ELSEVIER

Applied Mycology
and Biotechnology
An International Series
Volume 6. Bioinformatics
© 2006 Elsevier B. V. All rights reserved

Computational Methods in Genome Research

Manoj Bhasin and G. P. S. Raghava
Institute of Microbial Technology, Sector 39 A, Chandigarh, India (raghava@imtech.res.in)

Computational biology has revolutionized biological and medical research. In the last two decades, a large number of computer methods have been developed to analyze DNA, RNA and protein sequences. These computer methods are playing a vital role in extracting useful information from sequences of genomes. These computational methods have been developed by different academic groups all over the world to serve the biological community. The methods are available as stand alone programs or on-line web servers. Most of these software packages are available free for academicians (freeware). In this chapter, we have described the major computational methods available for biologists to extract information from sequences. This chapter covers computational methods f r i) genome annotation; ii) comparative genomics; iii) protein structure prediction; iv) functional classification of proteins; and v) identification of potential vaccine candidates. These software packages are not available from a single source so it is not easy for users to obtain the software of their interest. In order to overcome this problem, attempts have been made to collect and compile a list of free biological software programs that includes software at EMBL and Indiana University. A catalog of biological software (Biocatalog) is also available on the internet (Rodriguez-Tome 1998). Recently, a repository of free software in biology has been created at Institute of Microbial Technology, Chandigarh, India which contains more than 800 free software packages.

1. INTRODUCION

The development of nucleotide sequencing technologies (Sanger et al. 1977; Maxam and Gilbert 1977). has allowed biologists to look into the primary sequences of genes and study their genetic implications. This yielded sequence data has resulted in creation of nucleotide databases like the EMBL nucleotide database, Genbank and DNA Databank of Japan (DDBJ) (Benson et al. 2004; Miyazaki et al. 2004). In the last two decades, sequencing technology has undergone tremendous improvement that lead to the development of fully automatic sequencing from

Corresponding author: G. P. S. Raghava

manual sequencing, due to the advent of shotgun sequencing technology. This development in technology has shifted the focus of biologists from sequencing of individual genes to the complete genome of an organism. In 1995, the first whole-genome sequence (1830 kb) of thebacterium *Haemophilus influenzae* Rd was announced (Fleischmann et al. 1995). After 1995, the number of organisms whose genomes have been sequenced is increasing at an exponential rate. The latest estimates show that around 47 archaeal, 600 bacterial and 450 eukaryote sequencing projects are in existence or have been completed. High throughput sequencing has produced a large amount of data that increases the demand of databases to maintain these sequences. The databases allow for the quick retrieval of sequences to promote knowledge-based studies. In addition to databases of raw data, a large number of specific databases are also available over the internet. The experimental analysis of large-scale genomic data is nearly impossible as it is a labor and time-intensive task. So, there is a need for computational tools to process the information rapidly. Bioinformatics tools can provide useful information to facilitate the experimental analysis of genomic data.

At present, thousands of computer software packages, web servers and databases have been developed to solve biological problems. In the 1980s, most software and databases were developed for stand-alone systems and one needs to install the programs on an individual workstation. After 1990, most of the software and databases were developed for web servers, where multiple users can run these algorithms via the internet without installing them locally on their desktop machines. These software programs are scattered across numerous web sites and are not arranged in any systematic manner. Most of the material is available for free but not utilized intensively because of the difficulty in finding the appropriate packages for specific needs. Biological software can be divided in two categories, i) commercial software and ii) free software (freeware). Commercial software packages have been developed by various companies for commercial purposes depending on the demand of the market. These software packages are user-friendly and provide multiple functions from a single point. In addition to usability, integration and user-interface, they also provide documentation and customer support. The major limitations of these commercial software packages are high cost, non-availability of source code, and up-to-date algorithms. In most of the cases costs of commercial software are too high, and it becomes nearly impossible for the biological community of developing countries to afford them. In contrast, freeware programs are usually developed by academicians, researchers or scientists for the benefit of public. Freeware is essential for promoting science. There are a few disadvantages of freeware, which include i) non-availability of proper documentation; and ii) implementation and distribution of the software. In the literature, only a few articles are available on freeware repositories. There is no comprehensive review on biological software repositories. By recognizing the importance of freeware in biology, the European Molecular Biology Laboratory (EMBL) at Heidelberg and Indiana University have created software repositories to collect and distribute free biological software. In this chapter we summarize the existing bioinformatics tools including resources from where one can download them.

