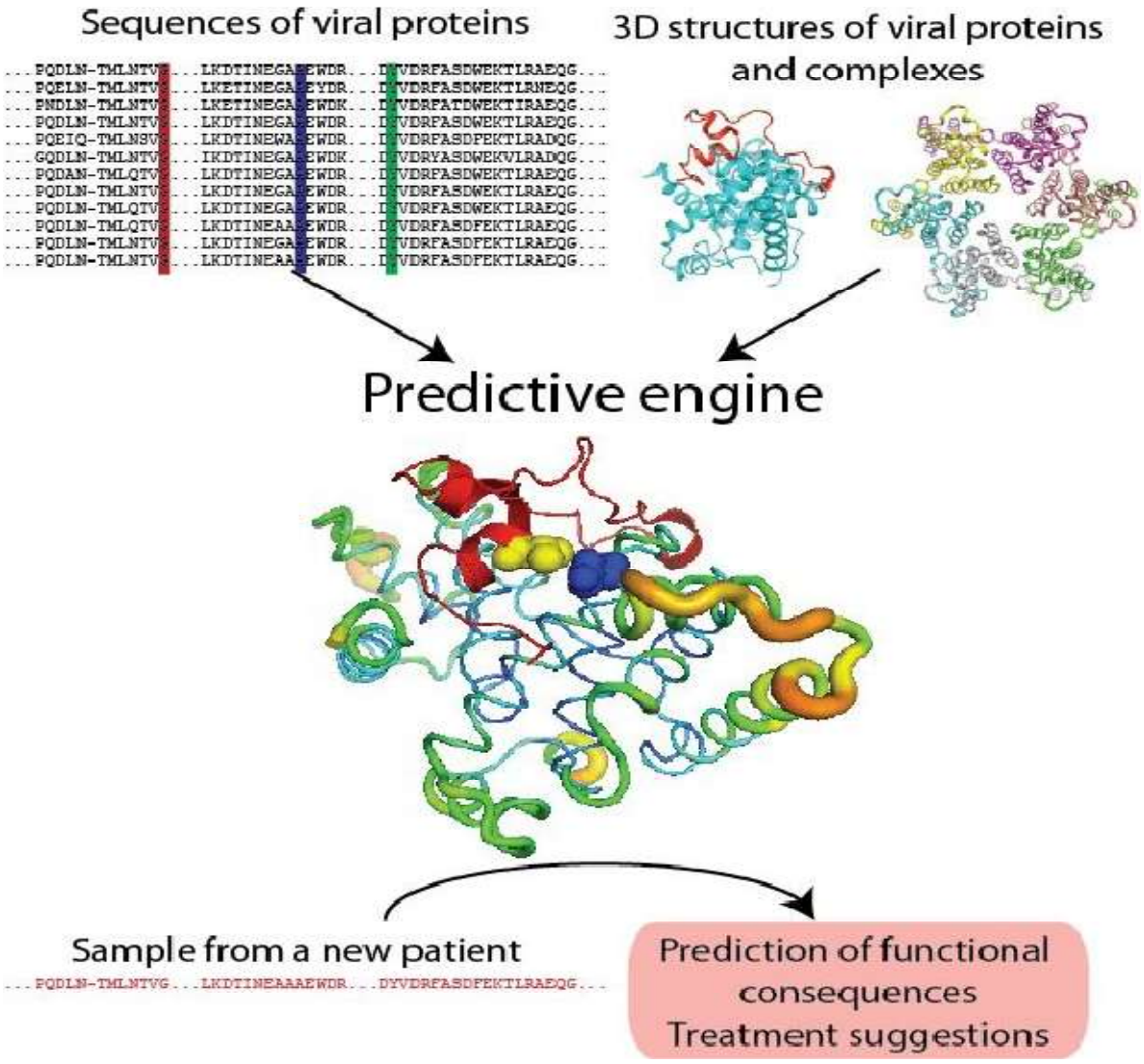

Human has 3.3 billion letters

Genes interact with each other



Protein structure and function




Sequences of viral proteins
3D structures of viral proteins and complexes
Predictive engine
Sample from a new patient
Prediction of functional consequences
Treatment suggestions

Applications of Bioinformatics

Knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology.

