Subject: Bioinformatics

Lesson: Applications of Bioinformatics

Lesson Developer: Arun Jagannath

College/ Department: Sri Venkateswara College, University of Delhi

Table of Contents

Chapter: Applications of Bioinformatics

- **Lead Identification Strategies**
    - **Structure-based approaches- Virtual Screening by Docking**
    - **Ligand-based approaches**
- **Lead Optimization**

- **Bioinformatics and Metabolomics**
- **Systems Biology: Application & Future Prospects**
    - **From OMICS to systems biology**
    - **Complexity of complex disorders**
    - **Systems Pharmacology**
- **Summary**
- **Exercise/ Practice**
- **Glossary**
- **References/ Bibliography/ Further Reading**

Introduction

Bioinformatics is the application of IT to address a biological data. Bioinformatics helps us in understanding biological processes and involves development and application of computational techniques to analyse and interpret a biological problem. Major research efforts in the area of bioinformatics and computational biology include sequence alignment, genome annotation, prediction of protein structure and drug discovery. The challenges facing bioinformatics and areas of potential applications are shown below-

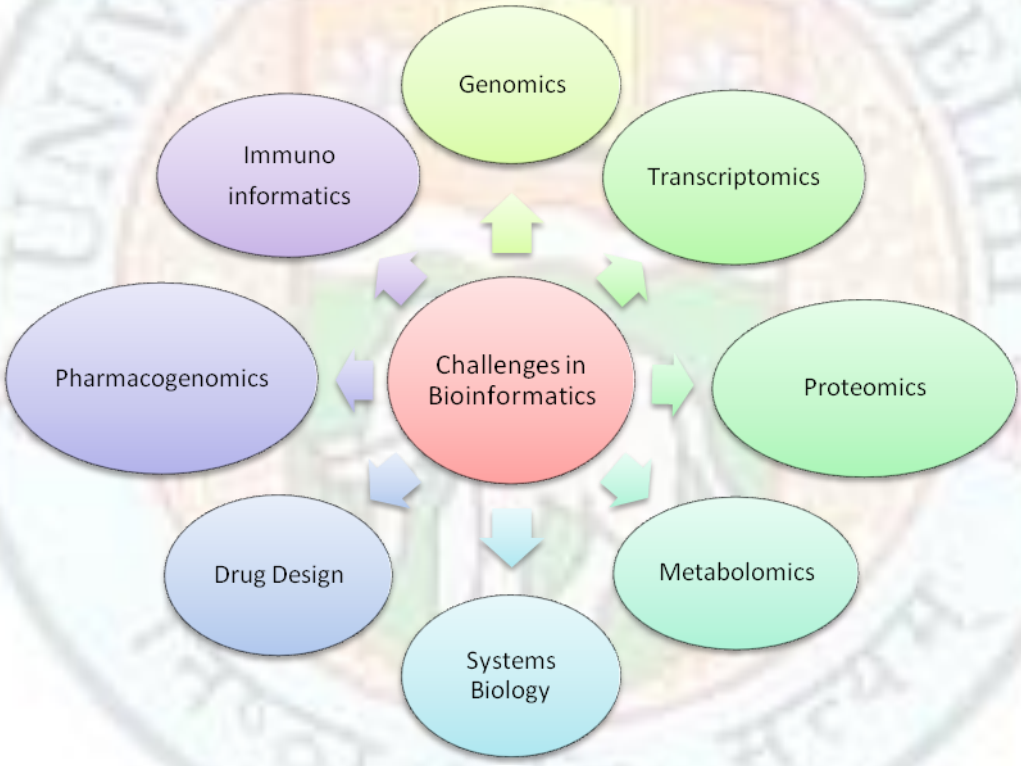

Figure : Challenges in bioinformatics

Source: Author

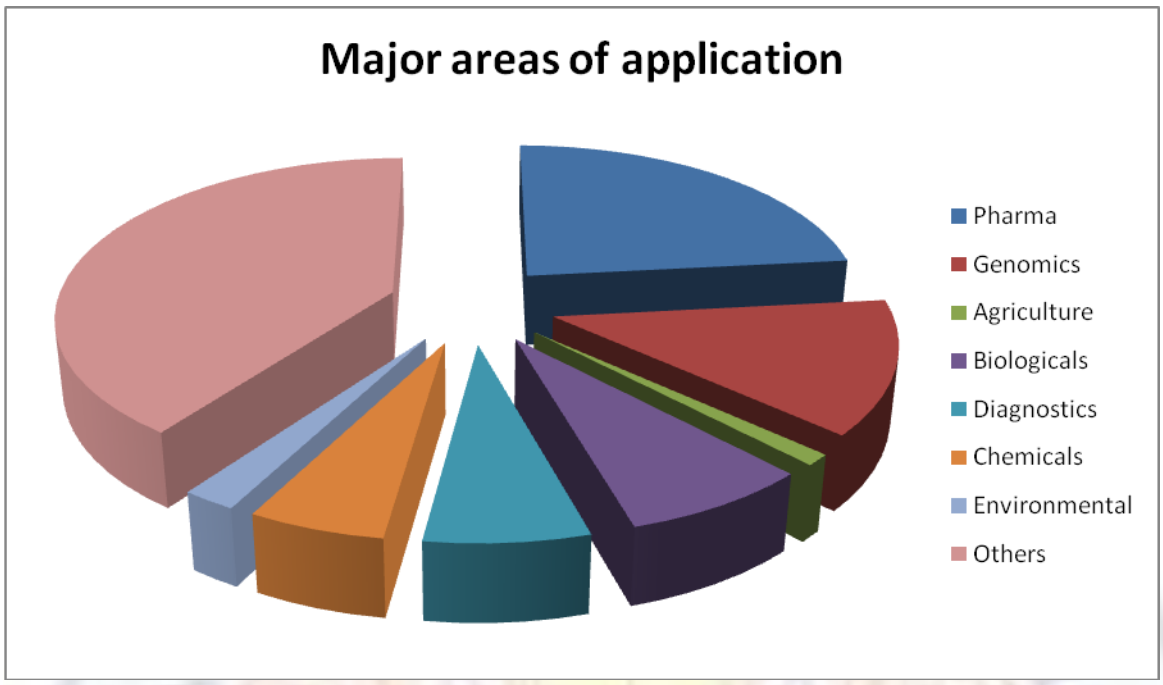

Figure: Major research areas in bioinformatics
Source: Author

Bioinformatics and Genomics

Genome and Genome Sequencing Projects

The word "genome" was coined by Hans Winkler from the German "Genom" in as early as 1926. The total DNA present in a given cell is called genome. In most cells, the genome is packed into two sets of chromosomes, one set from maternal and another one set from paternal inheritance. These chromosomes are composed of 3 billion base pairs of DNA. The four nucleotides (letters) that make up DNA are A, T, G, and C. Just like the alphabets in a sentence in a book make words to tell a story, same do letters of the four bases – A, T, G, C in our genomes.

Genomics is the study of the genomes that make up the genetic material of organism. Genome studies include sequencing of the complete DNA sequence in a genome and also include gene annotation for understanding the structural and functional aspects of the genome.

Genes are the parts of your genome that carry instructions to make the molecules, such as proteins that are responsible for both structural and functional aspects of our cells. The first organism that was completely sequenced was *Haemophilus influenzae* in 1995 that led to sequencing of many more organisms from both prokaryotic and eukaryotic world.

Table : Genome sizes of some organisms

Source: Author

Organism	Approximate Size (base pairs)	Number of genes estimated	Number of chromosomes
Homo sapiens (human)	3.2×10^9	~25,000	46
Mus musculus (mouse)	2.6×10^9	~25,000	40
Drosophila melanogaster (fruit fly)	137×10^6	13,000	8
Arabidopsis thaliana (plant)	100×10^6	25,000	10
Caenorhabditis elegans (roundworm)	97×10^6	19,000	12
Saccharomyces cerevisiae (yeast)	12.1×10^6	6000	32
Escherichia coli (bacteria)	4.6×10^6	3200	1
H. influenzae (bacteria)	1.8×10^6	1700	1

The Human Genome Project (HGP)

The Human Genome Project (HGP) was global effort undertaken by the U.S. Department of Energy and the National Institutes of Health with a primary goal of determining the

complete genome sequence in a human cell. It also aimed at identifying and mapping the genes and the non-genes regions in the human genome.

