

Computer Programs for Biological Problems: Is it Service or Science ?

G. P. S. Raghava

- **Simple Computer Programs**
 - Immunological methods
 - Methods in molecular biology
 - Other methods
- **Protein structure prediction**
 - Secondary & Supersecondary structure prediction
 - Supersecondary and Tertiary Structure prediction
- **Immunoinformatics: Tools for computer-aided vaccine design**
 - **B-cell epitope**
 - **T-cell epitope**
- **Genome annotation: Gene and Repeat prediction**
- **Functional annotation of proteins**
 - Subcellular localization
 - Classification of receptors
 - Analysis of Microarray Data
- **Work in Progress & Future**

Immunological Methods

➤ Computation of Ab/Ag Concentration from EISA d

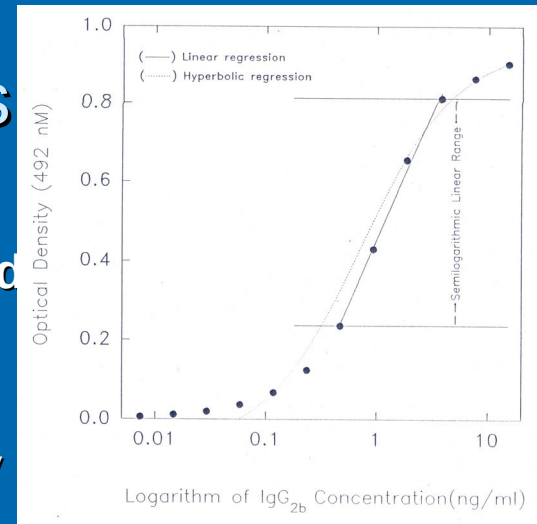
- Graphical Method
- Raghava et al., 1992, *J. Immuno. Methods* 153: 263

➤ Determination of affinity of Monoclonal Antibody

- Using non-competitive ELISA
- Serial dilution of both Ab and Ag concentration
- Law of mass equation
- Raghava and Agrewala (1994) *J. Immunoassay*, 15: 115

➤ Measurement and computation of IL-4 and Interferon- γ

- Ability to induce IgG1 and IgG2
- Agrewala et al. 1994, *J. Immunoassay*, 14: 83



Computer programs in GW-BASIC for PC, freely available

Methods in Molecular Biology

- GMAP: A program for mapping potential restriction sites
 - RE sites in ambiguous and non-ambiguous DNA sequence
 - Minimum number of silent mutations required for introducing a RE sites
 - Set theory for searching RE sites
 - **Raghava and Sahni (1994) Biotechniques 16:1116**
- DNASIZE: Improved estimation of DNA size from Gel Electrophoresis
 - Graphical method to improved prediction
 - **Raghava (1994) Biotechniques 17:100**
- DNAOPT: Optimization of gel conditions of gel electrophoresis and SDS-PAGE
 - Optimization of gel conditions
 - Sufficient distance between two fragments
 - Small fragment in range
 - **Raghava (1995) Biotechniques 18:274**

Other Methods

- **Hemolytic potency of drugs**
 - Raghava et al., (1994) *Biotechniques* 17: 1148
- **FPMAP: methods for classification and identification of microorganisms 16SrRNA**
 - graphical display of restriction and fragment map of genes;
 - compare the restriction and fragment map of genes
 - generate the fragment map of sequences in PHYLIP format
 - Raghava et al., (2000) *Biotechniques* 29:108-115
- **Nihalani, D., Raghava, G.P.S and Sahni, G (1997). Mapping of the plasminogen binding site of streptokinase with short synthetic peptides. Protein Science, 6:1284-92.**
- **Sarin, J., Raghava, G. P. S. and Chakraborti, P. K. (2003) Intrinsic contributions of polar amino acid residues towards thermal stability of an ABC-ATPase of mesophilic origin. Protein Science 12:2118-2120**