Computational studies of protein–ligand interactions

Knowledge of the three-dimensional structures of proteins

This informatics-based approach significantly reduces the time and cost necessary to develo
rugs with higher potency, fewer side effects, and less toxicity than using the traditional trial-and
rror approach.

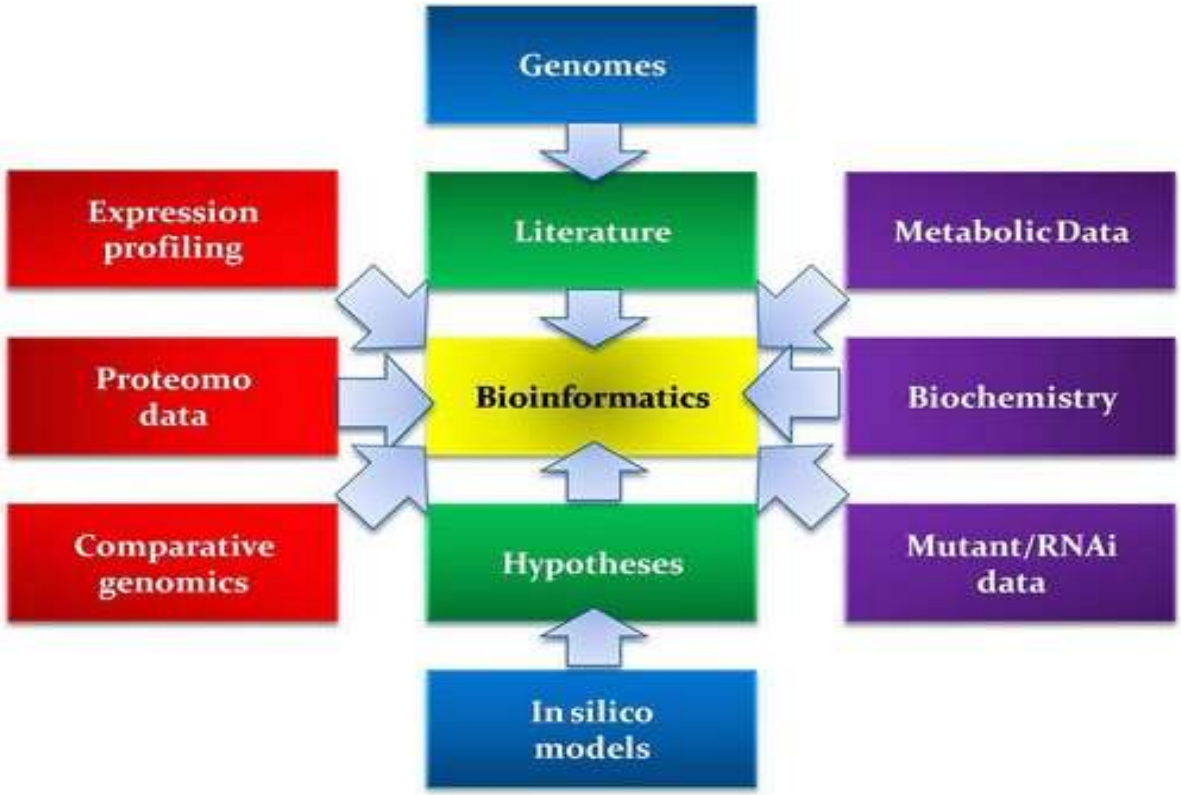
In forensics, results from molecular phylogenetic analysis have been accepted as evidence

Genomics and bioinformatics are now poised to revolutionize our healthcare system b
eveloping personalized and customizing medicine