Some key findings of the draft (2001) and complete (2004) human genome sequences included

1. Total number of genes in a human genome was estimated to be around 20,500.
2. Gene expression studies helped us in understanding some diseases and disorders in man.
3. Identification of primate specific genes in the human genome.
4. Identification of some vertebrate specific protein families.
5. The role of junk DNA was being elucidated.
6. It is estimated that only 483 targets in the human body accounted for all the pharmaceutical drugs in the global market.

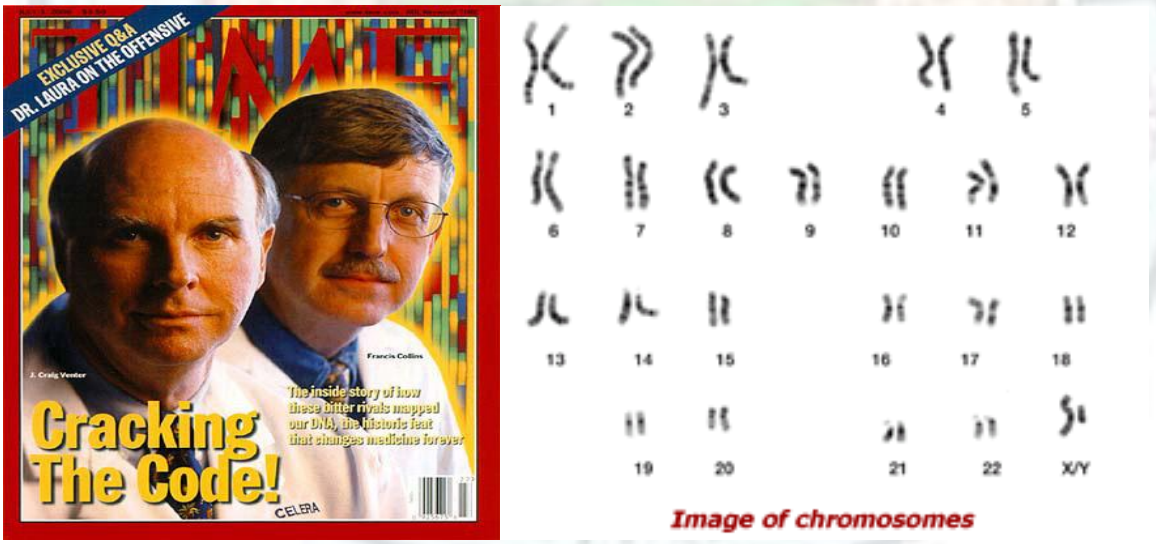

Figure: The Human Genome Project
Source: TIME Magazine

How was the whole genome sequenced?

The human genome was sequenced by two different methods – Hierarchical Genome Shotgun (HGS) Sequencing and Whole Genome Sequencing (WGS)

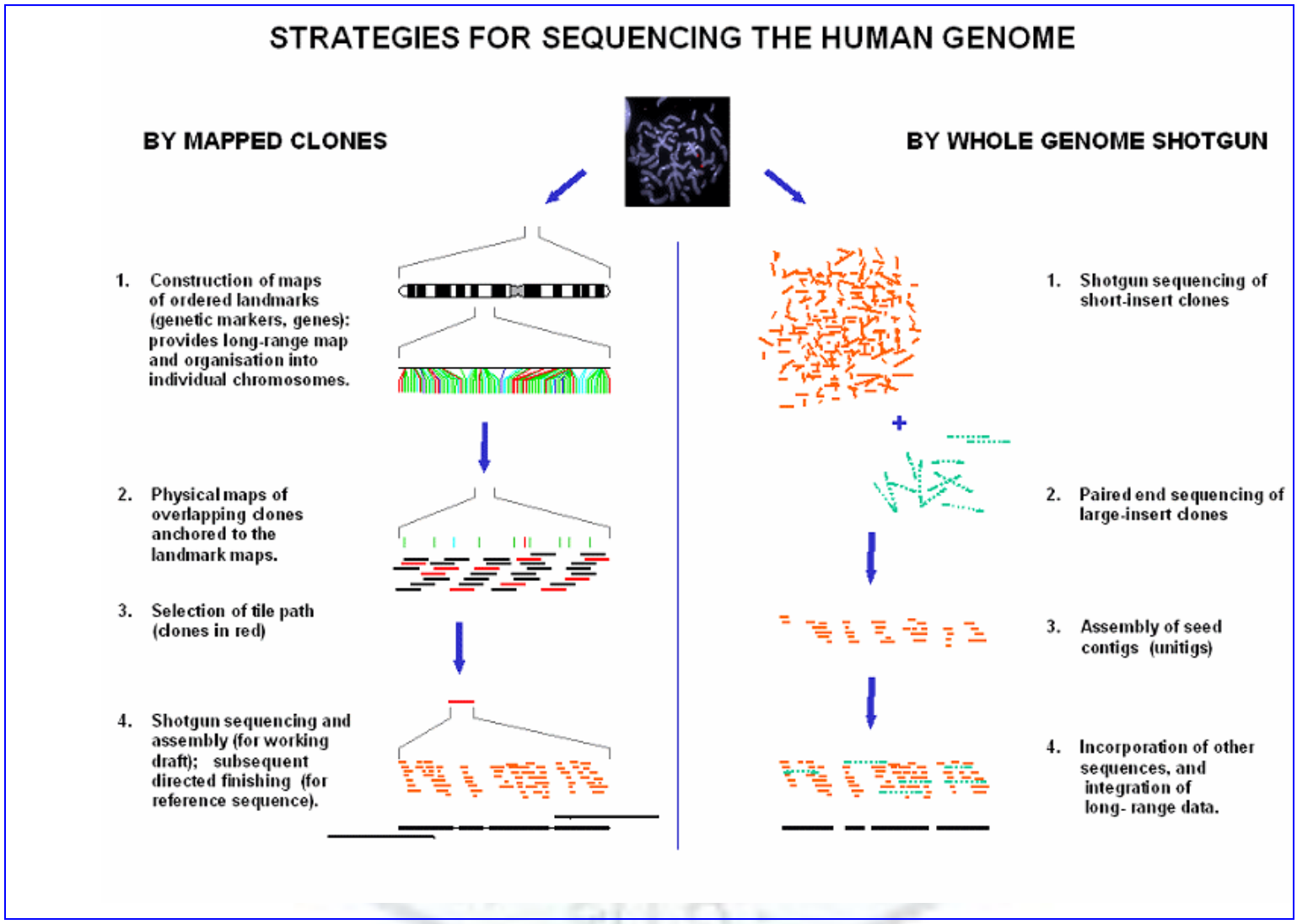

Figure: Two approaches for genome sequencing in the Human Genome Project

Why do we want to determine the sequence of DNA of an organism?

1. Genome variation among individuals in a population can lead to new ways to diagnose, treat, and someday prevent the thousands of disorders that affect mankind.
2. Genome studies help us to provide insights into understanding human disease biology.
3. Studies on nonhuman organisms' DNA sequences can contribute to solving challenges in health care and agriculture. Understanding the sequence of genomes can provide insights in the identification of unique and critical genes involved in the pathogenesis of microorganisms that invade us and can help identifying novel drug targets to offer new therapeutic interventions. Increasing knowledge about genomes of plants can reduce costs in agriculture, for example, by reducing the need for pesticides or by identification of factors for development of plants under stress.
4. HGP studies also included application of research on the the ethical, legal and social implications (ELSI) of the genomic research for individuals and communities.

Where are the genome data stored?

The genome sequence and the genes mapped are stored in databases available freely in the Internet. The National Centre for Biotechnology Information (NCBI) is a repository of the gene/protein sequences and stores in databases like GenBank. This large volume of biological data is then analyzed using computer programs specially written to study the structural and functional aspects of genome.

Prediction Methods

Computational approaches for prediction of genes is one of the major areas of research in bioinformatics. Finding genes by the traditional molecular biology techniques becomes time consuming process.Two classes of prediction methods for identifying genes from non genes in a genome are generally adopted: similarity or homology based searches and *ab initio* prediction.

Gene discovery in prokaryotic genomes becomes less time consuming as compared to prediction of protein coding regions in higher eukaryotic organisms due to the absence of intervening sequences called introns.

