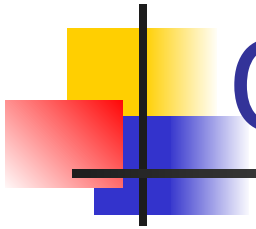
Advances in Computational Biology

Sanghamitra Bandyopadhyay
Professor, Machine Intelligence Unit
Director, Indian Statistical Institute, Kolkata
E-mail: sanghami@isical.ac.in
URL: http://www.isical.ac.in/~sanghami

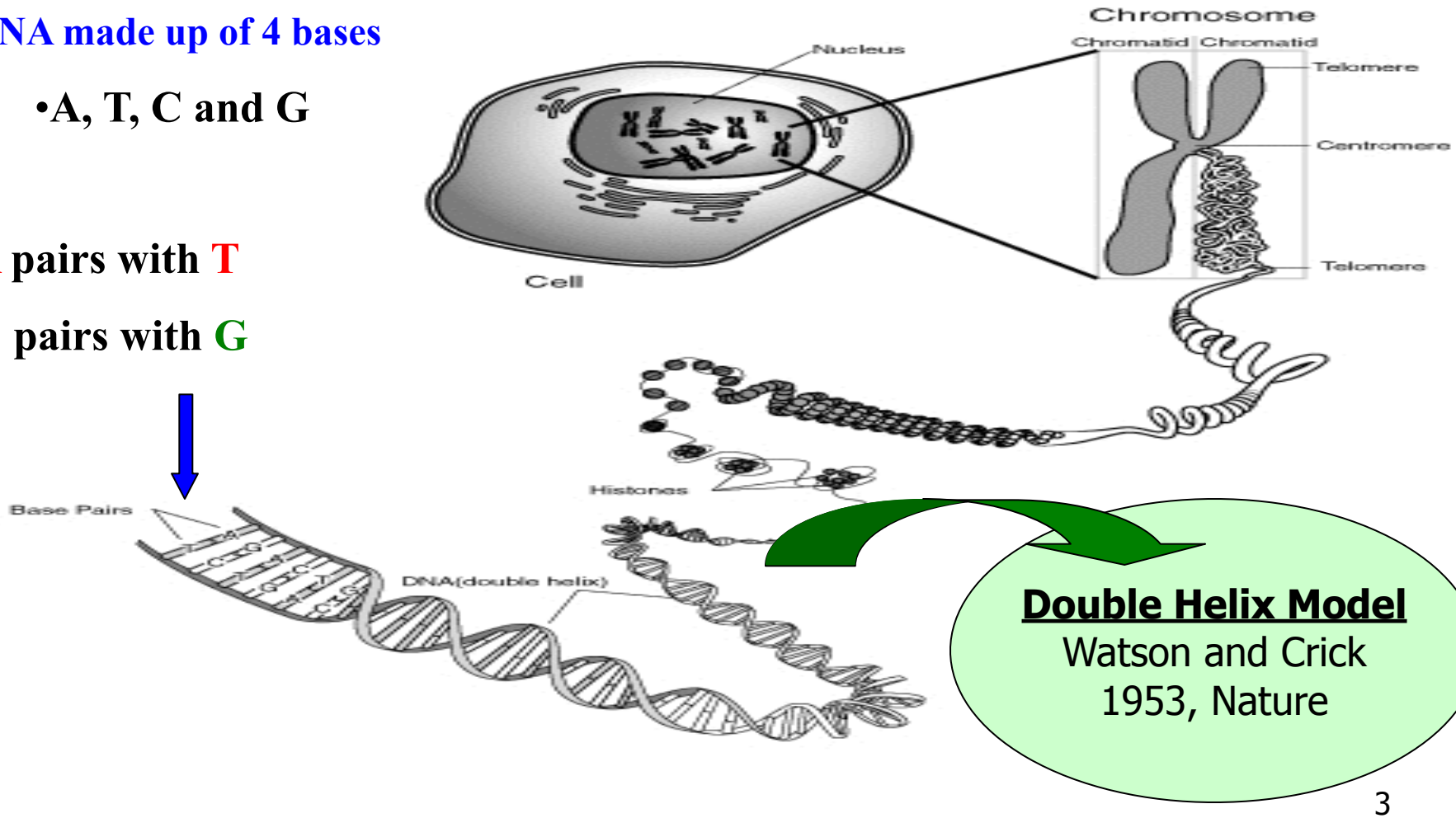


Outline

- Introduction
  - Cell
- Central dogma of molecular biology
  - Transcription
  - Translation
- Computational Biology
  - Introduction
  - Grand Challenges
  - Drug discovery
  - Bio-molecules involved in Cancer
- Future Missions
- Conclusions

Cell

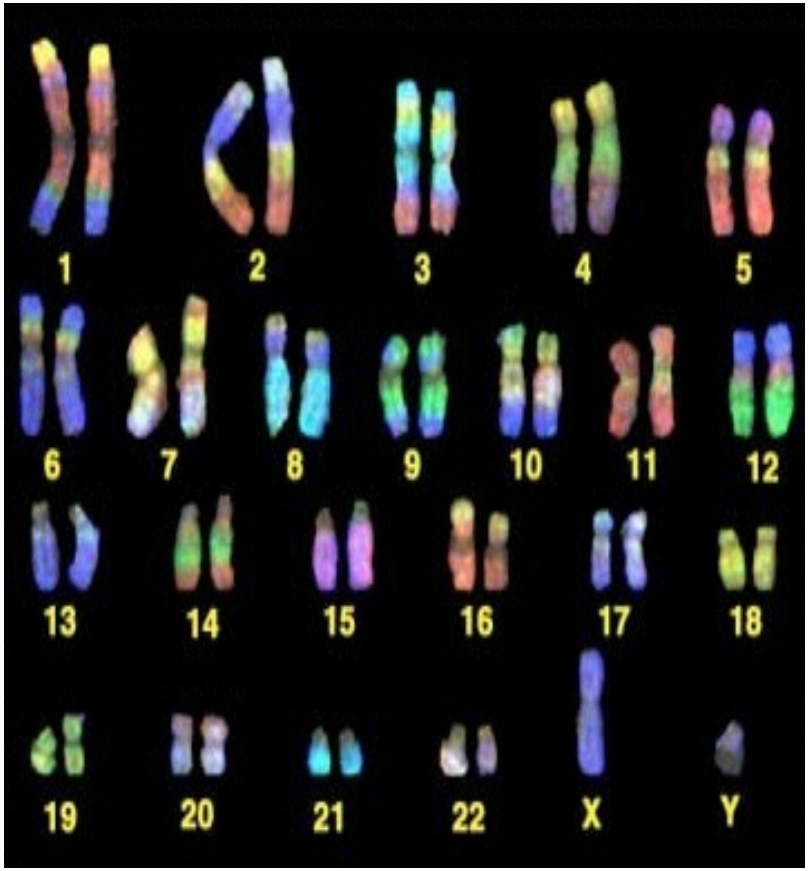
- **DNA made up of 4 bases**
  - **A, T, C and G**
- **A pairs with T**
- **C pairs with G**



Double Helix Model
Watson and Crick
1953, Nature

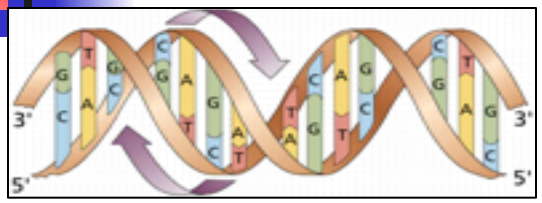
Chromosomes

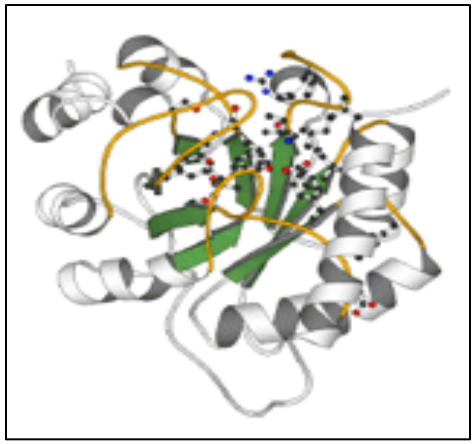
- Humans cells have
  - 22 pairs of chromosomes (autosomes)
  - and one pair of sex chromosomes
- 3 billion base pairs
  - Only about 2.5% occupied by protein coding genes



Ack: faculty.ksu.edu.sa

Central Dogma of Molecular Biology





DNA

transcription

mRNA

translation

Protein

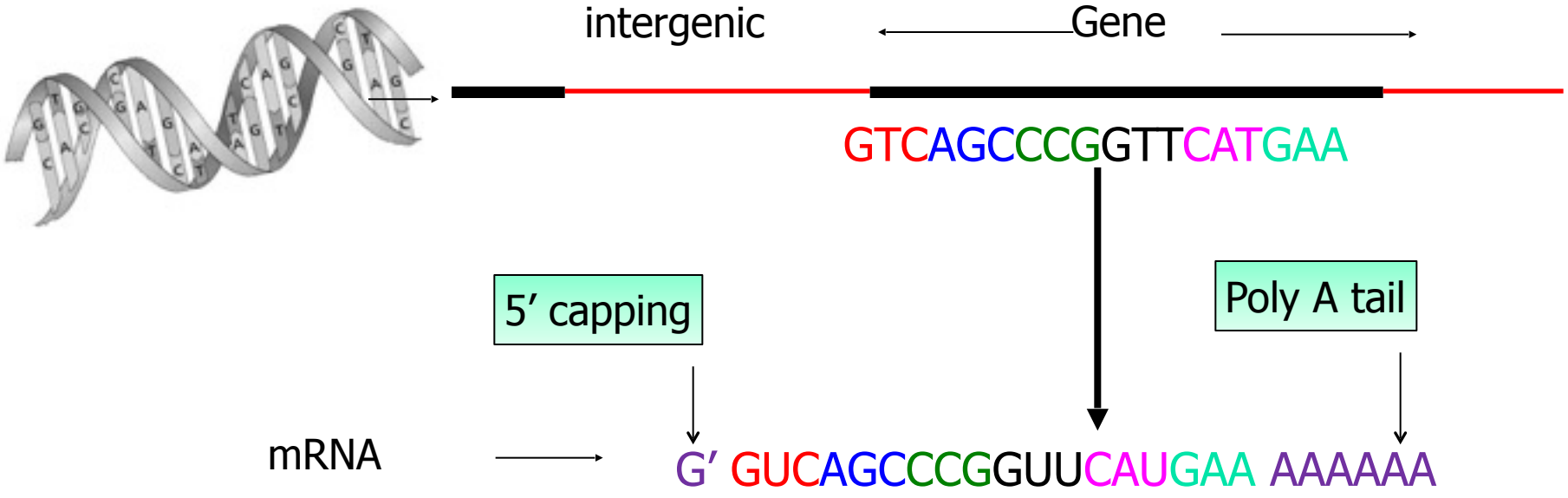
CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