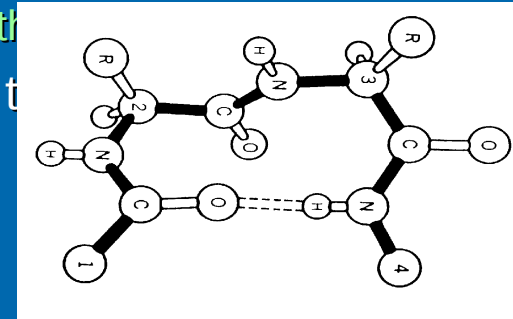
Protein Structure Prediction

➤ Regular Secondary Structure Prediction (α -helix β -sheet)

- APSSP2: Highly accurate method for secondary structure prediction
- Participate in all competitions like EVA, CAFASP and CASP (In top 5 methods)
- Combines memory based reasoning (MBR) and ANN methods

➤ Irregular secondary structure prediction methods (Tight turns)

- Betatpred: Consensus method for β -turns prediction
 - Statistical methods combined
 - Kaur and Raghava (2001) *Bioinformatics*
- Bteval : Benchmarking of β -turns prediction
 - Kaur and Raghava (2002) *J. Bioinformatics and Computational Biology*, 1:495:504
- BetaTpred2: Highly accurate method for predicting β -turns (ANN, SS, MA)
 - Multiple alignment and secondary structure information
 - Kaur and Raghava (2003) *Protein Sci* 12:627-34
- BetaTurns: Prediction of β -turn types in proteins
 - Evolutionary information
 - Kaur and Raghava (2004) *Bioinformatics* 20:2751-8.
- AlphaPred: Prediction of α -turns in proteins
 - Kaur and Raghava (2004) *Proteins: Structure, Function, and Genetics* 55:83-90
- GammaPred: Prediction of γ -turns in proteins
 - Kaur and Raghava (2004) *Protein Science*; 12:923-929.