Table: List of some gene prediction softwares

Source: Author

Homology based gene prediction tools			
Name	**Algorithm**	**Organism**	**Url**
GeneWise	Dynamic Programing	Human	http://www.sanger.ac.uk/resources/software/
ORFgene2	Dynamic Programing	Human,mouse,Drosophila,Arabidopsis	http://www.itb.cnr.it/sun/webgene/
PredictGenes	Dynamic Programing	Invertebrates,vertebrates,plants	http://www.cbrg.ethz.ch/Server
PROCRUSTES	Dynamic Programing	Vertebrates	http://www-hto.usc.edu/software/procrustes/

***Ab initio* based gene prediction tools**			
Name	**Algorithm**	**Organism**	**Url**
GeneMark	Hidden markov Model	Prokaryotes,eukaryotes	http://opal.biology.gatech.edu/GeneMark/
GENEFINDER	Dynamic Programing	Human,mouse,Drosophila, yeast	http://rulai.cshl.edu/tools/genefinder/
GENSCAN	Hidden markov Model,Dynamic Programing	Vertebrates,maize,Arabidopsis	http://genes.mit.edu/GENSCANinfo.html
GRAIL	Dynamic Programing	Human , mouse,	compbio.ornl.gov/Grail-bin/EmptyGrailForm

	,Neural Network	Arabidopsis, Drosophila	
HMMgene	CHMM	Vertebrates, *C.elegans*	http://www.cbs.dtu.dk/services/HMMgene/hmmgene1_1.php
ChemGenome	Physiochemical Model	Prokaryotes, Eukaryotes	http://www.scfbio-iitd.res.in/chemgenome
Genie	Hidden markov Model,Dynamic Programing	Drosophila,human	http://www.fruitfly.org/seq_tools/genie.html
GeneParser	Dynamic Programing ,Neural Network	Vertebrates	http://home.cc.umanitoba.ca/~psgendb/birchdoc/package/GENEPARSER.html

Comparative genomics and Functional Genomics

Comparative genomics is the analysis and comparison of genomes from two or more different organisms. Comparative genomics is studied to gain a better understanding of how a species has evolved and to study phylogenetic relationships among different organisms. One of the most widely used sequence similarity tool made available in the public domain is Basic Local Alignment Search Tool (BLAST). BLAST is a set of programs designed to perform sequence alignment on a pair of sequences (both nucleotide and protein sequence).

Figure: Comparative Genomics

Source: Katherine S. Pollard (2009) What Makes Us Human? Comparisons of the genomes of humans and chimpanzees are revealing those rare stretches of DNA that are ours alone. Scientific American

Functional genomics attempts to study gene functions and interactions. Functional genomics seeks to address questions about the function of DNA at the levels of genes, RNA at the levels of transcription and proteins at the structural and functional levels.

Pharmacogenomics

Pharmacogenomics analyzes how the genetic constitution affects a person's response to drugs and help us in the creation of personalized medicine to create and design drugs based on an individual's unique genetic makeup. Pharmacogenomics is used for the development of drugs to treat a wide range of health problems including diabetes, cancer, cardiovascular disorders, HIV, and tuberculosis.

Metagenomics
Metagenomics is a new and exciting field of biology for understanding biological diversity and offers a new view for looking at the microbial world. It is also referred to as

environmental genomics or community genomics. It provides solutions to fundamental questions in microbial ecology and genomic analysis of microorganisms.

INTERESTING FACTS ON GENOMICS

- Every cell of the human body contain complete set of DNA that make up the genome with the exception of egg and sperm cells that carry half of human genome.
- There are cells like red blood cells which have no DNA at all.
- The sequencing of the human genome was completed in 2003. Both female (blood) and male (sperm) samples were processed for human genome sequencing project.
- Genetic variation among the human, chimpanzee and gorilla shows that humans are more chimp-like than gorillas.
- A major part of our DNA whose function is unknown is referred to as junk DNA.
- The human genome is 3 billion bases of DNA made into 46 chromosomes (23 pairs autosomes & 1 pair of sex chromosome). It would take a century to just recite the complete sequence if done at a rate of one letter per sec for 24 hours a day.
- Our DNA differs from each other by only 0.2 percent (1 in 500 bases).

Next Generation Sequencing

The advancement of the field of molecular biology has been principally due to the capability to sequence DNA. Over the past eight years, massively parallel sequencing platforms have transformed the field by reducing the sequencing cost by more than two folds. Previously, Sanger sequencing ('first-generation' sequencing technology) has been the sole conventional technique used to sequence genomes of several organisms. In contrast, NGS platforms rely on high-throughput massively parallel sequencing involving unison sequencing of millions of DNA fragments from a single sample. The former facilitates the sequencing of an entire genome in less than a day. The speed, accessibility and the cost of newer sequencing technologies have accelerated the present –day biomedical research.

These technologies reveal large scale applications outspreading even genomic sequencing. The most regularly used NGS platforms in research and diagnostic labs today have been- the Life Technologies Ion Torrent Personal Genome Machine (PGM), the IlluminaMiSeq, and the Roche 454 Genome Sequencer. NGS platforms rapidly generate sequencing read data on the gigabase scale. So the NGS data analysis poses the major challenge as it can be time-consuming and require advanced skill to extract the maximum accurate information from sequence data. A massive computational effort is needed along with in-depth biological knowledge to interpret enormous NGS data.

Table : Next-Generation Sequencing Platforms

Source:

Next-generation sequencing technologies employ different techniques, but all have in common, the ability to sequence more DNA base pairs per sequencing run than earlier methods like Sanger sequencing.

Manufacturer	Technique	Run Time Per Read	Base Length	Cost (in 000s)
Helicos	ReversibleTerminator	8 days	32	$999
Illumina	ReversibleTerminator	4–9 days	75–100	$500–900
Ion Torrent	Real-time	<1 day	964	$50
Roche/454	Pyrosequencing	<1 day	330	$500–700
SOLiD	Sequencing By Ligation	7–14 days	50	$600–700

Adapted from Mol Ecol Resour 2011;11:759–69; Nat Rev Genet 2010;11:31–46; Am J Clin Path 2011;136:527–39.

Bioinformatics and Protein Structure Prediction

Proteins are linear polymer of amino acids joined by peptide bonds. Every protein adopts a unique three-dimensional structure to form a native state. It is this native 3D structure

that confers the protein to carry out its biological activity.Proteins play key roles in almost all the biological process in a cell. Proteins are important for the maintenance and structural integrity of cell.

Levels of protein architecture

There are four levels of protein structure. The primary structure of a protein is the arrangement of linear sequence of amino acids. The patterns of local conformation within the polypeptide are referred to as as secondary structure. The two most common types of secondary structure occurring in proteins are *α*-helices and *β*-sheets. These secondary structures are connected by loop regions. The tertiary structure represents the overall three dimensional structure of these elements and the protein folds into its native state. The quaternary structure includes the structure of a multimeric protein and interaction of its subunits. Figure illustrates the hierarchy in protein structure.

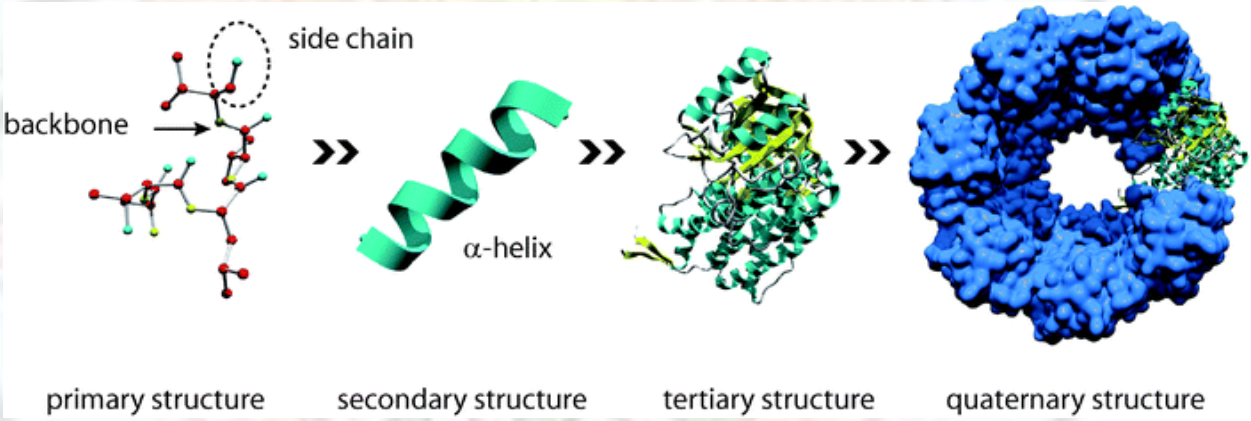

Figure: Levels of protein structural hierarchy

Source: DOI: 10.1039/B813273A (Tutorial Review) Chem. Soc. Rev., 2010, 39, 156-164

Explosion in the growth of Biological Sequence and Structure Data

Experimental determination of the tertiary structure of proteins involves the use of X-ray crystallography and NMR. In addition, computational techniques are exploited for the structural prediction of native structures of proteins. There has been an exponential growth of both the biological sequence and structure data, mainly due to the genome sequencing projects underway in different countries around the world. As of Oct 2013, there are 94,540 structures in the protein data bank (RCSB-PDB).