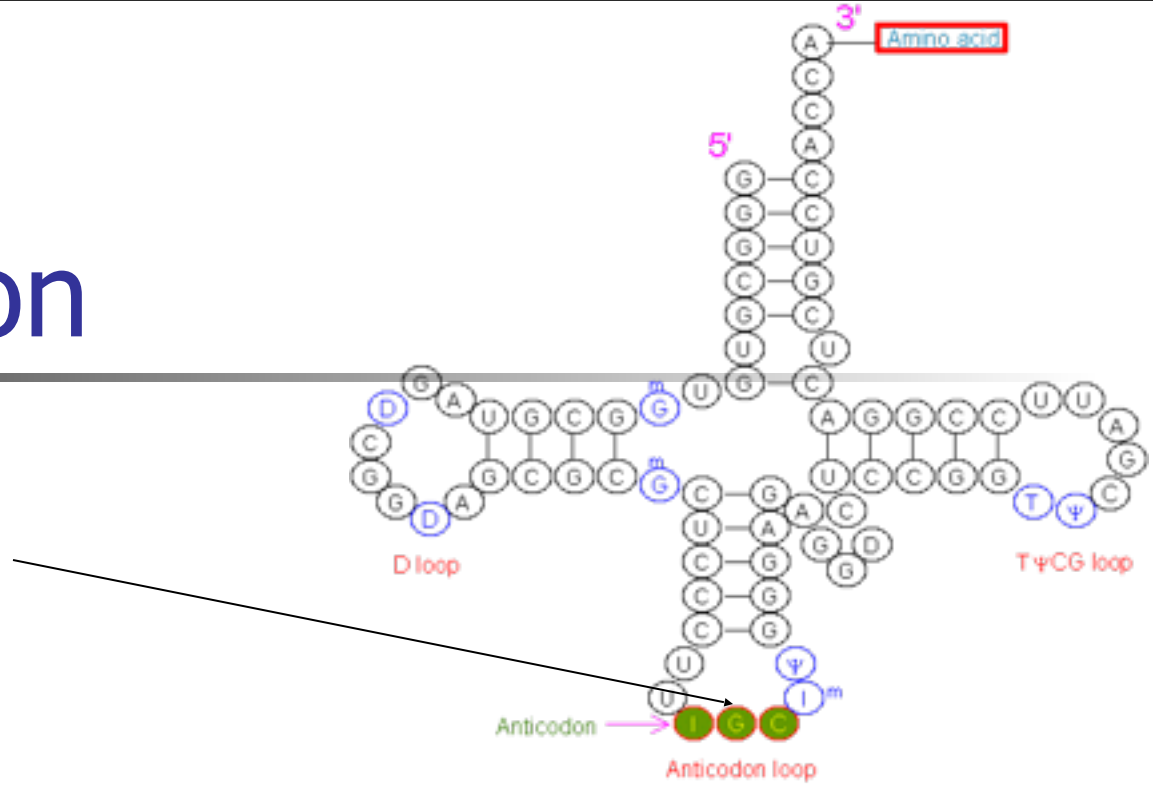
Transcription

- Process by which DNA forms RNA



Translation

tRNA carries the anticodon and the Corresponding amino acid



mRNA → GUCAGCCCGGUUCAUGAA

codon

amino acid → 20 standard amino acids

Protein → V S P V H E

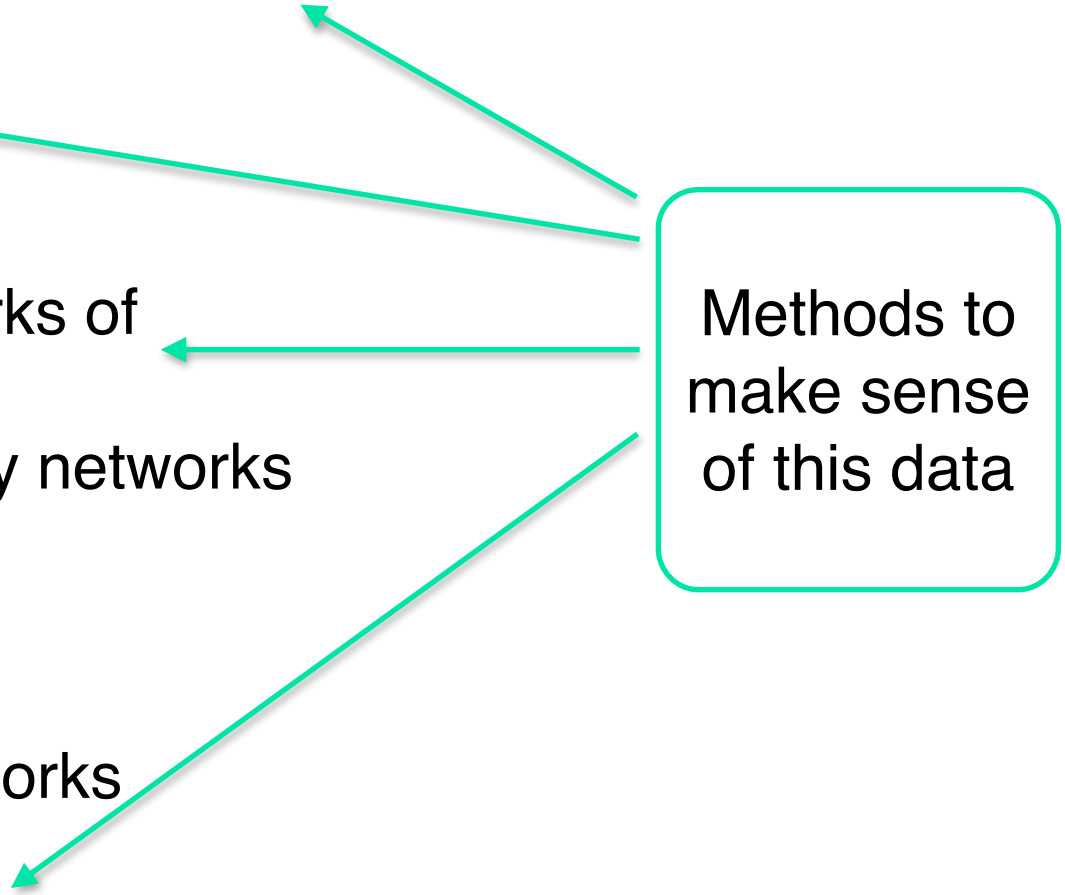
The Genetic Code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	STOP	STOP	**A**
	Leu	Ser	STOP	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

Most common start codon
AUG → Methionine
Alt start codons
GUG, UUG
(prokaryotes)

Levels of Data Generated

- Entire genomes, transcriptome and proteome
- Expression data
- Structural data
  - protein
  - RNA
- Interaction networks of
  - DNA-protein
  - gene regulatory networks
  - protein-protein
  - drug-disease
  - drug-drug
  - metabolic networks
- Epigenetic data



Computational Biology

- What is it?
  - Analysis, prediction, and modeling of biological data with the help of computers
  - Use of systems mathematics and computer simulations in Biology
  - Cells being treated as a system and studied at a systems level
- Why in Biology?
  - Huge, raw, unstructured data
  - Eliminates lot of expensive biological experiments
- Challenges
  - Biological system is extremely complex
  - Continually changing through evolution
  - Information to build a complete model of the cell is still not available
- Time will be the best judge
  - Already success stories are there

Ack: Palsson, 2000, Nature Biotech

The Challenges - Human Genome Project (HGP)

- Completed in 2003
  - 13-year project coordinated by the U.S. DoE and NIH
- Main project goals
  - identify all the approximately 20,000-25,000 genes in human DNA
  - determine the sequence of 3 billion base pairs making up human DNA
- So far about 3762 bacterial, 183 eukaryotic, and 181 archaea genomes sequenced
  - Human, rat chimpanzee, chicken, and many others
  - 56 plant genomes
- The Challenge
  - Cost of sequencing one genome about USD 250,000
    - Sanger sequencing to Next gen sequencing
    - Reduce it to USD 1000 – will become routine medical practice
  - Learn the language of DNA - its syntax and semantics, unravel its mystery

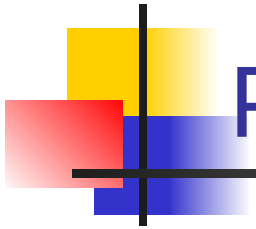
The Challenges – Assembling the genome

- DNA sequencing technology can sequence short strands
  - 1000 bps
- Genome
  - Millions of base pairs
- Shotgun sequencing → **ACTGTCAGCCTGATCGTAATCGCTAGCTGA**

Frag 1	A	C	T	A	T	G	C	G	C	T	A	A	T	C	C	T	T	-	-	-	-	-	-	-	-
Frag 2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	A	T	G	C	A	T	G

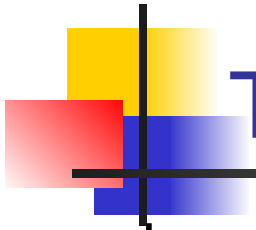
Frag 3	A	C	T	A	T	G	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Frag 4	-	-	-	-	-	-	-	G	C	T	A	A	T	C	C	T	T	G	A	T	G	C	A	T	G

- Align the four fragments to re-assemble the genome
- Next generation sequencing – more genomes will soon become available



Present Day Scenario

- Next Generation Sequencing Data
  - Low cost, less time than Sanger sequencing
  - Massively parallel
  - Generates millions of short reads or fragments in a single run
  - Sequences at large depth - many reads per character
  - Illumina, Roche 454, ABI SOLiD etc.
  - Huge amount of data
    - 600 Gb per run (Illumina HiSeq2000)
  - Use of Efficient Computational Methods
  - Storage, Indexing, Retrieval, Analysis
- RNA-seq data
- Single cell transcriptomic data



The Challenges – Sequence to Structure

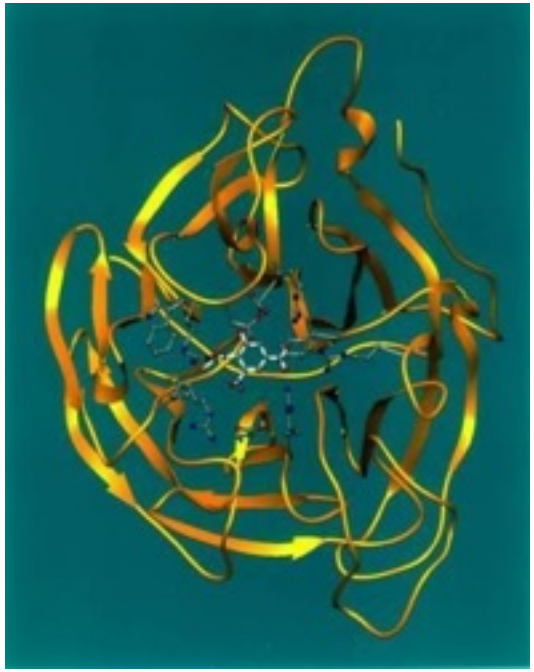
- Large number of sequences of bio-molecules are known
  - DNA, RNA, Protein
  - About 300,000 protein structures are known.
- What about their structure in 3-d?
  - Number of protein structures in PDB up to 2010 – 64640
  - Structure determines the function of a protein
- Predict structures from sequences
  - Important in drug design, design of novel enzymes
  - Machine learning methods
  - Search approach
    - Computation of protein free energy of each structure
    - Searching the prohibitively large space of possible protein structures
    - Finding the structure with global minimum energy
  - Homology with known proteins

The Challenges - Drug Discovery

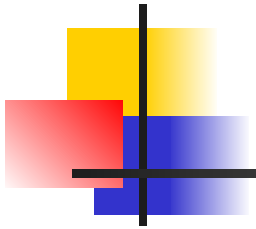
- Chance observation in nature
- Expensive high-throughput screening (HTS)
  - physically test thousands of diverse compounds a day
  - often with an expected hit rate on the order of 1% or less
  - still fewer expected to be real leads following further testing
- For a drug to come into the market
  - Typically it takes 10-15 years
  - Approximate cost more than a billion USD
- Use protein docking algorithms
- Rational drug design
  - design using the information about the 3D shape of target proteins

Rational Drug Design: Practical outcome

- Relenza: Influenza
  - discovered in 1989, a neuraminidase inhibitor designed to treat Influenza virus A
- Dorzolamide: anti-glaucoma agent used in eye drops
  - drug approved in 1995 designed with structure-based drug design - carbonic anhydrase (converts CO_2 to bicarbonate & protons) inhibitor
- Imatinib: leukemia
  - a tyrosine kinase inhibitor designed specifically for the bcr-abl fusion protein that is characteristic for Philadelphia chromosome-positive leukemias (chronic myelogenous leukemia and occasionally acute lymphocytic leukemia).