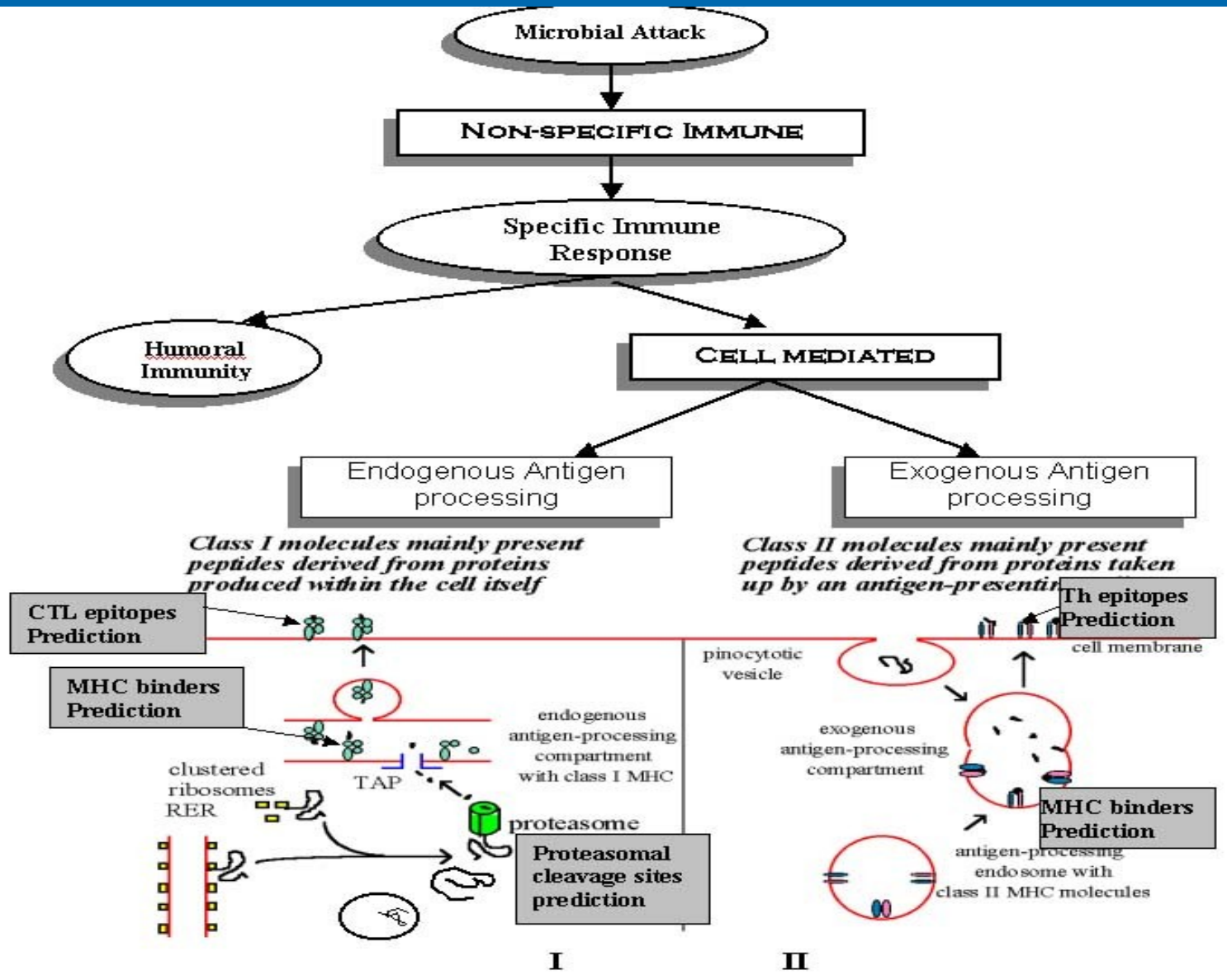
Protein Structure Prediction

- **BhairPred: Prediction of Supersecondary structure prediction**
 - Prediction of Beta Hairpins
 - Utilize ANN and SVM pattern recognition techniques
 - Secondary structure and surface accessibility used as input
 - [Manish et al. \(2005\) Nucleic Acids Research \(In press\)](#)
 - **TBBpred: Prediction of outer membrane proteins**
 - Prediction of trans membrane beta barrel proteins
 - Prediction of beta barrel regions
 - Application of ANN and SVM + Evolutionary information
 - [Natt et al. \(2004\) Proteins: 56:11-8](#)
 - **ARNHpred: Analysis and prediction side chain, backbone interactions**
 - Prediction of aromatic NH interactions
 - [Kaur and Raghava \(2004\) FEBS Letters 564:47-57 .](#)
 - **SARpred: Prediction of surface accessibility (real accessibility)**
 - Multiple alignment (PSIBLAST) and Secondary structure information
 - ANN: Two layered network (sequence-structure-structure)
 - [Garg et al., \(2005\) Proteins \(In Press\)](#)
 - **PepStr: Prediction of tertiary structure of Bioactive peptides**
- Performance of SARpred, Pepstr and BhairPred were checked on CASP6 proteins

Immunoinformatics: Tools for computer-aided vaccine design

- **Concept of vaccine and Drug**
 - **Drug:** Kill invaders/pathogens and/or Inhibit the growth of pathogens
 - **Vaccine:** Trained immune system to face various existing disease agents
- **Type of Vaccines**
 - Whole Organism of Pathogen (MTb, 4000 proteins)
 - Target proteins/antigens which can activate immune system
 - Subunit Vaccine: Antigenic regions which can simulate T and B cell response
- **Limitations of present methods of subunit vaccine design**
 - Developed for one or two MHC alleles (not suitable for large population)
 - Do not consider pathways of antigen processing
 - No single source of known epitopes
- **Initiatives taken by BIC at IMTECH**
 - In 2000, BIC take initiative to overcome some of limitations
 - To understand complete mechanism of antigen processing
 - Develop comprehensive databases