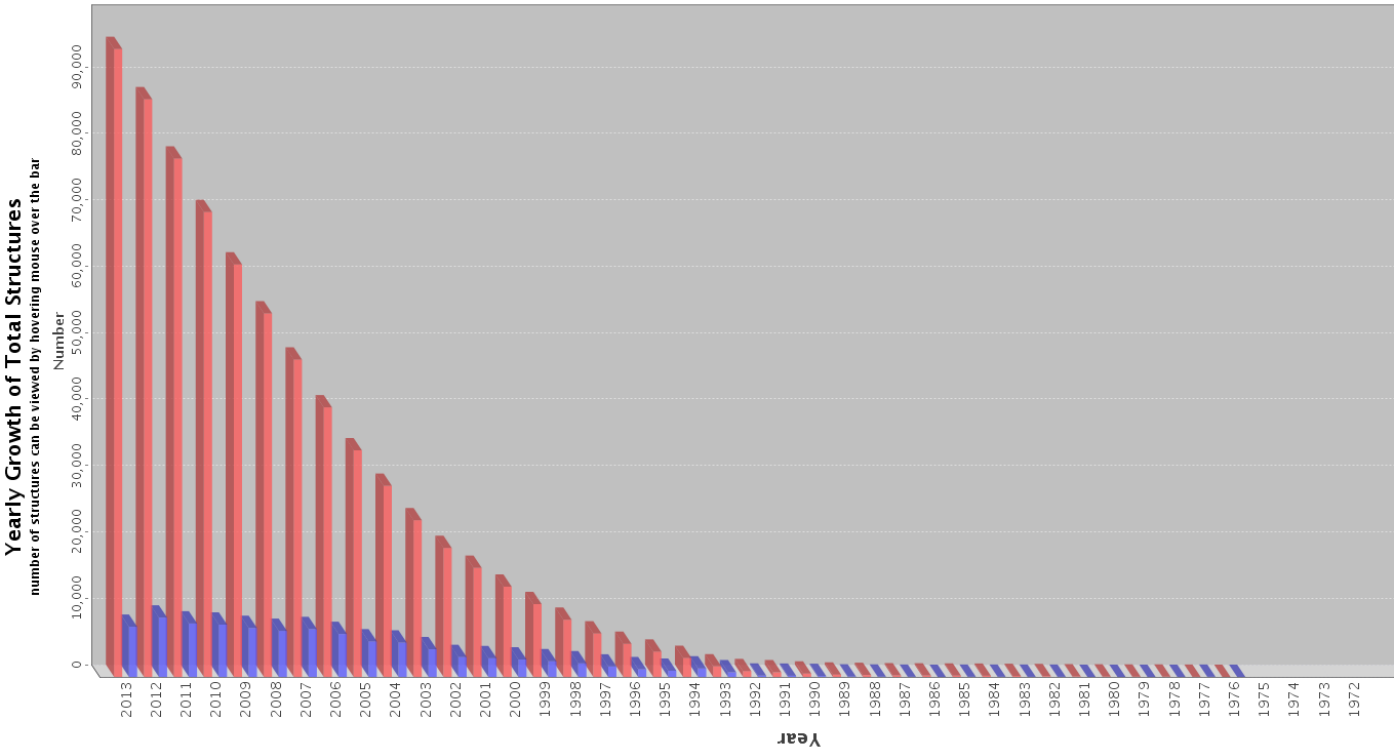

Figure: Growth of structures in PDB. The red bar indicates the growth of structures totally while blue bar indicates the number of structures in PDB in that particular year.

Source: www.rcsb.org

Computational approaches to protein structure prediction

There are three different methods of protein 3D structure prediction using computational approaches

1. Comparative Protein Modeling or Homology Modeling

Homology modeling predicts the structure of a protein based on the assumption that homologous proteins share very similar structure, as during the course of evolution, structures are more conserved than amino acid sequences. So a model is generated based on the good alignment between query sequence and the template. In general we can predict a model when sequence identity is more than 30%. Highly homologous sequences will generate a more accurate model.

Table: Some protein structure prediction softwares/tools

Tool	Prediction method
3D-JIGSAW	Homology Modeling
CPHModel	Homology Modeling
SWISSMODEL	Homology Modeling
ESyPred3D	Homology Modeling
MODELLER	Homology Modeling
PHYRE	Threading or Fold Recognition
BHAGEERATH	*Ab-initio* method
I-TASSER	*Ab-intio* method
ROBETTA	*Ab-intio* method
Rosetta@home	*Ab-intio* method

2. Protein Threading

If two sequences show no detectable sequence similarity, threading or fold recognition is employed to model a protein. Threading predicts the structure for a protein by matching its sequence to each member of a library of known folds and seeing if there is a statistically significant fit with any of them.

3. *Ab initio* method

Ab initio protein modeling is a database independent approach based exploring the physical properties of amino acids rather than previously solved structure. Ab-initio modeling takes into consideration that a protein native structure has minimum global free energy.

Bioinformatics in Drug Discovery and Development

Drug Discovery & Development: A Difficult and Expensive problem

Drug discovery and development is costly and time-consuming process. It involves a multidisciplinary effort to design novel and effective drugs. It takes around 14 years for a drug to enter the market with a cost of around 800 million US dollars.

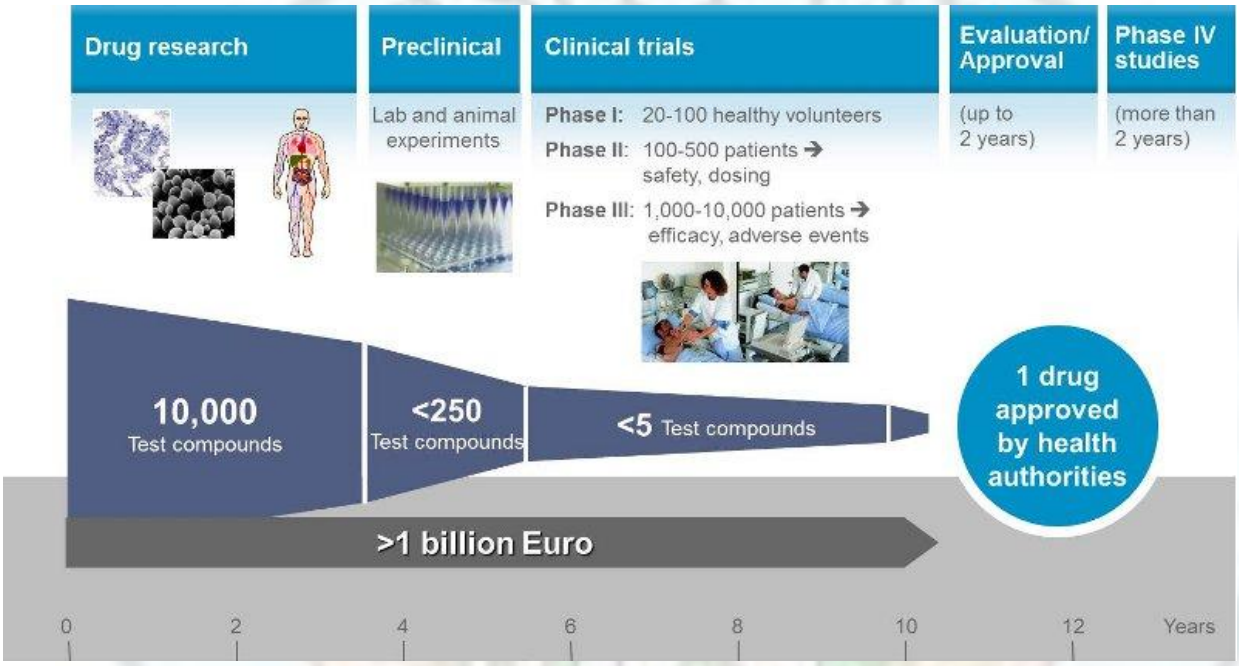

Figure : Time and the cost involved in a drug discovery process

Source: (http://www.bayerpharma.com/en/research-and development/processes/index.php)

In order to cut down the cost and time involved in the process of drug discovery, design, development and optimization, there is a growing need to apply computational approaches in various stages of drug-discovery pipeline.