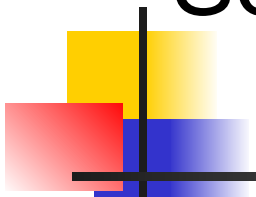


Flu drug **Relenza** – designed by Australian team of scientists headed by Prof Peter Coleman



COMPUTATIONAL METHODS IN BIOLOGY

Sequence Alignment

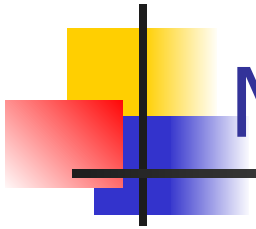


- FOGSAA: Fast optimal global sequence alignment algorithm [Scientific Reports, 2013]
  - Based on the concept of branch and bound and look ahead of estimated scores
  - Almost 70%-90% faster than Needleman Wunsch for genome sequences
  - 25%-40% faster for protein sequences

- ConLSH [2nd ASE International Conference on Big Data Science and Computing, Stanford University, CA, USA, 2014.]:
  - method for large scale sequence alignment
  - using context based locality sensitive hashing
  - based on FOGSAA

- Towards the design of local algorithms applicable to NGS data

microRNA
A Small Biomolecule Implicated in Cancer



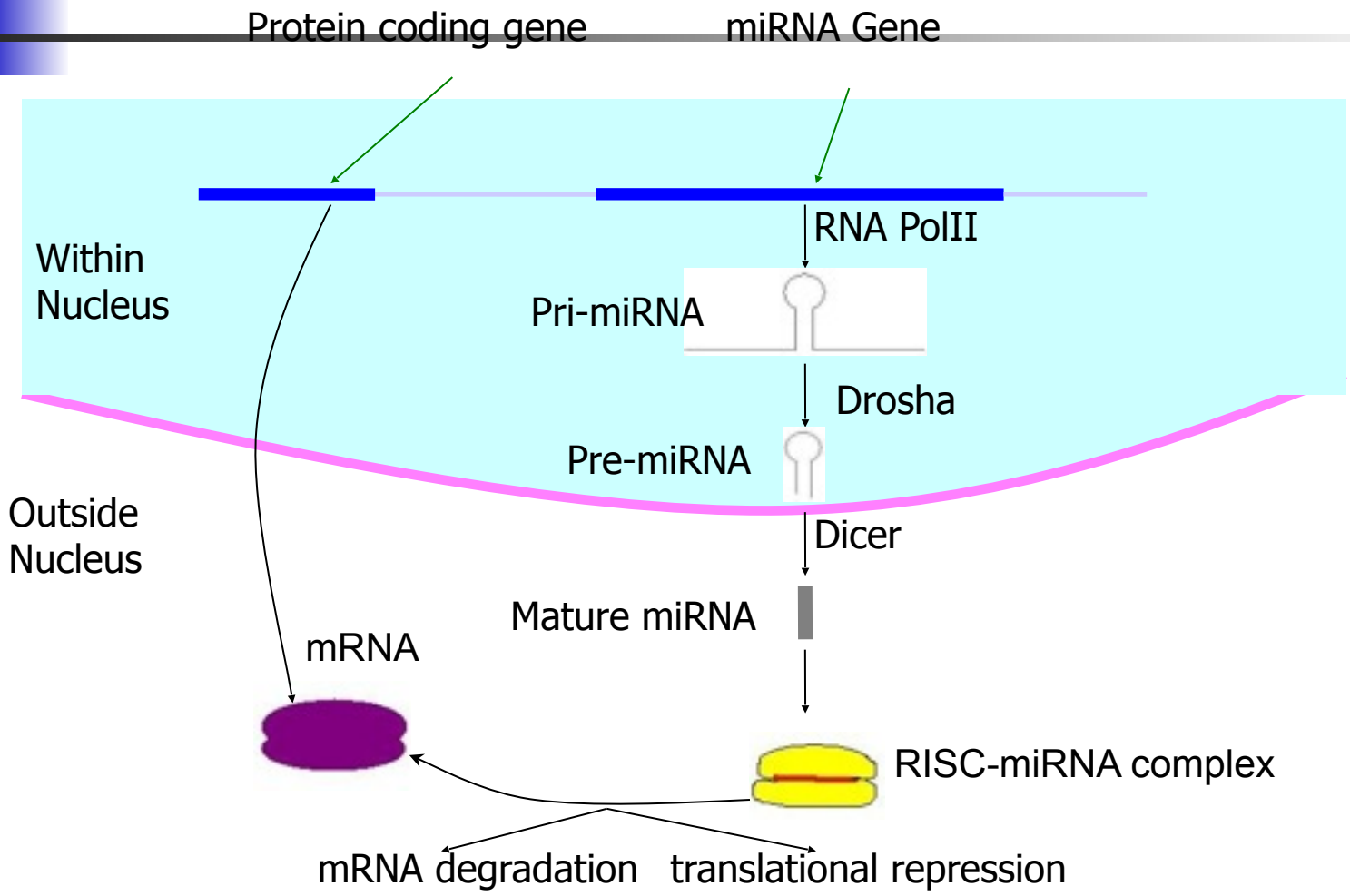
MicroRNA

- Small endogenous RNA molecules
  - about 22 nucleotides long
- Regulates the expression of other genes
  - by mRNA degradation
  - translational repression

Important Roles of miRNA

- Regulate cellular processes
  - Cell differentiation
  - Development
  - Genomic stability
- Metabolic regulation (miR-375) & insulin secretion.
- Post-transcriptional regulation of gene expression.
- Viral infection (HIV/HPV infection).
- Implicated in different types of cancers
  - leukemia, lung, colon, breast, cervical, pancreas, thyroid, prostate…

Biogenesis of miRNA



Deregulation of microRNA Causes Cancer

Some miRNA Statistics

- 1,600 precursor and 2,042 mature human miRNAs in miRBase Release 19 (August 2012): http://www.mirbase.org/cgi-bin/browse.pl?org=hsa
- More than 600 miRNAs are known to be significantly dysregulated in various cancer malignancies
  - Expression of miRNAs are altered in most human malignancies
- More than10,000 biologically validated miRNA-cancer relationships have been identified.
- Many miRNAs have strong oncogenic (ex: miR-21, miR-145, miR-155) or tumor suppressor (ex: let-7a-2, miR-16, miR-143) characteristics

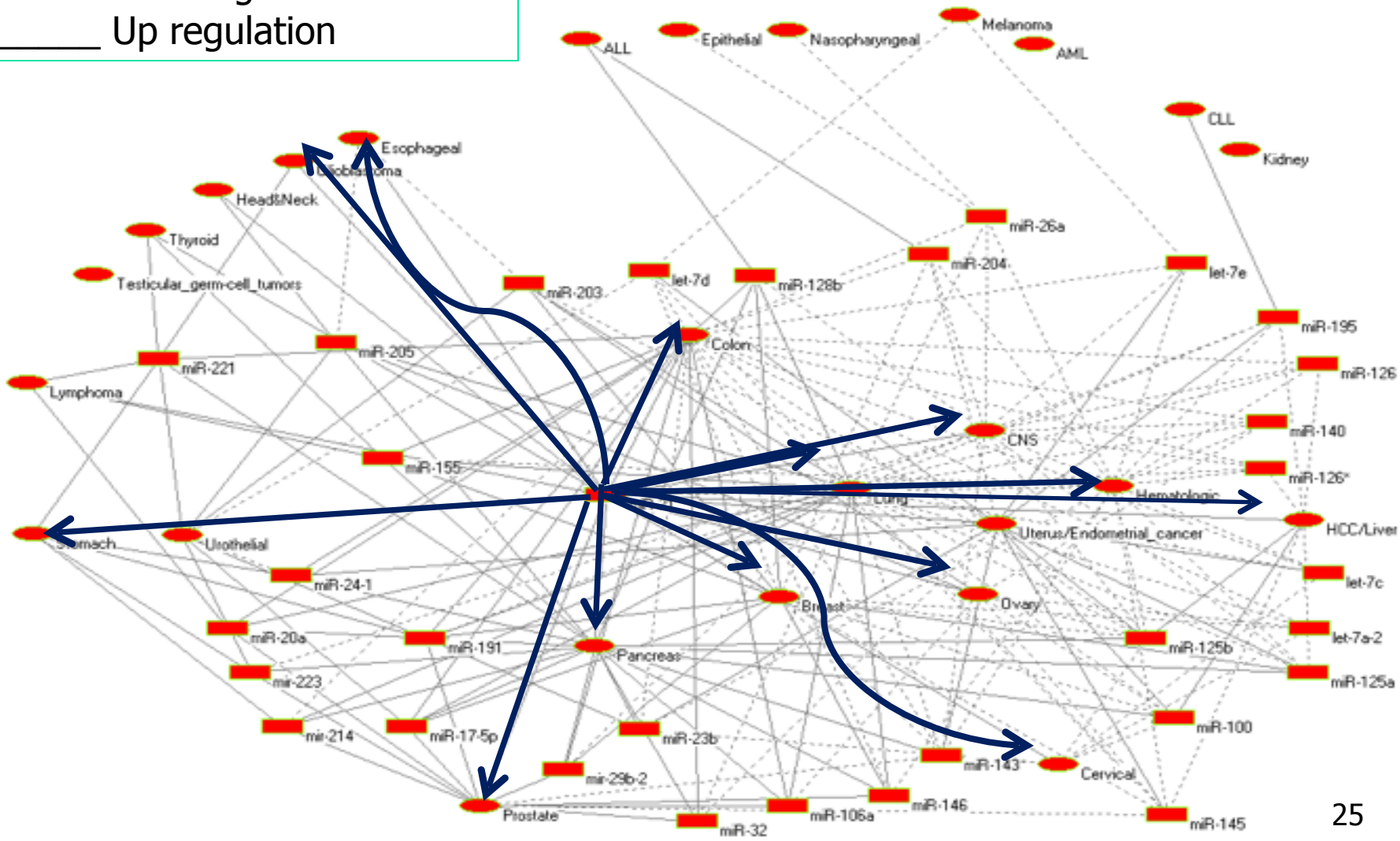
Cancer-miRNA network

Bandyopadhyay and Mitra, *BMC Silence*, 2010

- >200 miRNAs involved in cancer
- 1000 cancer-miRNA relations known

------ Downregulation

_____ Up regulation



In Silico microRNA Target Detection

MicroRNA Target Prediction Algorithms

- Traditional algorithms (mainly sequence based):
  - miRanda, 2003-2008
  - PicTar, 2006
  - TargetScan, 2004-2009
- Structural interaction based algorithms:
  - PITA, 2007
  - STARmiR, 2007