Immunoinformatics: Concept



Immunoinformatics: Databases Developed

MHCBN

- **A comprehensive database of mhc binding/ non-binding peptides, TAP binders and T-cell epitopes**
- **Largest database of T-cell epitopes (> 24,000 peptides)**
- **A set of data analysis tools e.g immunological BLAST, peptide mapping.**
- **Bhasin et al. (2003) Bioinformatics 19:665**

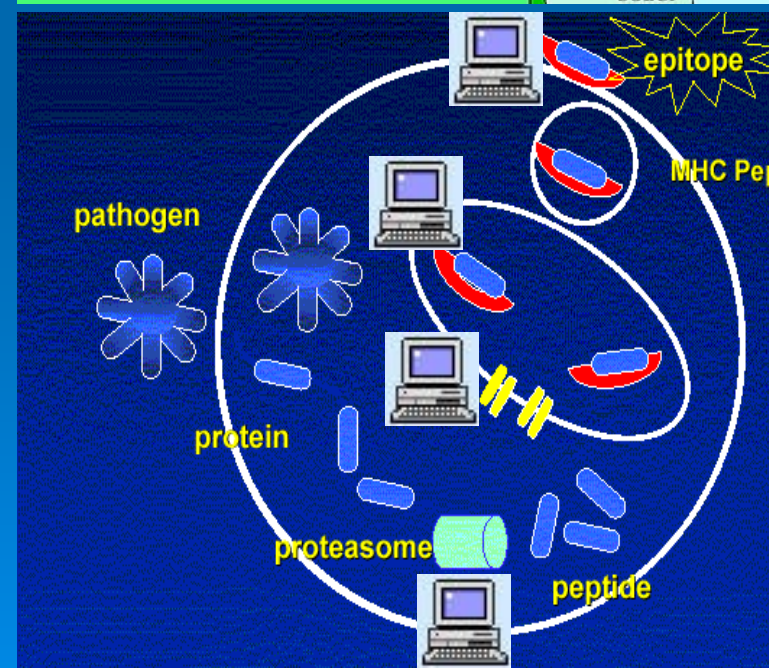
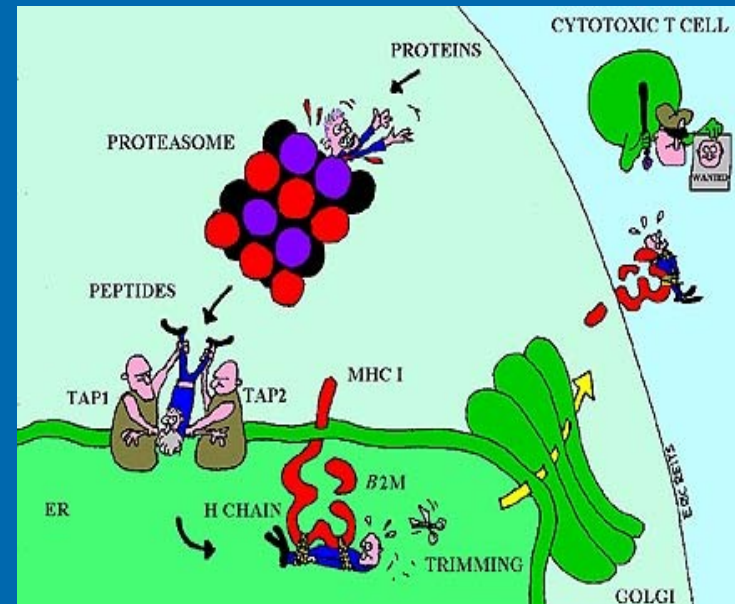
Bcipep

- **A database B cell epitopes**
- **Reference database of 3000 B cell epitopes.**
- **Hyperlinked to sequence databases**
- **Facilitate the mapping of T cell epitopes on B cell epitopes.**
- **Saha et al. (2005) BMC Genomics**

Both databases distributed by European Bioinformatics Institute (EBI), UK. Only databases from India distributed by EBI