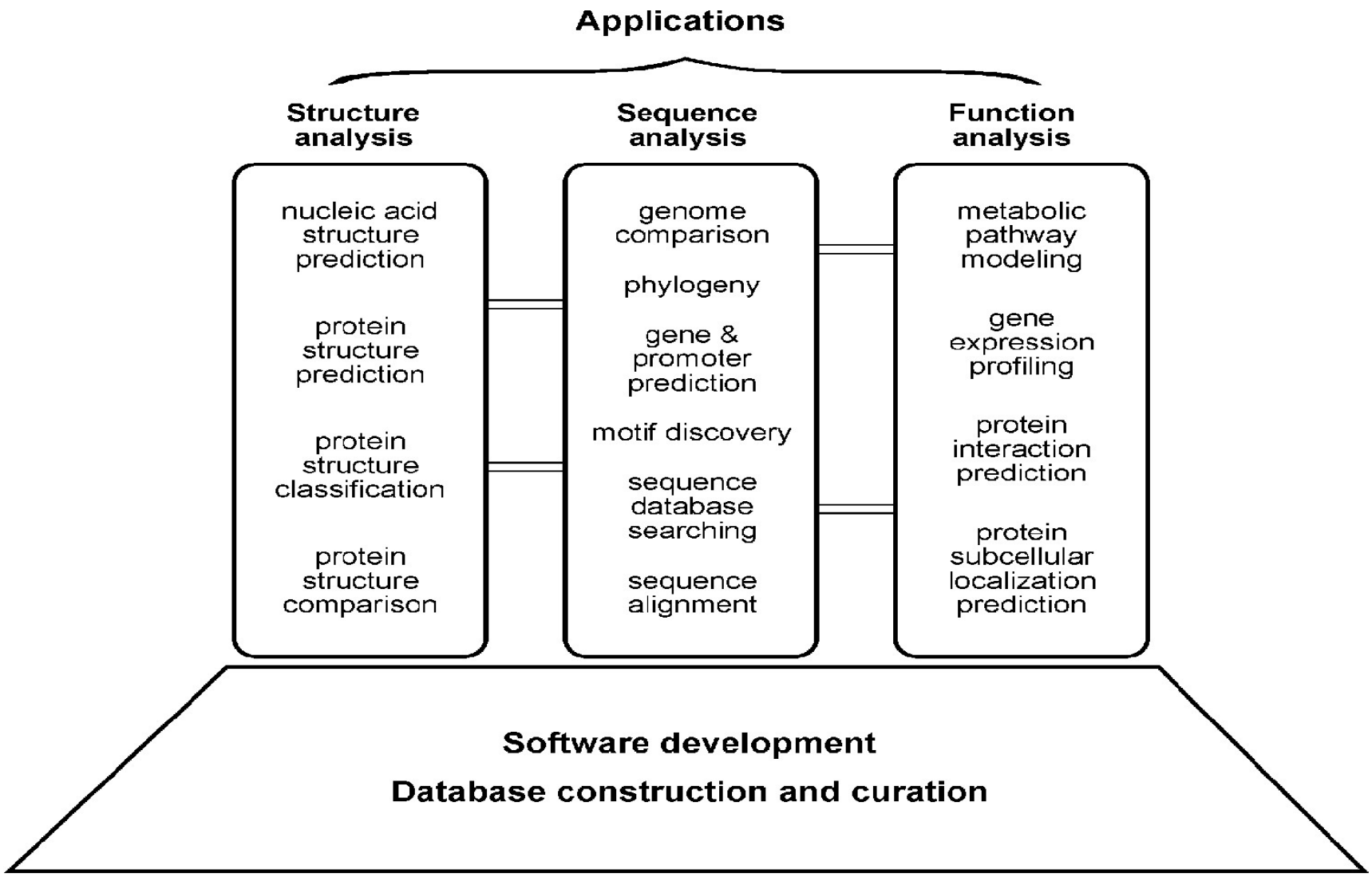
Applications of Bioinformatics

• Bioinformatics tools are being used in agriculture as well.

• Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance.



Applications of Bioinformatics



Limitations of Bioinformatics

- Bioinformatics is Bioinformatics and experimental biology are independent, but complementary, activities.
- Bioinformatics depends on experimental science to produce raw data for analysis. It, in turn, provides useful interpretation of experimental data and important leads for further experimental research.
- Quality of bioinformatics predictions depends on the quality of data and the sophistication of the algorithms being used.
- Sequence data from high throughput analysis often contain errors. If the sequences are wrong or annotations incorrect, the results from the downstream analysis are misleading as well.
- They often make incorrect predictions that make no sense when placed in a biological context. Errors in sequence alignment.