MicroRNA Target Prediction Algorithms cont…

- Machine learning based algorithms:
  - NBmiRTar, 2007
  - MirTarget2, 2008
  - TargetMiner, 2009 ♣

♣ Sanghamitra Bandyopadhyay and Ramkrishna Mitra, Bioinformatics, 2009

www.isical.ac.in/~bioinfo_miu

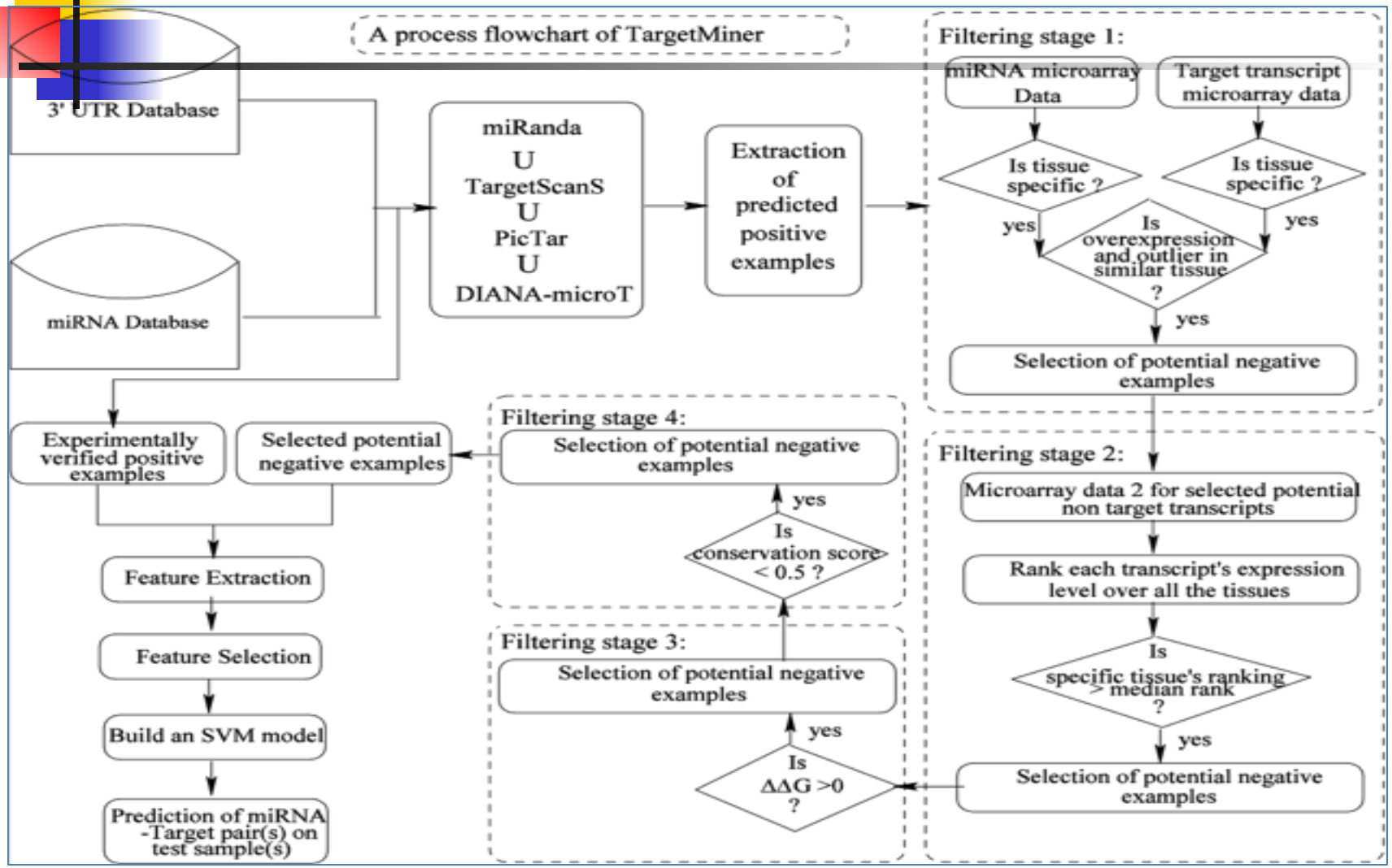
Target Prediction Problem : Machine Learning Approach

Data Set			Class
miR1	⟷	mRNA1	Target
miR1	⟷	mRNA2	Non-Target
miR2	⟷	mRNA3	Target
miR3	⟷	mRNA2	Non-Target
miR2	⟷	mRNA4	Non-Target
miR1	⟷	**mRNA4**	**???**

Classification Problem of two Classes

- Feature extraction
- Feature selection
- Classification – SVM, decision tree, random forest classifier

TargetMiner



*

Performance comparison

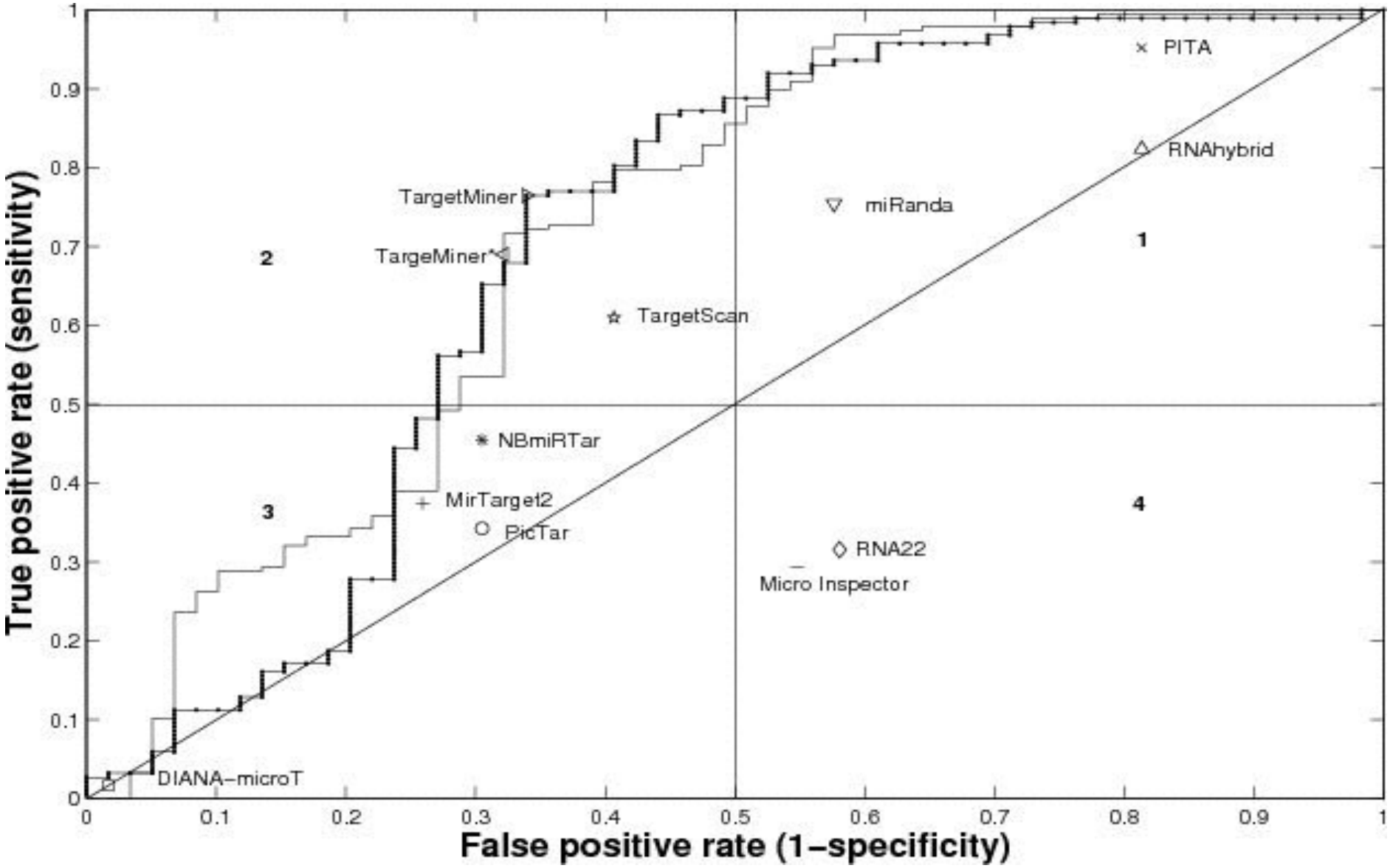
	miRanda	NBmiRTar	PicTar	TargetScan	TargetMiner	TargetMiner*
MCC	0.167	0.129	0.034	0.174	0.384	0.321
ACA	0.589	0.574	0.518	0.601	0.713	0.684

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \qquad Specificity = \frac{TN}{TN+FP} \times 100\% \qquad MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{[TP+FP][TP+FN][TN+FP][TN+FN]}}$$

$$\text{ACA} = 1/c \sum_{i=1}^{c} accuracy\ of\ class\ i, where\ c = number\ of\ classes$$