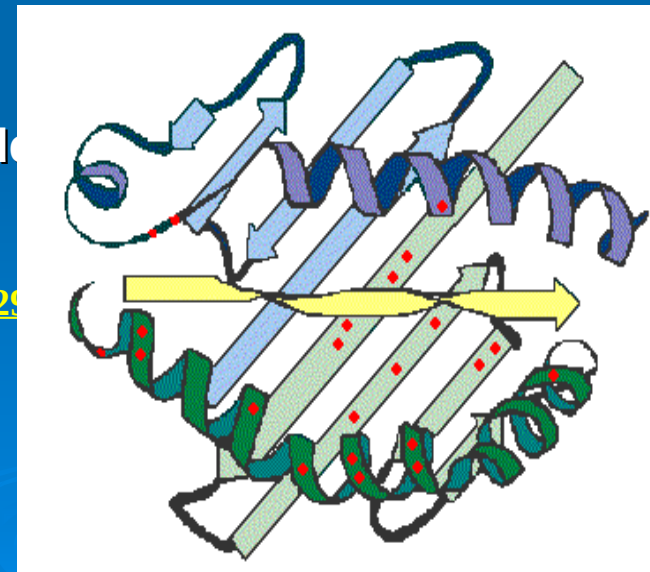
Immunoinformatics: Prediction of CTL Epitopes

- **Propred1**: Promiscuous binders for 47 MHC class I alleles
 - Cleavage site at C-terminal
 - Singh and Raghava (2003) *Bioinformatics* 19:1109
- **nHLApred**: Promiscuous binders for 67 alleles using ANN and QM
- **TAPpred**: Analysis and prediction of TAP binders
 - Bhasin and Raghava (2004) *Protein Science* 13:596
- **Pcleavage**: Proteasome and Immuno-proteasome cleavage site.
 - Trained and test on in vitro and in vivo data
 - Bhasin and Raghava (2005) *NAR (In Press)*
- **CTLpred**: Direct method for CTL Epitopes
 - Can discriminate CTL epitopes and Non-epitope MHC class I binders
 - Bhasin and Raghava (2004) *Vaccine* 22:3195