Where are Computational Techniques used?

Computer-aided drug design (CADD) is a popular term that describes many computational approaches used at various stages of a drug design project. It constitutes development of online repositories of the chemical compounds for generation of hits, programs for prefiltering compounds with remarkable physicochemical characteristics, as well as tools for systematic assessment of potential lead candidates before they are synthesized and tested in animal models.

Target Identification and Validation

The identification of new drug targets implicated in disease remains one of the major challenges in the drug discovery process. Target identification can be carried out by classical biochemical methods or computational systems biology approaches. Target validation includes evaluating a biomolecule physiologically and pharmacologically and also at the molecular, cellular, or whole organism level. It has been reported that all current drugs with a known mode-of-action act through 324 distinct molecular drug targets.

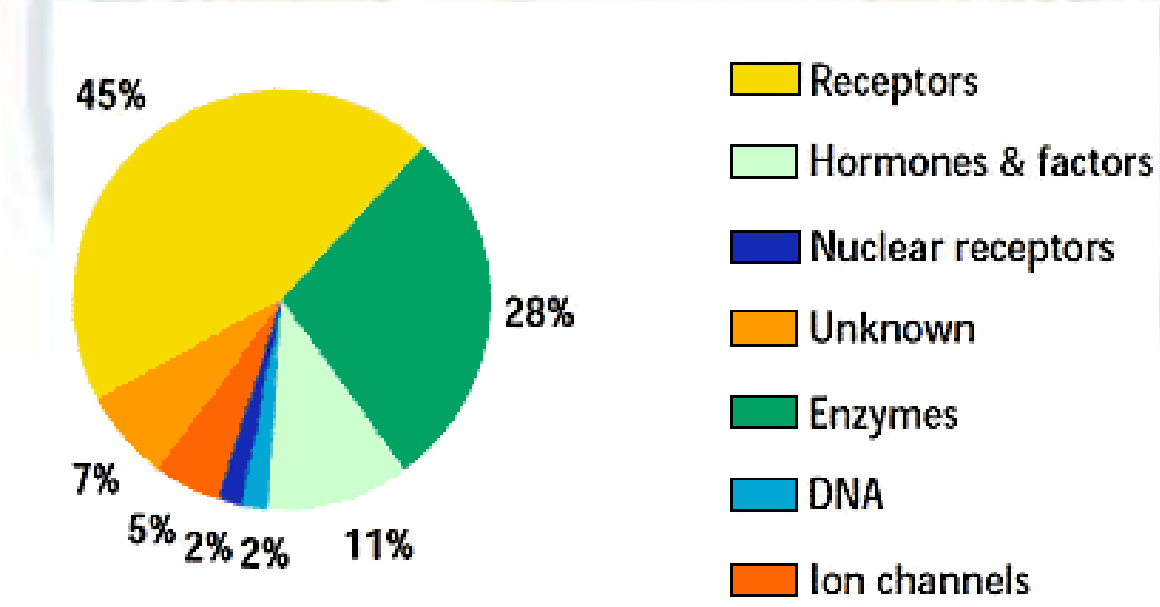

Figure : Classification of drug targets**.** More than 60% of the drug targets are membrane receptor proteins and enzymes.

Source: Author

Table: Some enzymes as drug targets and drugs developed.

Enzymes	Drugs
Cyclooxygenase	Aspirin
Angiotensin converting enzyme	Captopril
Dihydrofolate reductase	Methotrexate
HIV protease	Saquinavir
Xanthine Oxidase	Allopurinol
Carbonic anhydrase	Acetazolamide
Reverse Transcriptase	AZT(Retrovir)

Target Structure Prediction

The drug targets generally selected for drug discovery are proteins. Most of the structure of proteins has been determined experimentally by X-ray crystallography or NMR spectroscopy. The structure of the protein target can also be modeled computationally using one or combination of the three approaches- Homology Modelling, Threading and ab-initio approaches.

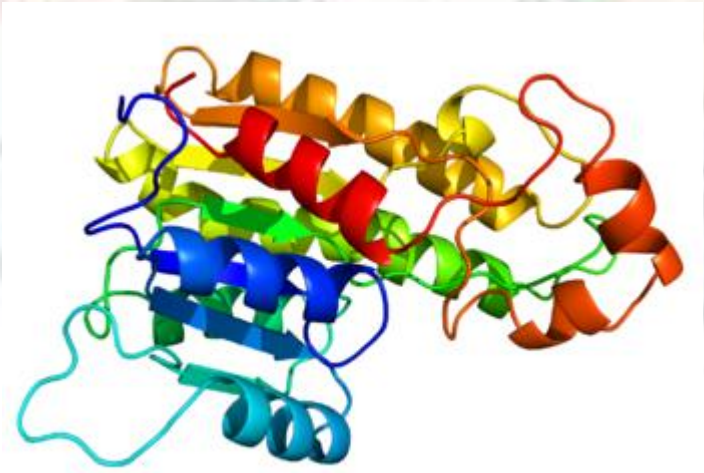

Figure: A model of the dehydrogenase reductase SDR family 7B (DHRS7B) protein predicted by homology modeling (SWISS-MODEL).

Source: Author

Binding Site Identification and Characterization

Identification of binding site in a protein target structure play a very crucial role for protein-ligand interactions and structure based design endeavours. Many computational methods have been developed for the binding site prediction with the given a 3D structure of a protein. These methods probe the protein surface for cavities or pockets that are most likely to represent binding site. Prediction methods can be divided into two categories: energy-based and geometry-based methods. Energy-based methods find pockets by computing the interaction energy between protein atoms and a small-molecule probe. By contrast, geometry-based methods are based on the assumption that binding sites for ligands are located in the crevices on the protein surface.

Table : Main algorithms used for prediction of binding sites along with some binding site prediction softwares and tools

Method type	Software/Tool	URL
Geometric	CASTp	http://sts.bioengr.uic.edu/castp/http://sts.bioengr.uic.edu/castp/
Geometric	LigASite	http://www.bigre.ulb.ac.be/Users/benoit/LigASite/index.php?home
Geometric	CAVER	http://caver.cz/
Geometric	FPocket	http://fpocket.sourceforge.net/
Geometric	McVol	http://www.bisb.uni-bayreuth.de/index.php?page=data/mcvol/mcvol http://metapocket.eml.org/
Geometric	SURFNET-Consurf	http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html
Geometric	VOIDOO	http://xray.bmc.uu.se/usf/voidoo.html
Geometric & Energy based	SiteMap	http://www.schrodinger.com/productpage/14/20/ http://cssb.biology.gatech.edu/findsite
Energy based	SITEHOUND	http://scbx.mssm.edu/sitehound/sitehound-web/Input.html http://bioinfo3d.cs.tau.ac.il/MolAxis/
Energy based	Autoligand	http://autodock.scripps.edu/resources/autoligandhttp://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.ht

		ml
Energy based	GRID	http://www.moldiscovery.com/soft_grid.php http://xray.bmc.uu.se/usf/voidoo.html
Energy based	Q-SiteFinder	www.modelling.leeds.ac.uk/qsitefinder/

.