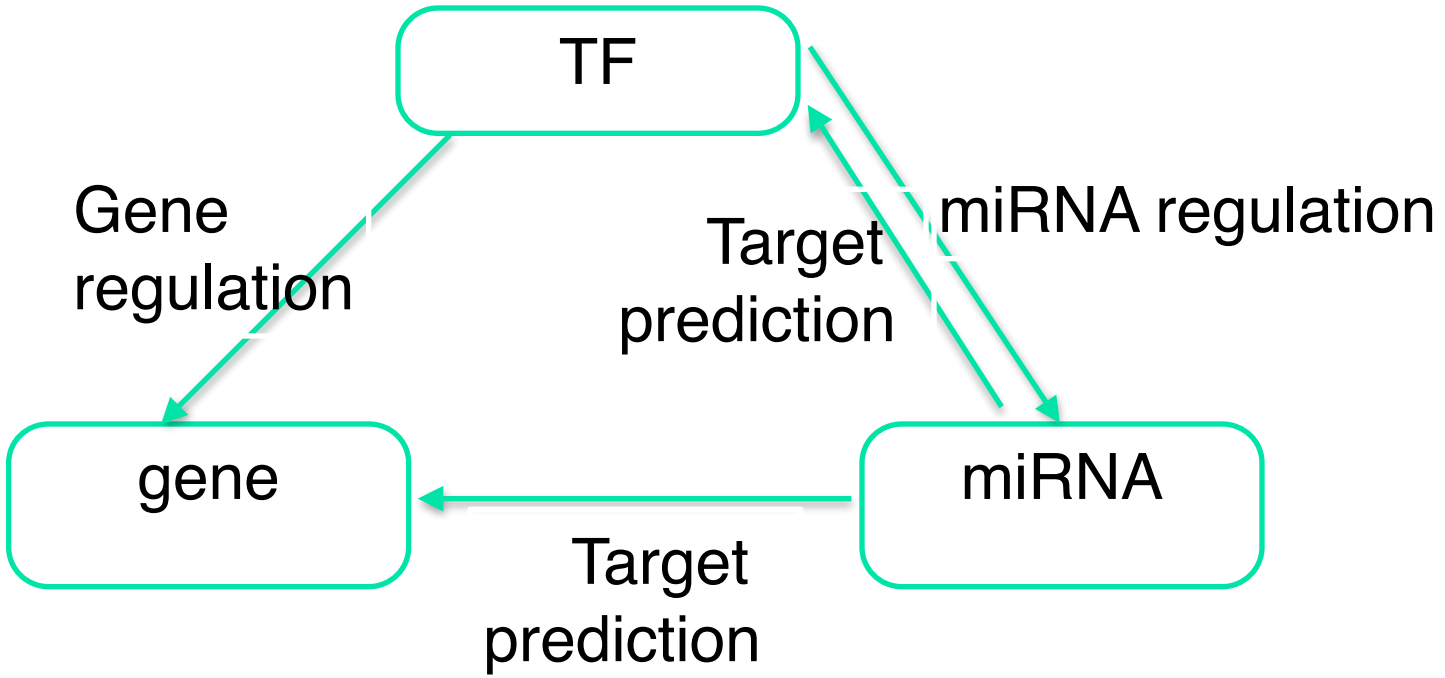
*

Performance comparison



*

MicroRNA Induced Regulatory Network



Molecular BioSystems, 2013

Analyzing the Network Topology: Breast Cancer Specific Network

- miR-210: Expression of miR-210 correlates significantly with tumor hypoxia, independent marker of breast cancer [Camps *et al.*, 2008]
- miR-149: Involved in apoptosis resistive pathway[Hu *et al.*, 2011]
- miR-10B: Oncomir with role in tumor invasion and metastasis [Ma *et al.*, 2007]
- miR-155: Oncomir [Jiang *et al.*, 2010]

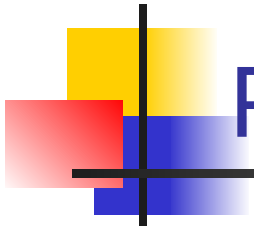
miR-202, miR-155: Is it important?

mir-155 later validated independently [Genes and Cancer]

Analyzing the Network Topology: Colorectal Cancer Network

- miR-378: Oncomir specific to colorectal cancer [Feng *et al.*, 2011]
- miR-34B: Early screening marker of colorectal cancer [Kalimutho *et al.*, 2011]

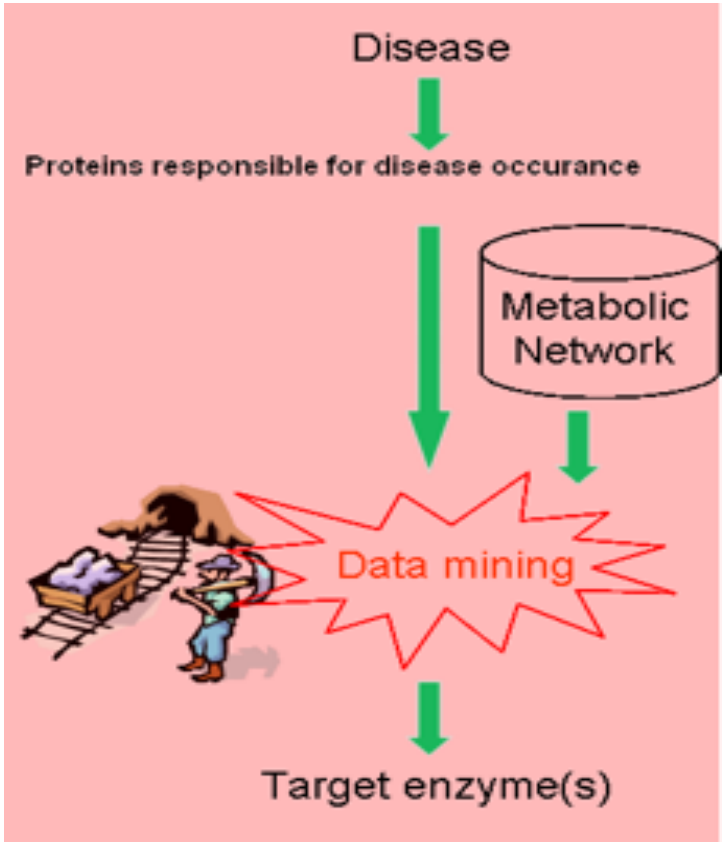
miR-149: Possible biomarker?



Rational Drug Design

- Steps in RDD
  - Step 1: Looking for protein targets
  - Step 2: Identify the active site
  - Step 3: Design molecules for blocking the active site
  - Step 4: Analysis of the properties of the designed molecules

Drug Target Identification

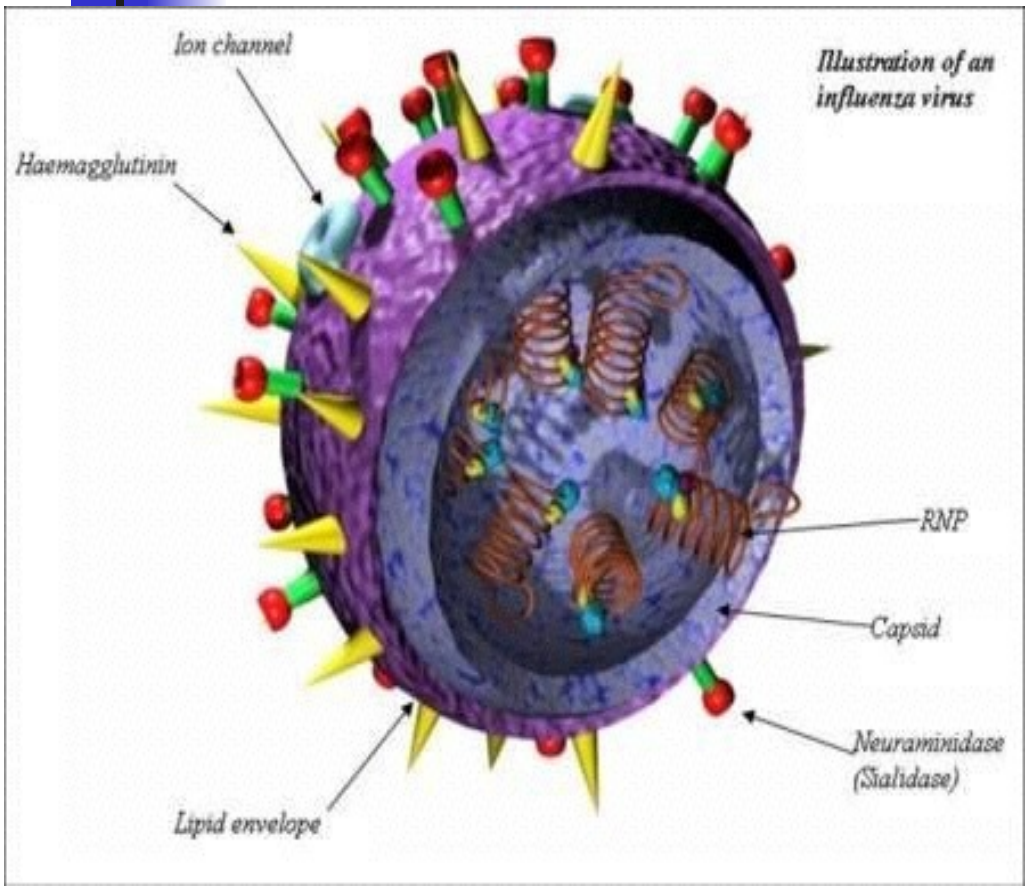


- Genome and proteome analysis
- Metabolic network analysis

Drug Targets for Pathogens

- Proteins responsible for pathogenesis of the causative organism
- Proteins responsible for vital processes of the causative organism
- Proteins altering the normal vital processes of a patient
- At the same time, normal proteins of the host should not be targeted

Identification of Potential Drug Targets



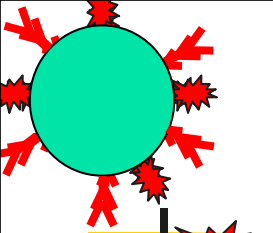
Flu virus

Influenza viruses are named according to the proteins sticking out of their virus coat.

There are two types of protein = N and H.

N and H have special shapes to perform specific jobs for the virus.

Source http://www.gtac.edu.au/site/bioinformatics/bio_task_10/bio_task_10.ppt



Virus

N cuts the links between the viruses and the cell surface so virus particles are free to go and infect more cells.

H attaches to cell surface proteins so virus can enter

Proteins on cell surface

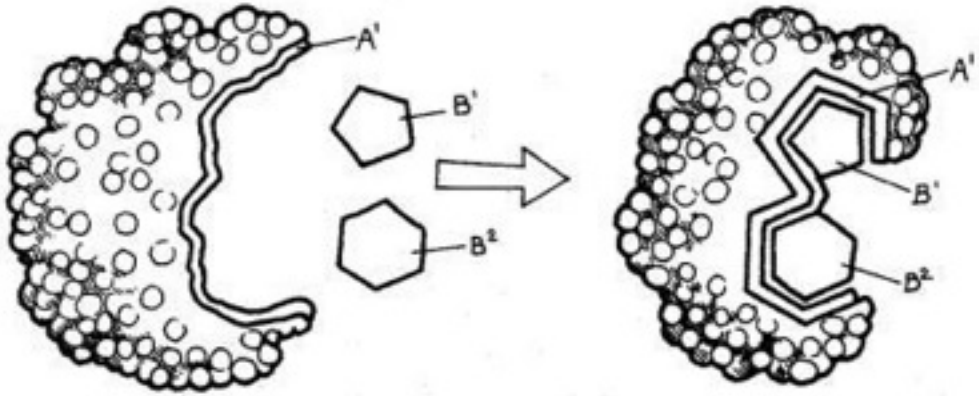
Virus genes are released into the cell.

The lung cell is 'tricked' into using these genes to make new virus particles.