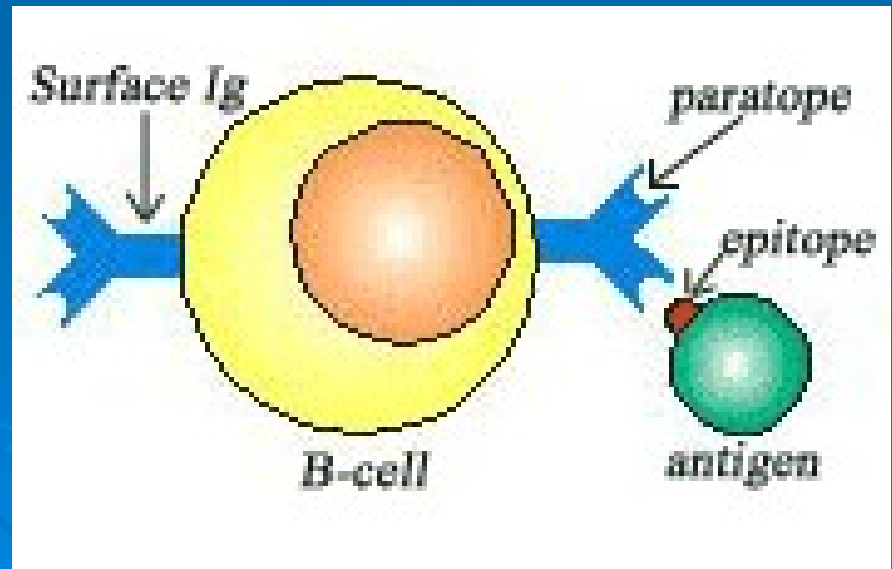
Immunoinformatics: T Helper Epitopes

- **Propred: Promiscuous of binders for 51 MHC Class II binders**
 - Virtual matrices
 - **Singh and Raghava (2001) *Bioinformatics* 17:1236**
- **HLADR4pred: Prediction of HLA-DRB1*0401 binding peptides**
 - Dominating MHC class II allele
 - ANN and SVM techniques
 - **Bhasin and Raghava (2004) *Bioinformatics* 12:421.**
- **MHC2Pred: Prediction of MHC class II binders for 41 alleles**
 - Human and mouse
 - Support vector machine (SVM) technique
 - Extension of HLADR4pred
- **MMBpred: Prediction of Mutated MHC Binders**
 - Mutations required to increase affinity
 - Mutation required for make a binder promiscuous
 - **Bhasin and Raghava (2003) *Hybrid Hybridomics*, 22:229**
- **MOT : Matrix optimization technique for binding core**
- **MHCBench: Benchmarking of methods for MHC binders**



Immunoinformatics: B-cell Epitopes

- **BCEpred: Prediction of Continuous B-cell epitopes**
 - Benchmarking of existing methods
 - Evaluation of Physico-chemical properties
 - Poor performance slightly better than random
 - Combine all properties and achieve accuracy around 58%
 - **Saha and Raghava (2004) ICARIS 197-204.**
- **ABCpred: ANN based method for B-cell epitope prediction**
 - Extract all epitopes from BCIPPEP (around 2400)
 - 700 non-redundant epitopes used for testing and training
 - Recurrent neural network
 - Accuracy 66% achieved



Genome annotation: Gene/Repeat prediction

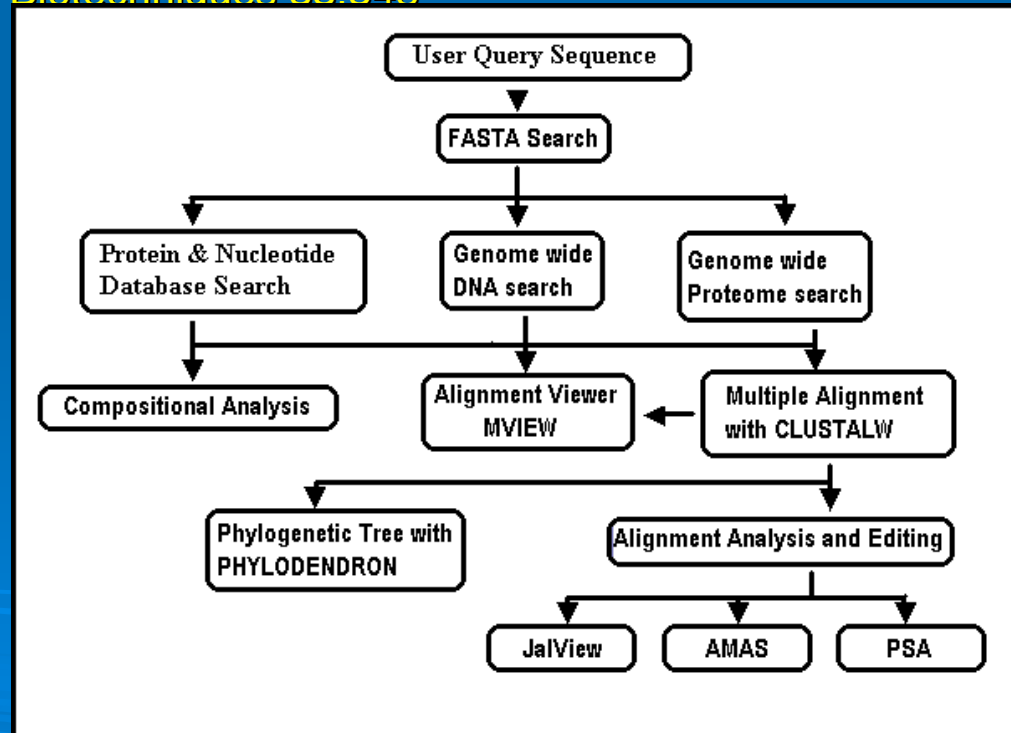
- **FTGpred: Prediction of Prokaryotic genes**
 - **Ab initio method for gene prediction**
 - **Based on FFT technique**
 - **Issac et al. (2002) Bioinformatics 18:197**
- **EGpred: Prediction of eukaryotic genes**
 - BLASTX search against RefSeq database
 - BLASTN search against intron database
 - probable intron and exon regions are compared to filter/remove wrong exons;
 - NNSPLICE program is used to reassign splicing signal site positions
 - finally *ab initio* predictions are combined with exons derived
 - **Issac and Raghava (2004) Genome Research 14:1756**
- **GeneBench: Benchmarking of gene finders**
 - Collection of different datasets
 - Tools for evaluating a method
 - Creation of own datasets
- **SRF: Spectral Repeat finder**
 - FFT based repeat finder
 - **Sharma et al. (2004) Bioinformatics 20: 1405**