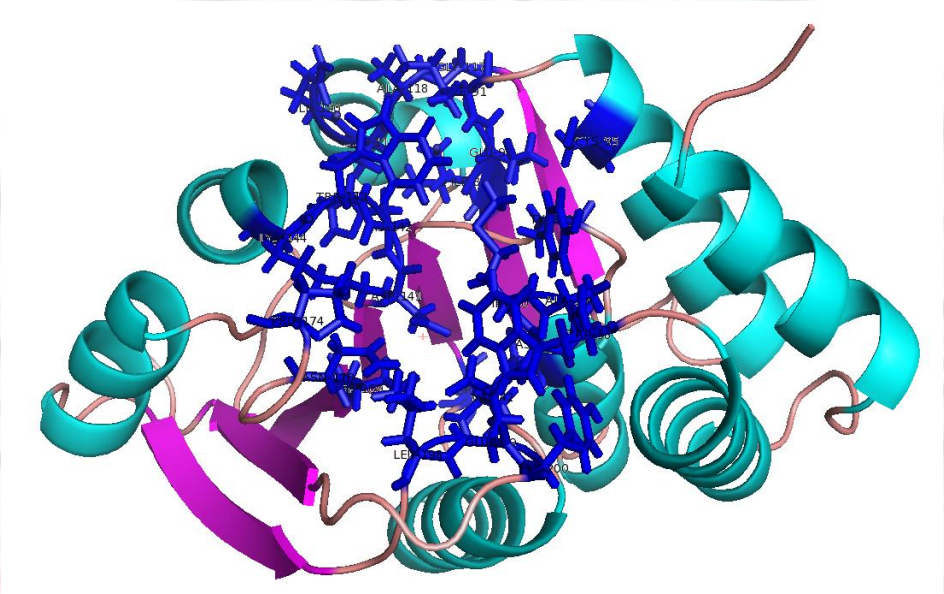

Figure : Binding site prediction of human Catechol-O-Methyl transferase. The binding site residues are shown in blue.

Source: Author

Lead Identification Strategies

Computational approaches for drug design are broadly divided into two categories: Structure-based approaches and ligand-based approaches.

Structure-based approaches- Virtual Screening by Docking

Once the structure and binding site of the target protein is known, there are several ways to generate a lead molecule based on the target which include virtual screening, and de novo generation.

Virtual screening is an approach which is widely used for structure-based drug design where large library of compound databases like DrugBank, ZINC is screened against a drug target in order to identify those chemical structures which are most likely to bind well within the active site of target protein. In structure-based virtual screening, ligands are docked into the active site of protein where a scoring function is used to estimate the binding affinity of the ligand to the protein. Knowing the preferred orientation of the ligand in receptor helps to predict the strength of association or binding affinity or binding free energy. The best binding pose has the minimal binding energy which is a function of several terms and is given by:

$$\mathbf{\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{tor} + \Delta G_{sol}}$$

where ΔG_{bind} is binding free energy, ΔG_{vdw} is van der Waals interaction energy, ΔG_{hbond} is H-bonding potential, ΔG_{elec} is electrostatic interaction energy, ΔG_{tor} is no. of rotatable bonds and ΔG_{sol} is solvation binding energy.

Lead Evaluation

Lead structures are also evaluated for their likelihood to be orally bioavailable using the "Lipinsky's Rule of 5". Lipinsky's rule of 5 states that lead molecules generally have less than five hydrogen bond donors and less than ten hydrogen bond acceptors, a molecular weight less than 500, and a calculated log of the partition coefficient (clogP) less than 5. Additional filters such as number of rotatable bonds (which should be less than ten), number of chiral centres, stability and the ease of synthesis, can also factor into the decision to proceed with a particular candidate lead. Finally, leads are taken up for synthesis and biochemical evaluation.

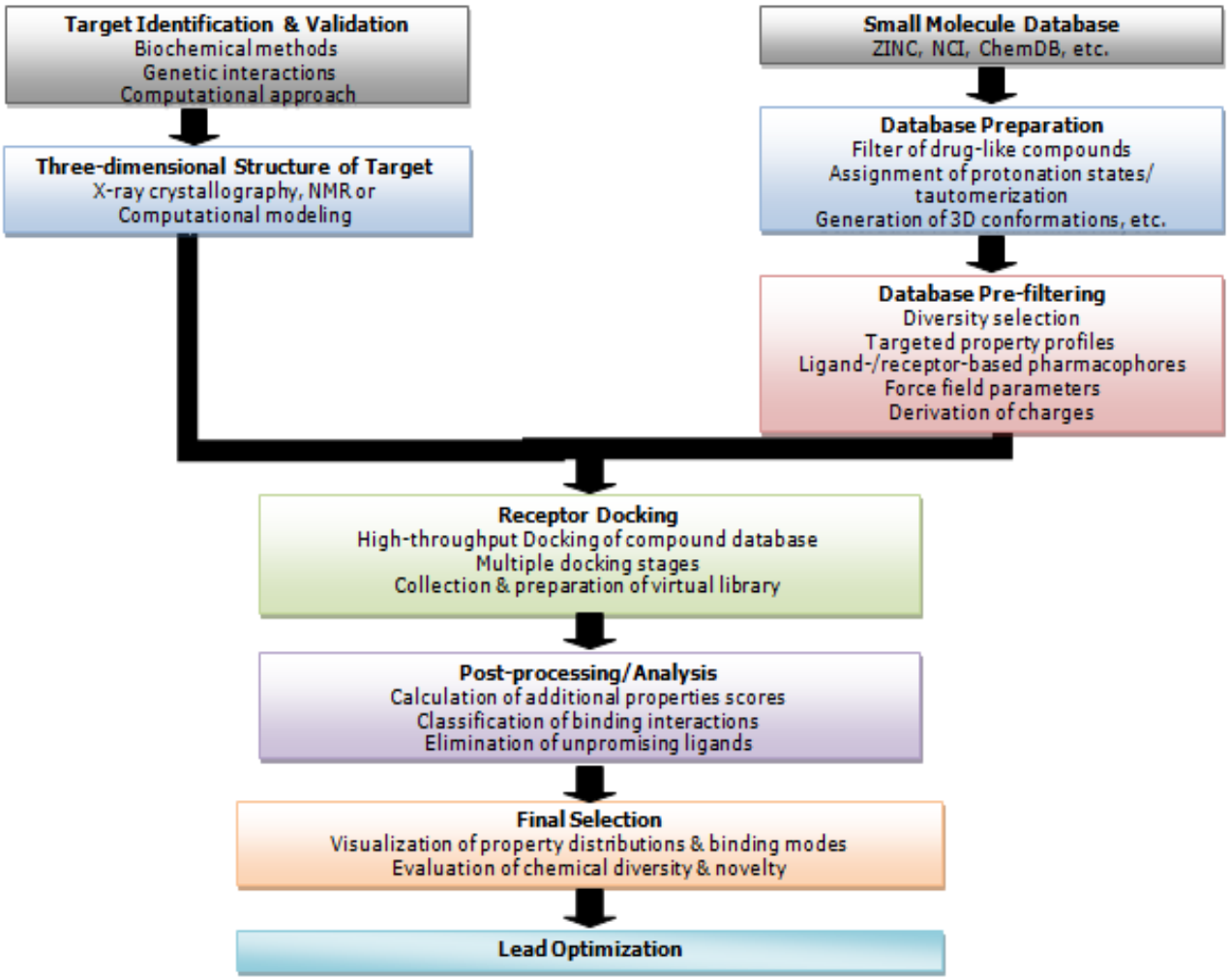

Figure : Virtual screening pipeline

Source: Author

Table : List of few protein-ligand docking softwares

Source: Author

Software	**Methodology**
AutoDock	AutoDock is a popular docking program in C used to predict the bound conformations of a small, flexible ligand to a macromolecular target whose structure is known.

FlexX	Docking process used Pattern recognition technique called pose clustering algorithm.
GOLD	Genetic optimization algorithm for automated ligand docking.
Patch Dock	Patch Dock is more commonly used for generation of complexes based on protein-protein docking and protein-lgand docking.
DOCK	An algorithm to address both rigid body and flexible docking
ParDOCK	An all-atom energy based Monte Carlo, docking method for rigid docking of ligands to targets.
FRED - Fast exhaustive docking	A program that uses an exhaustive search algorithm to dock molecules from a multiconformer database into a receptor site
GLIDE	Glide approximates a complete systematic search of the conformational and positional space of the docked ligand and then refines by Monte carlo sampling.

In de novo generation, ligand molecules are built within the binding pocket by assembling small fragments such as benzene rings, carbonyl groups, amino groups, etc., where they are positioned, scored, and linked *in silico*. The final compounds are then taken up for synthesis. This method results in novel molecules not present in any databases.