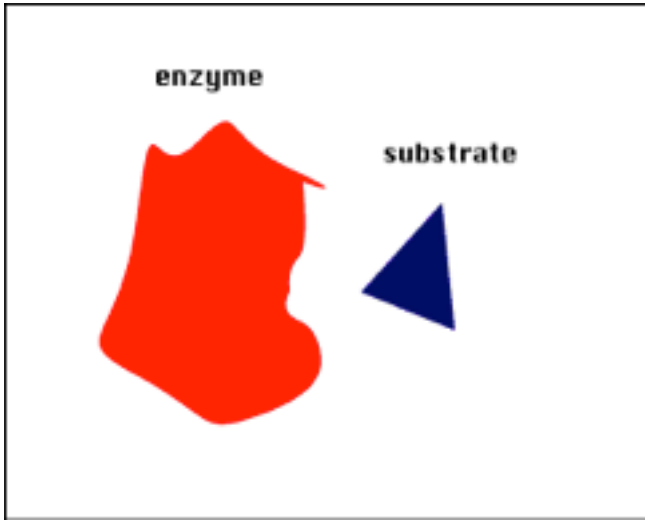
Active site

- Specific sites in proteins where all the action happens.
- Each protein has a specific shape so it will only perform a specific job.
- Example – An enzyme that increases the rate of a reaction

Joining things together

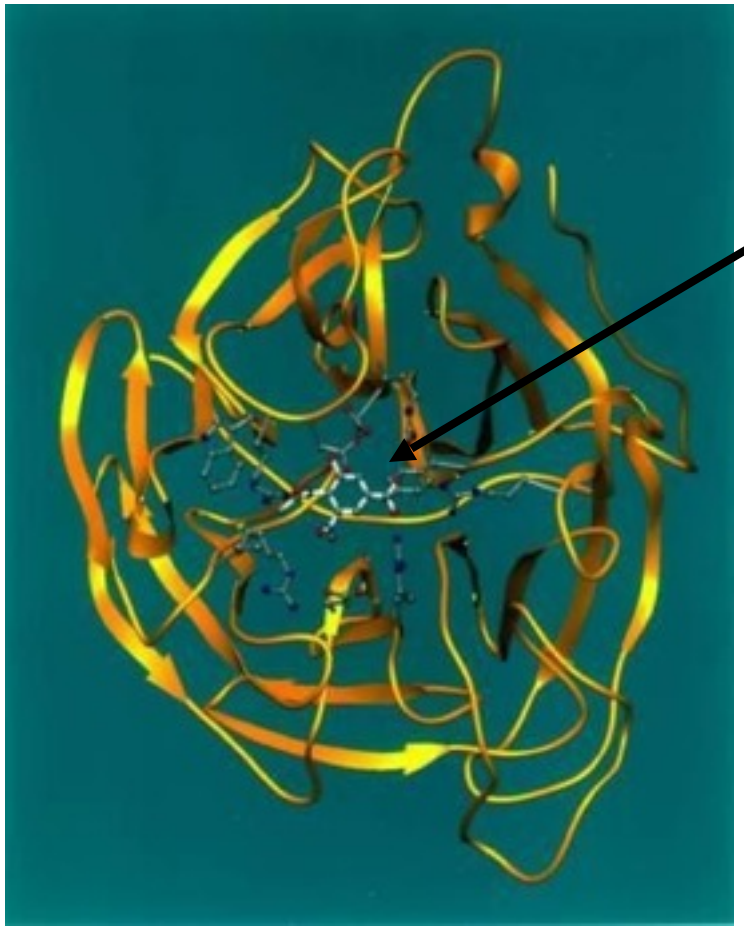


Ripping things apart

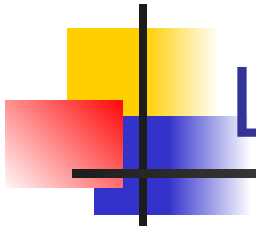


http://chsweb.lr.k12.nj.us/mstanley/outlines/enzymesap/Enzymesap.html
http://academic.brooklyn.cuny.edu/biology/bio4fv/page/active_.html

Drug Relenza blocking the active site of protein Neuraminidase



RELENZA



Ligand Design

- Given the active site geometry of a target protein – design a ligand with complementary geometry and certain electrochemical properties
  - low energy components
  - high oral bioavailability
  - possessing certain pharmacological properties so that it does not participate in non-specific bonding
  - more likely to be synthesizable.
- Genetic algorithm/simulated annealing used for optimization of the energy components and other measures
  - Interaction energy, Internal energy, Electrostatic energy
- Multiobjective optimization strategy used

Scoring function components

- Energy
  - van der Waals energy = $[(C_n / r^6) - (C_m / r^{12})]$
  - Electrostatic energy = $(q_1 q_2)/(4\pi\varepsilon_0 r^2)$
    - $\varepsilon_0 = 8.854185 \times 10^{-12}$ coulomb^2/(N m^2)
  - Bond stretching energy = $[k_l \times (l_{xy} - l_{xy,0})^2]/2$
  - Angle bending energy = $[k_\theta \times (\theta - \theta_0)^2]/2$
  - Torsional energy = $k_\varphi \times (1 - \cos n \times (\varphi - \varphi_0))$
- Tanimoto Coefficient

$$t = \frac{\sum_{i=1}^{881} x_{i,ref-scaffold} x_{i,lig}}{\sum_{i=1}^{881} x_{i,ref-scaffold}^2 + \sum_{i=1}^{881} x_{i,lig}^2 + \sum_{i=1}^{881} x_{i,ref-scaffold} x_{i,lig}}$$