Genome annotation: Comparative genomics

➤ **GWFASTA: Genome Wide FASTA Search**

- Standard FASTA search against nucleotide and protein sequences databases
- Search against nucleotide sequences of genomes (finished/unfinished)
- Search against protein sequences of proteomes (annotated only)
- **Issac and Raghava (2002) Biotechniques 33:548**

➤ **GWBLAST: Genome wide BLAST search**



Functional annotation of proteins: Subcellular localization

- **PSLpred: Sub cellular localization of prokaryotic proteins**
 - 5 major sub cellular localization
 - SVM based method
 - Accuracy of classification of final model 91%
 - [Bhasin and Raghava \(2005\) Bioinformatics 21: 2522](#)
- **ESLpred: Subcellular localization of Eukaryotic proteins**
 - SVM based method
 - Amino acid, Dipetide and properties composition
 - Sequence profile (PSIBLAST)
 - [Bhasin and Raghava \(2004\) Nucleic Acids Research 32:W414.](#)
- **HSLpred: Sub cellular localization of Human proteins**
 - Need to develop organism specific methods
 - Proteins belongs to same location have same type of composition
 - Higher eukaryote proteins are different than lower eukaryote in same location
 - 84% accuracy for human proteins
 - [Garg et al. \(2005\) Journal of Biological Chemistry 280:14427-](#)

Functional annotation of proteins: Classification of receptors

➤ **Nrpred: Classification of nuclear receptors**

- BLAST can easily identify the NR proteins (6 conserved domains)
- BLAST fails in classification of NR proteins
- SVM based method developed to identify four class of NR proteins
- Uses composition of amino acids
- **Bhasin and Raghava (2004) Journal of Biological Chemistry 279: 23262**

➤ **GPCRpred: Prediction of Families and Subfamilies of G-protein-coupled receptors**

- Predict GPCR proteins & class
- > 80% in Class A, further classify

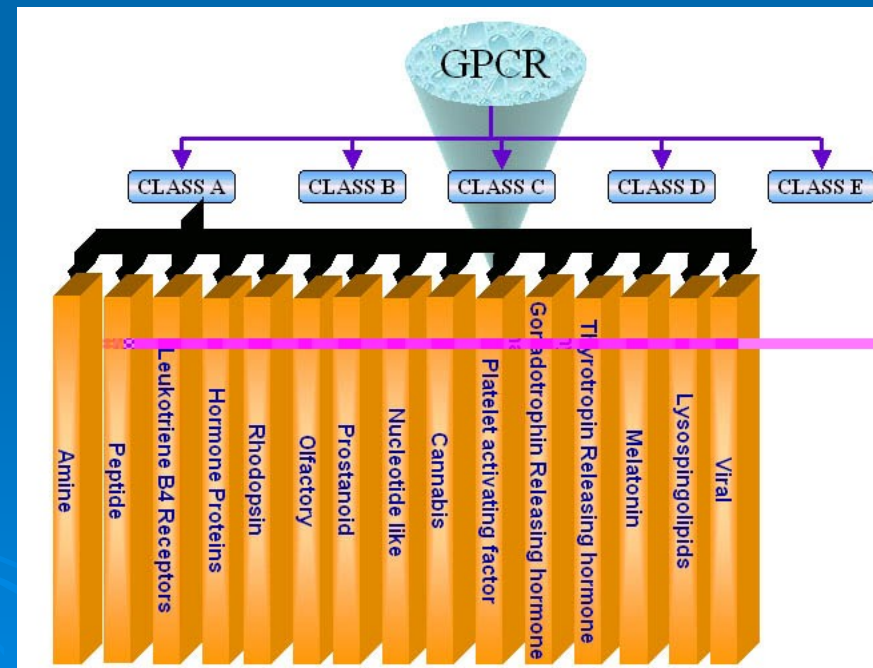
Bhasin and Raghava(2004) Nucleic Acids Research 32:W383