Ligand-based approaches

In cases where the 3-D structure of a target protein is not available, drugs can be designed using the known ligands of a target protein as the starting point. Some of the frequently used methods are quantitative structure-activity relationships (QSAR) and pharmacophore models. Pharmacophore modeling is used to identify common structural features of ligands such as hydrogen-bond donors, hydrogen bond acceptors, aromatic rings, hydrophobic regions, etc. which can then be used to screen for molecules with these features.

QSAR relates chemical structure to biological activity using mathematical models. A QSAR model can then be used to predict the biological activity of molecules based on their structure information. A QSAR model is also used to effectively screen potentially active molecules from a database.

Lead Optimization

Another important part of drug design is lead optimization whereby researchers design, synthesize, and retest analogues of primary lead compounds. Chemical modifications of the lead compound are carried out to make them more effective and safer. Lead optimization involves generation of analogues of the initial leads to find compounds that have acceptable pharmaceutical properties and can enter the next phases of drug development and clinical testing.

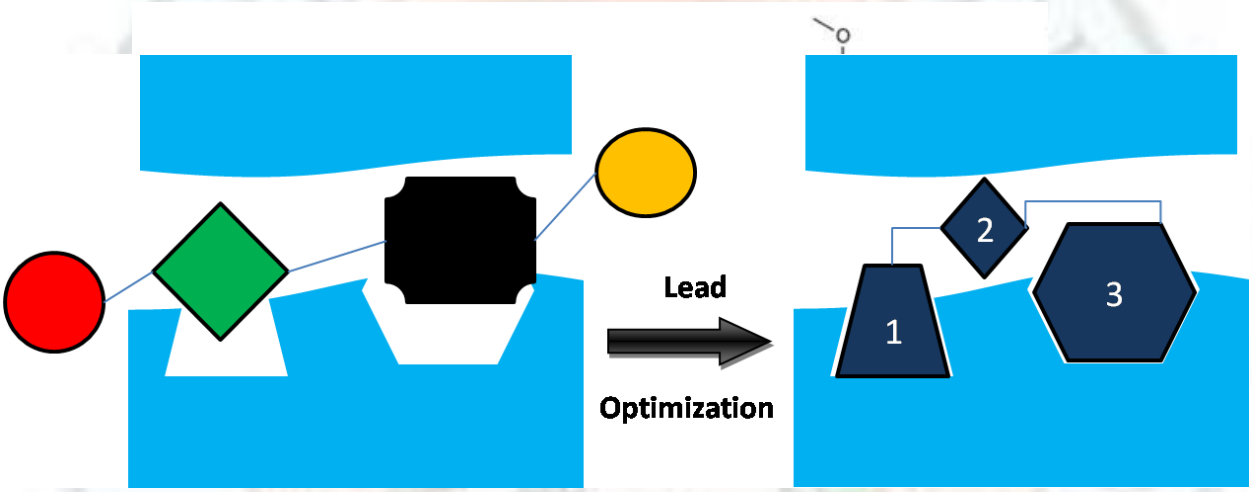

Figure: Initial hit derived from virtual screening approach which, after lead optimization, shows better binding.

Source: Author

Lead optimization needs involvement of both computational approaches and medicinal chemistry. The process requires exploration of the entire chemical space to check for structural features important for *in vitro* pharmacological activity and further improvement of physicochemical properties and metabolic profile. Chemically, this can be achieved using combinatorial chemistry approach, bioisosterism, rapid parallel synthesis and many other approaches. Computationally, there are two basic strategies for lead optimization, namely structure-based (high-throughput molecular docking, fragment-based screening) and ligand-based approaches (QSAR, Combinatorial chemistry, Pharmacophore modeling). Computationally, some of the softwares have tools for lead optimization, for example, Lead-optimization docking and CombiGlide in Schrödinger suite and Muse from Tripos.

After lead evaluation, compounds are taken up for synthesis and further biological evaluation which includes animal testing, human clinical trials and FDA approval, etc after which the drug enters the market or goes back to the drug design process.

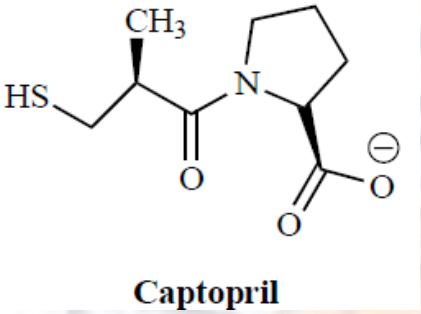

Figure :Discovery of Captopril as antihypertensive agent - **An example of one of the early successful applications of computer-aided drug design**

Table: Several drug design softwares are now available in public domain. Table lists some popular softwares for drug designing.
Source: Author

Name of the Software	Description	URL
Insight II, Discovery Studio	Molecular modeling and drug design program developed by Accelrys Inc.	www.accelrys.com
SYBYL	Molecular modeling program by Tripos	www.tripos.com
OSDD	Translational platform for drug discovery	www.osdd.net
Sanjeevini	A complete drug design software from SCFBio, IIT Delhi	www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp
Molecular Operating Environment (MOE)	A drug discovery software package by Chemical	www.chemcomp.com

	Computing group	
Schrödinger	Schrodinger provides wide range of modules like Glide, Prime, CombiGlide, etc. for drug discovery	www.schrodinger.com

Bioinformatics and Metabolomics

Metabolome represents the entire set of small-molecule metabolites (metabolic intermediates, hormones, other signalling molecules and secondary metabolites) in a cell that are formed from the cellular process in organisms and study of metabolome comprises of Metabolomics . Metabolomics is a direct measure of cellular physiology. HMDB is public domain database comprising of metabolites, drugs and food components that can be found in the human body (www.hmdb.ca).

Systems Biology: Application & Future Prospects

"Ask five different astrophysicists to define a black hole, the saying goes, and you'll get five different answers. But ask five biomedical researchers to define systems biology, and you'll get 10 different answers . . . or maybe more." (Courtesy: NIH)

Systems biology, an interdisciplinary scientific field, takes account of both biological and computational information to explain any living system. In contrast to traditional fields of life sciences, it has been quite challenging to summarize systems biology in a single line. Noticeably, the functionality of a system (organ or tissue or even a cell) at the biological level is the result of the combined effort of multiple effectors (genes or proteins) instead of a single effector where all these effectors work in unison to produce a desirable outcome. For instance, most of the human diseases are a consequence of dysfunctioning of several genes (polygenic). Hence, it becomes indispensable to study and understand the crosstalk between multiple genes and pathways so as to precisely pinpoint the probable causal agent. Systems biology can be used as an efficient tool to understand the

physiology of the human body at the molecular level. Hence, a deeper insight will aid in preventing many diseases.

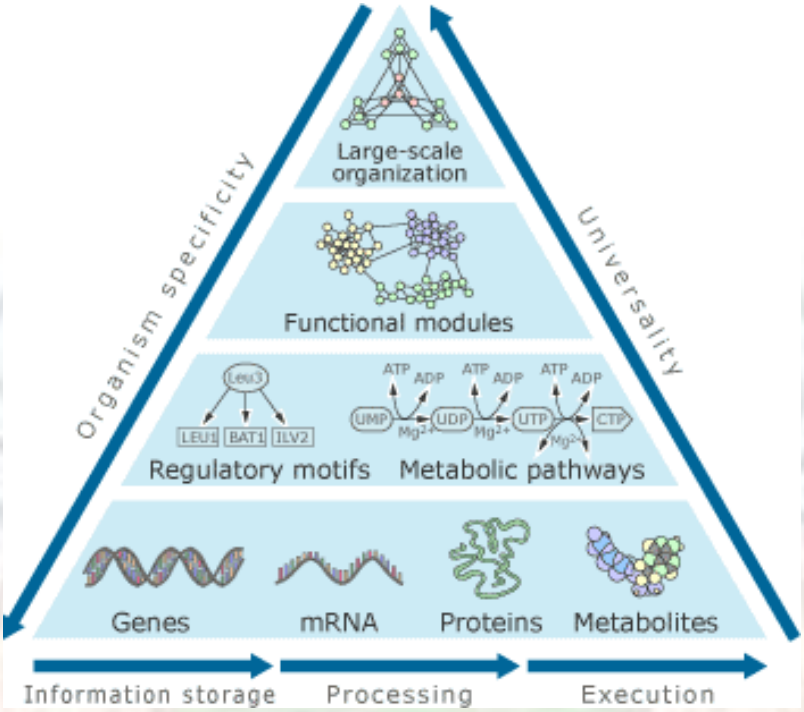

Figure: "Systems biology...is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different....It means changing our philosophy, in the full sense of the term"~Denis Noble

Source: Z. N. Oltvai and A. L. Barabási- Life's Complexity Pyramid, Science (2002).