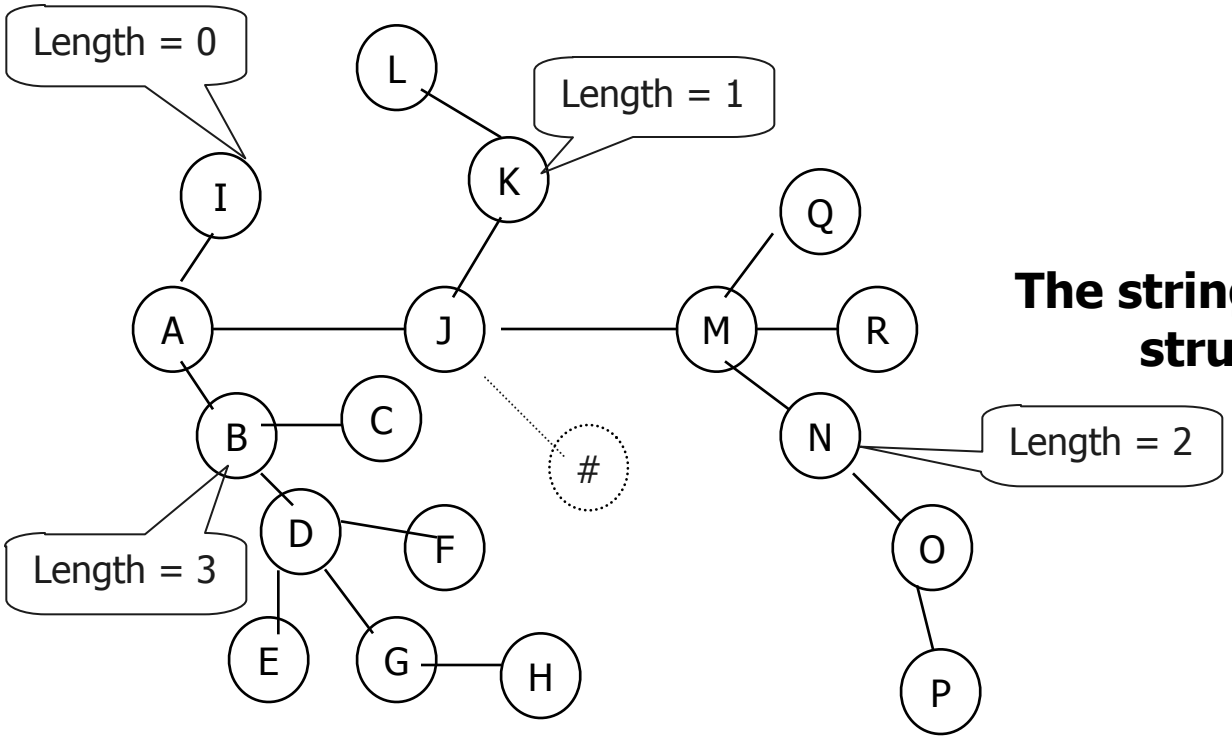
- Oral Bioavailability Measure

ENCODING STRATEGY

B	L	LC	LL	LU	LM	LML	LMU	….	U	UC	UL	UU	UM	UL	….	B	L

Lower tree structure

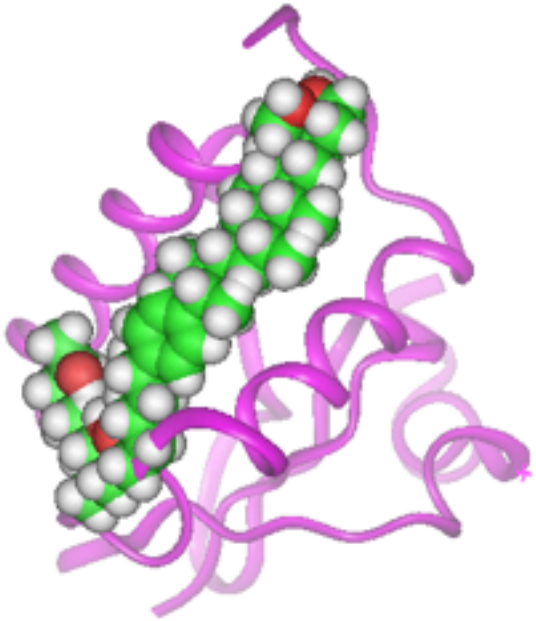
Upper tree structure



The string representation of this structure is as follows

AB3#CDEFG#H#I0J#0K1#L#MN2##0P##Q0R

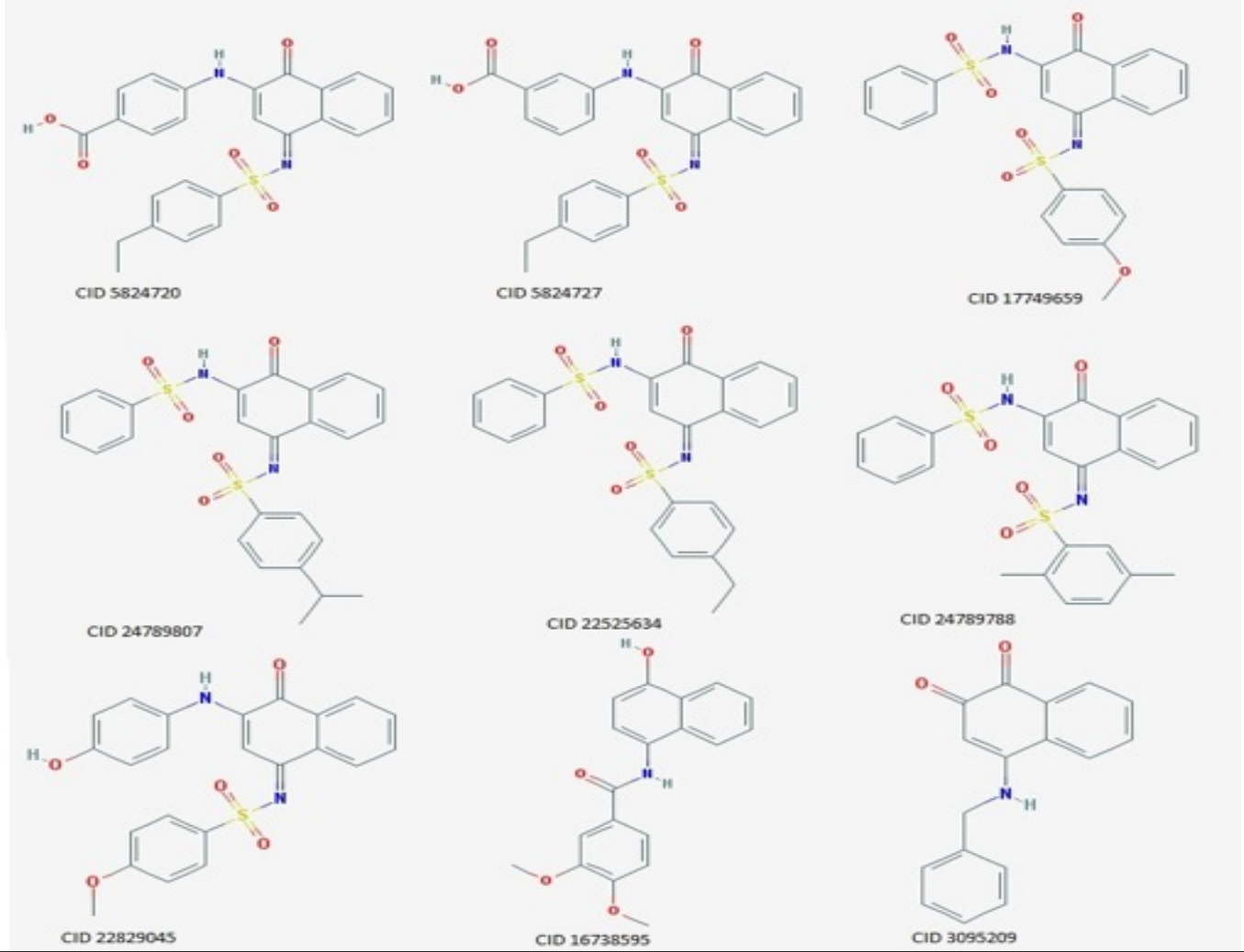
Results for HIV-1Nef Protein



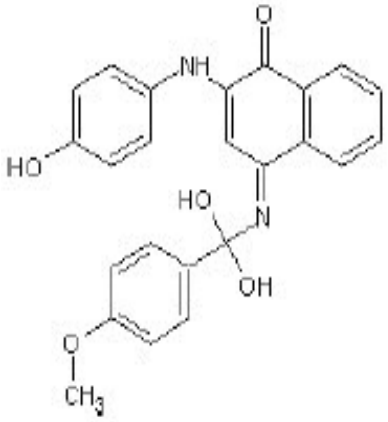
Docking Scores

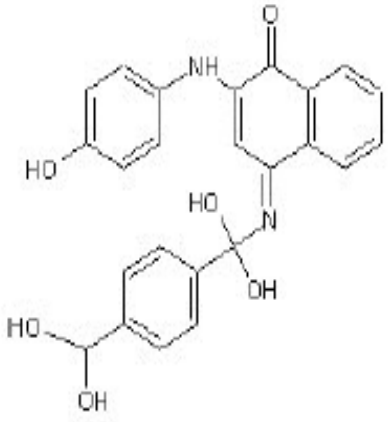
Ligand Identifier	VGA	LigBuilder	NEWLEAD	NSGA-II	MOLig
Ligand1	-1.67	-2.74	-3.64	-4.36	-4.96
Ligand2	-1.16	-2.11	-3.77	-4.11	-4.65
Ligand3	-1.27	-2.67	-3.21	-4.97	-4.11
Ligand4	-1.69	-2.59	-3.94	-4.51	-4.97
Ligand5	-2.08	-3.67	-3.47	-4.85	-4.82
Ligand6	-2.84	-3.47	-2.87	-4.28	-4.96
Ligand7	-1.07	-4.22	-2.61	-4.96	-4.24
Ligand8	-1.76	-4.68	-2.11	-5.74	-4.11
Ligand9	-1.92	-4.71	-2.89	-5.49	-5.99
Ligand10	-2.8	-3.36	-2.74	-5.61	-4.67
Ligand11	-2.17	-3.12	-2.61	-5.29	-5.74
Ligand12	-1.68	-3.92	-2.94	-5.11	-5.91
Ligand13	-2.11	-3.17	-2.98	-5.67	-5.96
Ligand14	-2.54	-2.95	-2.11	-5.97	-6.11
Ligand15	-1.32	-2.77	-2.39	-5.22	-6.97
Ligand16	-1.04	-2.12	-3.41	-5.11	-5.28
Ligand17	-1.29	-2.56	-2.99	-5.91	-6.47
Ligand18	−1.85	-2.39	-3.46	-5.22	-6.87
Ligand19	-3.28	-1.98	-3.75	-5.92	-6.8
Ligand20	-3.01	-1.96	-3.66	-5.67	-6.94
Maximum	-1.04	-1.96	-2.11	-4.11	-4.11
Minimum	-3.28	-4.71	-3.94	-5.97	-6.97
Mean	-1.9275	-3.058	-3.0775	-5.1985	-5.5265

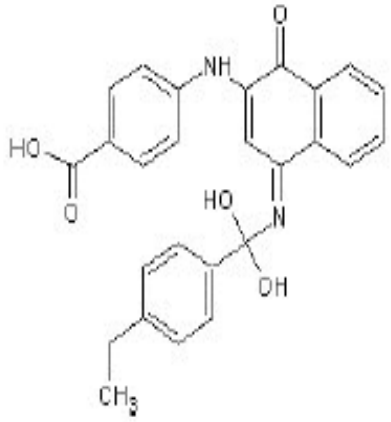
Known reconstructed inhibitors

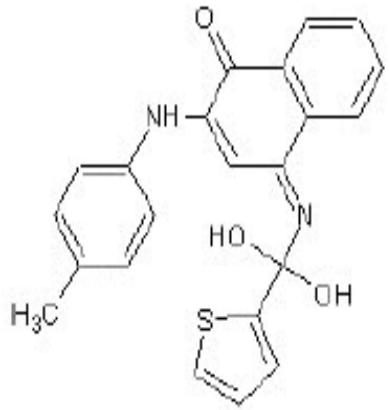


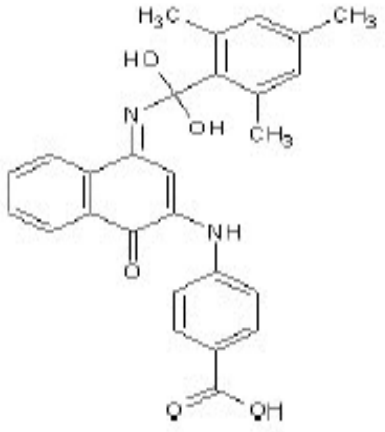
Predicted Probable Binders

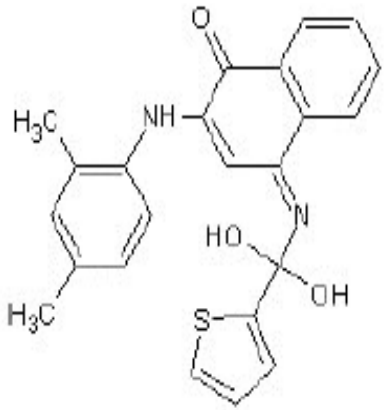




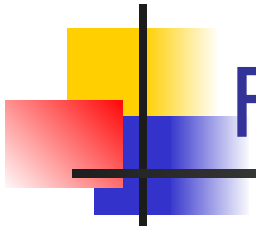








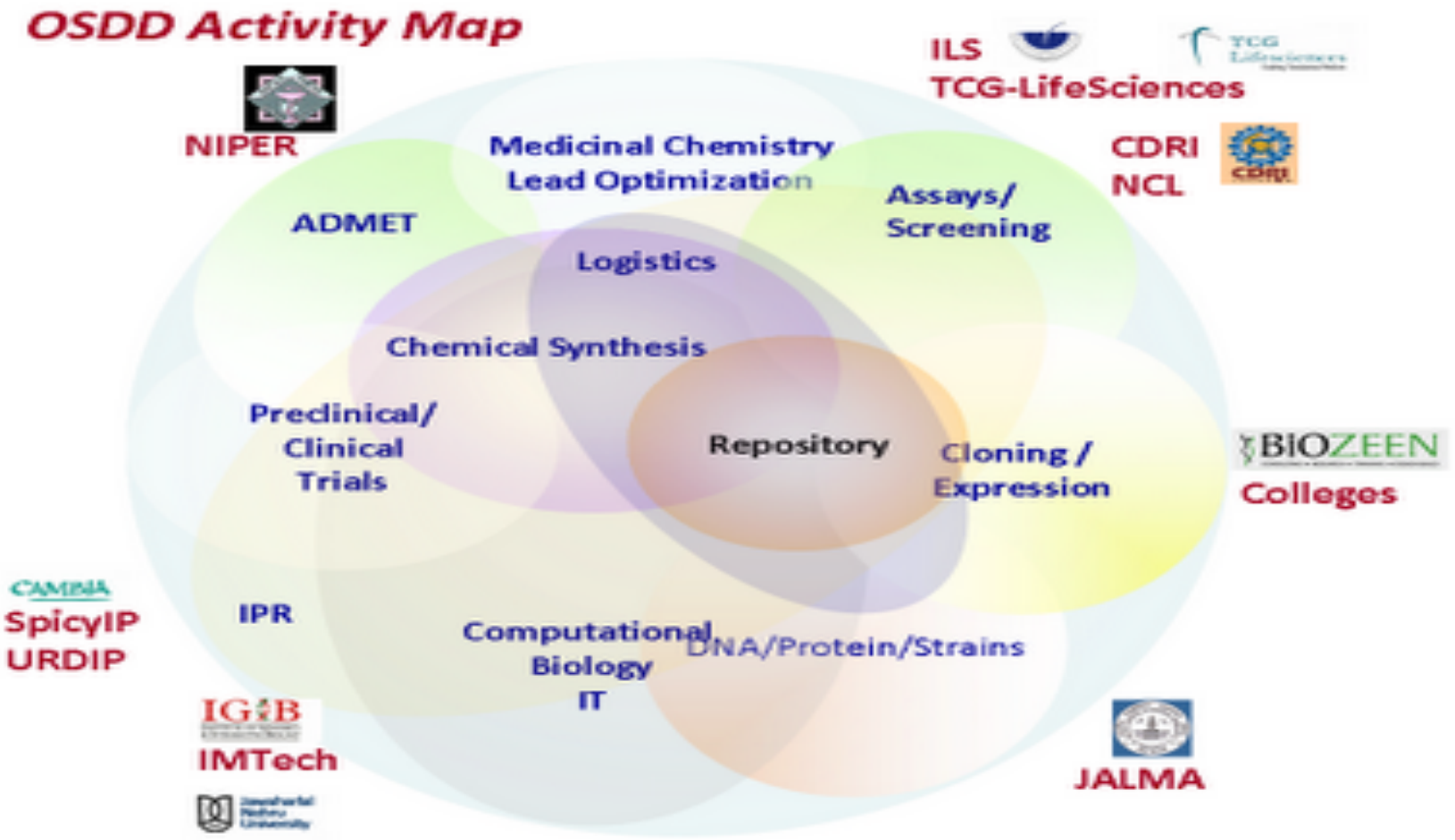
Future Missions



Future mission – OSDD

- OSDD – Open Source Drug Discovery
- A CSIR initiative
- Objectives
  - to provide affordable healthcare to the developing world
  - discovering novel therapies for neglected tropical diseases like Malaria, Tuberculosis, etc.