➤ **GPCRclass: Amine type of GPCR**

Major drug targets, 4 classes, Accuracy 96.4%

Acetylcholine; adrenoceptor; dopamine; serotonin

Bhasin and Raghava(2005) Nucleic Acids Research (In press)



Functional annotation of proteins: Analysis of Microarray data

- **LGEpred: Prediction of gene expression from amino acid composition of its proteins**
 - ❖ Analyze gene expression of *Saccharomyces cerevisiae*
 - ❖ Positive correlation between composition (Ala, Gly, Arg & Val) gene expression
 - ❖ Negative correlation for Asp, Leu, Asn & Ser
 - ❖ SVM based method for prediction of gene expression
 - ❖ Correlation 0,72, between predicted and actual expression
 - ❖ Amino acid composition with expression profile improves accuracy of function prediction
 - ❖ Membrane proteins have poor correlation between A.A. composition and expression
 - ❖ **Raghava and Han (2005) BMC Bioinformatics 6:1057**
- **Correlation and prediction of gene expression from its nucleotide composition**
 - Composition of G, C and G+C shows positive correlation with gene expression
 - Negative correlation for A, T and A+T
 - Inverse correlation between composition of a nucleotide at genome level
 - Correlation 0.87, between predicted and experimentally
- **Gene expression from codon biasness in gene and genome**
 - Major codon shows positive correlation
 - Correlation 0.85 between predicted and actual expression

Limitations: Only predict gene expression in a given condition, trained on one condition will not work in other condition

Summary of Major Publications

Name of Journal	Impact Factor of Journal (ISI 2003)	Number of Publication	
		In Last 5 Years	Total
Genome Research	9.6	1	1
Bioinformatics	6.7	10	10
Nucleic Acids Res.	6.6	9	9
Journal Biol. Chemistry	6.5	2	2
BMC Bioinformatics	4.9*	2	2
Proteins	4.3	3	3
Protein Science	3.8	4	5
FEBS Lett.	3.6	1	1
Vaccine	3.0	1	1
BMC Genomics	3.0*	1	1
J. Immuno. Methods	2.8	0	1
Biotechniques	2.4	2	6
Others	-	5	11

* Unofficial Impact factor of 2003

Work in Progress

- BTXpred: Prediction of bacterial toxins
- NTXpred: Classification of neurotoxins
- Mitpred: Prediction of mitochondrial proteins
- SRTpred: Identification of classical and non-classical secretory proteins
- AC2Dgel: Analysis and comparison of 2D gels
- VICMPred: Prediction of gram negative bacterial functional proteins
- HLA_Affi: Prediction affinity (real value) of HLA-A2 binders
- HaptenDB: Database of Haptens
- Functional annotation of Malaria

Acknowledgement

➤ Colleagues & Collobrators

- G. C. Varshney
- Girish Sahni
- J. N. Agrewala
- Amit Ghosh
- Chetan Premani
- Balvinder Singh
- Pradip Chakraborti
- Pushpa Agrawal
- G. C. Mishra
- Anish Joshi

➤ PhD Students

- Harpreet Singh
- Harpreert Kaur
- Manoj Bhasin
- Sudipto Saha
- Manish Kumar
- Sneh Lata

➤ Project assistants and Staff

- Aarti Garg
- Navjyot K. Natt
- Amit Kush
- Rajesh Solanki
- Mahender Singh
- Ruchi Verma