From OMICS to systems biology

The basic building blocks and biological processes of the body's physiology are highly systematized and complex. To unveil this cellular and molecular intricacy we need to employ a systemic approach. Progress in systems biology has been driven by advances in domains such as mRNA expression, proteomics, and sequencing. These high throughput technologies have generated huge amount of genetic data that renders the modern science to a new idiom called 'omics' (genomics, proteomics or metabolomics). A systemic understanding is greatly needed to retrieve valuable information from such large datasets to interpret it in relative context. Hence, presently both academic and profit organizations are using systems

approach to identify mechanisms implicated in disease, select novel drug targets and discover associated biomarkers.

Complexity of complex disorders

Most of the common human diseases are polygenic in nature which involves crosstalk of multiple proteins in different pathways, also known as complex or multi factorial disorders. Interplay between various genetic factors along with several environmental factors (epigenetic) increases the complexity to a much higher extent. Therefore to rephrase such disease mechanisms we need to understand the overall 'biological network' (i.e gene regulatory network, cell signalling network, physical or functional interactome of proteins etc.). Network based approaches takes account of cellular functions as a broad spectrum of molecular interactions of various genes or proteins to determine prospective therapeutic drug targets.

Systems Pharmacology

Multiple proteins in our body often share common binding sites forvarious drugs. As a result of lack of target specificity "one drug-one target" design approach falls apart. An alternative to such a limitation can be overcome via precise insight into the initial drug design method and its adverse physiological effects. A consideration of the physical and chemical properties of a designed drug accompanied by identification of its targets in different biological pathways can be cost effective and time saving. Besides, a foremost objective of 'systems pharmacology' tends to explore the poly-pharmacological aspects of drug design for complex human disorders.

Next Generation Sequencing

The advancement of the field of molecular biology has been principally due to the capability to sequence DNA. Recent introduction of massively parallel sequencing platforms have transformed the field by reducing the sequencing cost by more than two folds. Previously, Sanger sequencing ('first-generation' sequencing technology) has been the sole conventional technique used to sequence genomes of several organisms. In contrast, NGS platforms rely on high-throughput massively parallel sequencing involving unison sequencing of millions of DNA fragments from a single sample. The former facilitates the sequencing of an entire genome in less than a day. The speed, accessibility and the cost of

newer sequencing technologies have accelerated the present –day biomedical research. These technologies reveal large scale applications outspreading even genomic sequencing. The most regularly used NGS platforms in research and diagnostic labs today have been- the Life Technologies Ion Torrent Personal Genome Machine (PGM), the IlluminaMiSeq, and the Roche 454 Genome Sequencer. NGS platforms rapidly generate high volume of data on the gigabase scale. So the NGS data analysis poses the major challenge as it can be time-consuming and require advanced skill to extract the maximum accurate information from sequence data. A massive computational effort is needed along with in-depth biological knowledge to interpret enormous NGS data.

Table :Next-Generation Sequencing Platforms

Source: *Adapted from MolEcolResour 2011;11:759–69; Nat Rev Genet 2010;11:31–46; Am J Clin Path 2011;136:527–39.*

Next-generation sequencing technologies employ different techniques, but all have in common, the ability to sequence more DNA base pairs per sequencing run than earlier methods like Sanger sequencing.				
Manufacturer	**Technique**	**Run Time Per Read**	**Base Length**	**Cost (in 000s)**
Helicos	ReversibleTerminator	8 days	32	$999
Illumina	ReversibleTerminator	4–9 days	75–100	$500–900
Ion Torrent	Real-time	<1 day	964	$50
Roche/454	Pyrosequencing	<1 day	330	$500–700
SOLiD	Sequencing By Ligation	7–14 days	50	$600–700

Summary

1. Bioinformatics is application of IT to address biological problems.

2. Bioinformatics and its related fields like Genomics, Proteomics, Transcriptomics, Metabolomics and Systems biology finds useful applications in agriculture, health sector and environmental issues.
3. The three major thrust areas of research include genome and transcriptome and proteome analysis, protein structure prediction and computer aided drug design.
4. Many softwares/tools are being developed and are available freely over the internet to locate genes in a genome and predict structures of protein.
5. Bioinformatics and computational biology help in reducing the cost and time for designing new drugs and are nowadays routinely now used in pharmaceutical companies.

Exercises

1. Discuss the major research areas in the field of bioinformatics
2. What is the difference between the gene organization in prokaryotes and eukaryotes?
3. Differentiate between comparative genomics and functional genomics
4. What is pharmacogenomics?
5. What are the levels of hierarchy in protein structure?
6. Name two experimental methods of protein structure determination.
7. What are the three methods of protein structure prediction by computational approaches?
8. What are the two approaches used in computer aided drug design to design new inhibitor/lead molecules?
9. Define pharmacophore.
10. Name some commonly used docking softwares.
11. What is systems biology? Why do you need to study systems biology?

Glossary

Bioinformatics	Managing and analyzing biological data using computing methods
CADD	Computer Aided Drug Design(CADD) refers to all computational approaches to discover, design and optimize drugs.
Databases	Collection of records of similar content that can be easily assessed and managed
Docking	Docking is the process of determination of binding between two molecules using some force fields. Molecular docking suggests favorable poses of a molecule with respect to the other after binding. The poses are ranked and scored by a mathematical function called scoring function
Functional Genomics	Functional genomics deals with the study of genes and its functions.
Genome	Genome is the total DNA present in a cell.
Gene	Gene is the region of DNA that codes for m-RNA that is translated to protein
Genomics	Genomics is the study of DNA structure, function and expression.
HGP	The Human Genome Project (HGP) is global effort with a primary goal of determining the human genome sequence and identifying genes and the non-genes in the DNA sequence.
Lead compound	Lead compound is often used in drug discovery to describe a chemical compound that has pharmacological or biological activity.
Lead Optimization	Lead optimization approaches aims at enhancing the effectiveness and safety of the most promising compounds .
Metabolomics	Metabolome is the complete set of small molecule metabolites present in the system. Hence, metabolomics is the study of dynamics of metabolites in the body.
Metagenomics	Metagenomics is the study of microbial genomes in a community. It includes the application of genomics to environmental processes. .
Pharmacophore	A pharmacophore is 3D spatial arrangement of functional groups or chemical features describing the conformation of pharmacologically active components of a drug.
Systems biology	Systems biology is an approach in which a set of biological processes or components are studied as an integrated system, where processes or

components interact with one another. Systems biology gives an integrated view of the interacting components in the system and thus facilitates understanding the effects of a biological process on systems level.

Virtual Screening

Virtual screening (VS) is the computational screening of a compound library with respect to a target. The virtual library is subjected to a number of screening methods to select the compounds of desirable properties and affinity towards the target.

References

1. Rastogi SC, Mendiratta N, Rastogi P (2011) Bioinformatics: Concepts, Skills & Applications. CBS Publishers & Distributors Pvt. Ltd. ISBN: 81-239-1482-2.
2. Mount DM (2004) Bioinformatics: Sequence and Genome Analysis 2. Cold Spring Harbor Laboratory Press. ISBN: 0-87969-712-1.
3. Ghosh Z, Mallick B (2012) Bioinformatics: Principles and Applications. Oxford University Press. ISBN-13: 978-0-19-569230-3.
4. Campbell AM, Heyer LJ (2006) Discovering Genomics, Proteomics, and Bioinformatics. CSHL Press. ISBN: 0-8053-8219-4.
5. Young DC (2009) Computational Drug Design. John Wiley & Sons, Inc. ISBN: 978-0-470-12685-1.
6. Tramontano A (2006) Protein structure Prediction: Concepts and Applications. WILEY-VCH. ISBN-13: 978-3-527-31167-5.
7. Hiroaki Kitano (2001) Foundation of Systems Biology. MIT Press. ISBN: 0-262-11266-3.