Future mission – OSDD How it works…



source: http://www.osdd.net/how-does-osdd-work

Future mission – OSDD who are involved…

- Nearly 400 motivated students, with the help of a few faculty and scientists, get together voluntarily and work single-mindedly [Source: THE HINDU]

- Average age – around 20s [Source: Business Line]

Future mission – OSDD

- Challenges:
  - ~1,000 mycobacterium tuberculosis genes out of ~ 4,000 have been annotated
  - re-annotation of the ~ 4,000- mycobacterium tuberculosis genes that determine how the TB bug lives and infects humans.
  - to unlock previously undiscovered details of tuberculosis
  - gene map should be similar to a Google map or a Wikipedia article that can be modified and updated as new information emerges on the features of the genome.

Source: http://www.osdd.net/news-updates

Future mission – Sophisticated drug delivery

- Effective treatment requires effective drug administration
  - Temporal control
    - Drug release timing
  - Spatial control
    - Location of drug release

Future mission – Temporal control (also known as sustained-release)

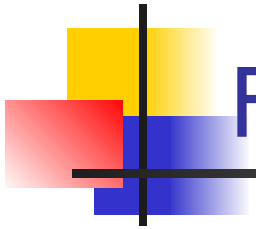
- Remember
  - If the active compound has a long half-life (over 6 hours), it is sustained on its own.
  - If the pharmacological activity of the active compound is not related to its blood levels, time releasing has no purpose.
  - If the absorption of the active compound involves an active transport, the development of a time-release product may be problematic.

Source: Wikipedia

Future mission – challenges in Gene therapy

- Gene delivery and activation –
  - Gene therapy will work only if we can deliver a normal gene to a large number of cells - say, several million - in a tissue.
- Introducing changes into the germline –
  - Delivering a gene to the wrong tissue would be inefficient and could cause health problems for the patient.
- Immune response –
  - Gene delivery vectors must be able to escape the body's natural surveillance systems. Failure to do so can cause serious illness or even death.
- Disrupting important genes in target cells
  - If the gene stitches itself into an inappropriate location or disrupting another gene may cause health problems for the patient.

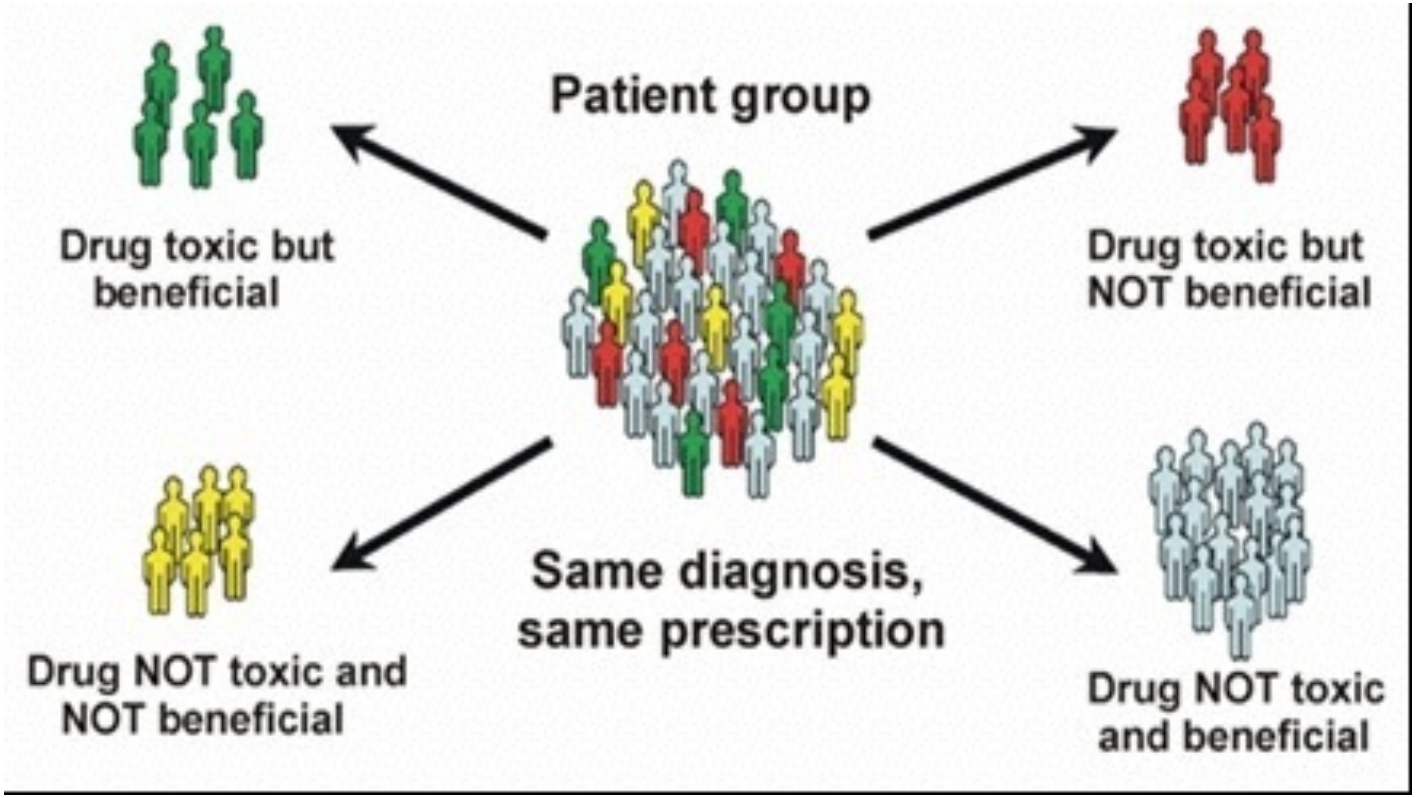
Source: http://learn.genetics.utah.edu/content/tech/genetherapy/gtchallenges/



Future mission – Personal genomics

- Personal genomics is a branch of genomics where individual genomes are genotyped and analyzed using bioinformatics tools.
- Motivation
  - rapid drop in the cost of sequencing a human genome
  - continual development of new, faster, cheaper DNA sequencing technologies such as "next generation DNA sequencing".
  - According to The National Human Genome Research Institute, part of the U.S. National Institute of Health
    - sequence a human-sized genome for US$100,000 by 2009
    - sequence a human-sized genome for US$1,000 by 2014

Future mission – Personalized medicine



References

- U. Maulik, S. Bandyopadhyay and A. Mukhopadhyay, Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics, Springer, Heidelberg, Germany, 2011.
- S. Bandyopadhyay and R. Mitra (2009). TargetMiner: microRNA target prediction with systematic identification of tissue specific negative examples. Bioinformatics, **25**, 2625-2631.
- S. Bandyopadhyay and S. Sengupta, "IVGA3D: De novo ligand design using a variable sized tree representation,"Protein and Peptide Letters, 17(12), 1495-516, 2010.
- S. Sengupta and S Bandyopadhyay, "De novo Ligand Design of Potential RecA Inhibitors using Multiobjective Optimization," IEEE Transaction on Computational Biology and bioinformatics, 2012.
- A. Chakraborty and S. Bandyopadhyay, "FOGSAA: Fast Optimal Global Sequence Alignment Algorithm", Scientific Reports (Nature Publishing Group) (accepted)
- Baskerville, S. and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA, **11**, 241–247.
- Bartel D. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, **116**, 281-297.
- Grimson et al., A. (2007). MicroRNA targeting specificity in mammals: Determination beyond see pairing. Molecular Cell, **27**, 91–105.

References cont…

- Kertesz et al., M. (2007). The role of site accessibility in microRNA target recognition. Nature Genetics, **39**, 1278–1284.
- Khvorova A., Reynolds A., and Jayasena S. (2003). Functional siRNAs and miRNAs exhibit strand bias, Cell, **115**, 209–216.
- Krek, A. et al. (2005) Combinatorial microRNA target predictions. Nat. Genet. **3**, 495–500
- Long et al., D. (2007). Potent effect of target structure on microRNA funcction. Nat. Struct. Mol. Biol., **14**, 287–294
- Robins et al., H. (2005). Incorporating structure to predict microRNA targets. Proc. Natl. Acad. Sci. USA, **102**, 4006–4009.
- Martinez et al., J. (2002). Singlestranded antisense siRNAs guide target RNA cleavage in RNAi, Cell, **110**, 563–574.
- Papadopoulos et al., G. L. (2009). The database of experimentally supported targets: a functional update of tarbase. Nucleic Acids Research, **37**, D155–D158.
- Selbach et al., M. (2008). Widespread changes in protein synthesis induced by microRNAs, Nature 2008, **455**, 58–63.

References cont…

- Smith TF., Waterman MS. (1981). Identification of common molecular subsequences. J Mol Biol, **147**, 195-197.
- Vapnik, VN. (1995). The Nature of Statistical Learning Theory. Berlin: Springer-Verlag.
- Vapnik, VN. (1998) Statistical Learning Theory. Wiley-Interscience 1998.
- Xiao et al., F. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res., **37**, D105–D110.
- Yousef et al., M. (2007). Naive bayes for microRNA target predictions-machine learning for microRNA targets. Bioinformatics, **23**, 2987–2992.
- S. Arnold, P. Franken, and M. P. Strub. The crystal structure of HIV-1 Nef protein bound to Fyn Kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. Structure, 5:1361–1372, 1997.
- S. Bandyopadhyay, A. Bagchi, and U. Maulik. Active Site Driven Ligand Design: An Evolutionary Approach. Journal of Bioinformatics and Computational Biology, 3:1053–1070, 2005.

References 61 cont…

- G. B. Dreyer, D. M. Lambert, T. D. Meek, T. J. Carr, T. A. Tomaszek, Jr. A. V. Fernandez, H. Bartus, E. Cacciavillani, and A. M. Hassell. Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease: structure-activity analysis using enzyme kinetics, x-ray crystallography, and infected T-cell assays. Biochemistry, 31:6646–6659, 1992.
- G. Goh and J. A. Foster. Evolving molecules for drug design using genetic algorithm via Molecular Tree. In Int. Conf. Genet. Evol. Comput., pages (27–33), 2000.
- A. Hillisch and R. Hilgenfeld, editors. Modern Methods of Drug Discovery. Springer, Germany, 2003.
- J. H. Holland. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, 1975.
- V. Tschinke and N. C. Cohen. The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypothesis. Journal of Medicinal Chemistry, 45A:1834–1837, 2006.
- J. M. Yang and C. Y. Kao. Flexible Ligand Docking Using a Robust Evolutionary Algorithm. Journal of Computational Chemistry, 21:988–998, 2000.

Thank you!!