2. ANALYSIS OF NUCLEOTIDE AND PROTEIN SEQUENCES

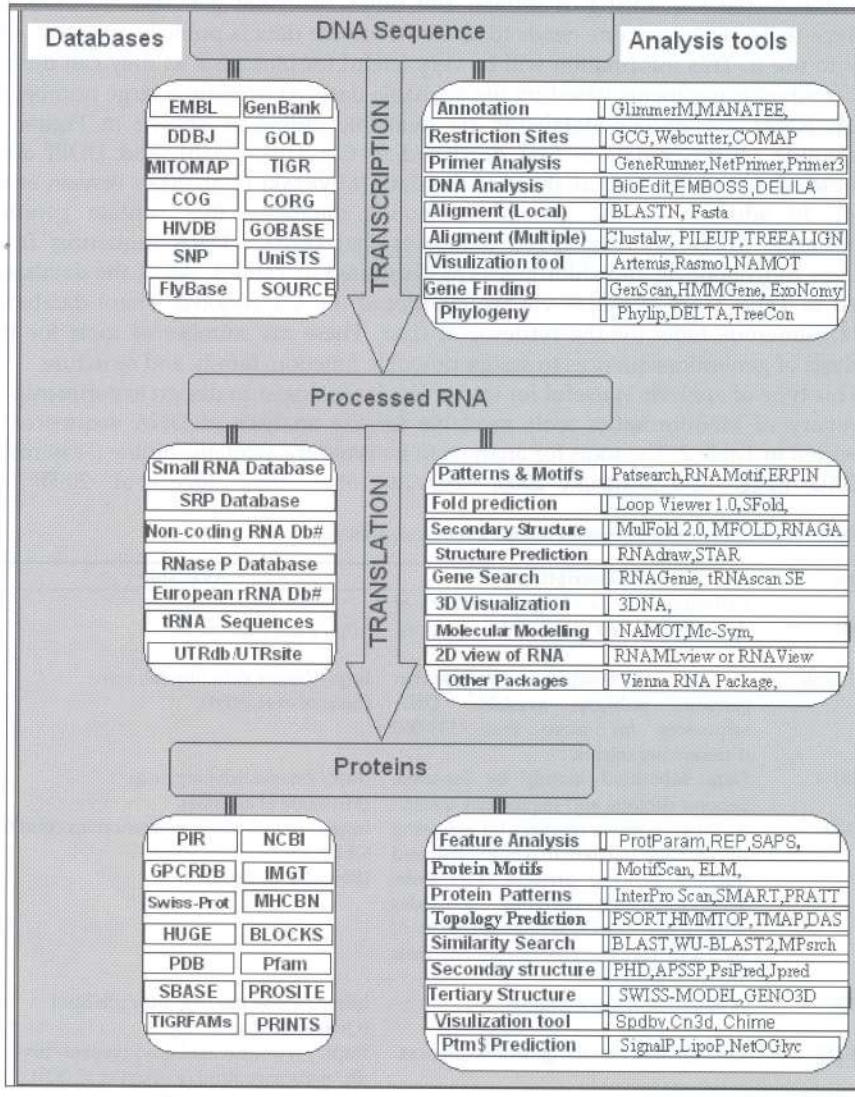
In this era of genomics, genetic sequences are being generated at an exponential rate due to the sequencing of human and other microbial genomes. In order to analyze these sequences one needs to know where the data is publicly available and how to use it. This information will be very useful for biologists as they can design and conduct experiments based on the available data. At present, a large number of DNA, RNA and protein databases are available online as shown in Figure 1. Nucleotide sequence data is first deposited to GenBank, EMBL and DDBJ after automated annotation and analysis of the data (Miyazaki et al. 2004; Benson et al. 2004). In addition to the general nucleotide databases that contain genome information, there are specialized databases that contain DNA sequences from viruses, mitochondria and chloroplasts as illustrated in Table 1. All of the databases listed in Table 1 are available via the internet at the URLs noted. These databases have numerous tools for the retrieval of data. There are number of tools for the analysis of genomic sequences to assign probable function, family and structure.

This type of analysis is useful for the molecular biologist to design experiments. A summary of bioinformatics tools available for the analysis of DNA sequences is provided in Table 2. The tools for analysis of features are available online (Majoros et al. 2003; Rozen and Skaletsky 2000; Schneider et al. 1982; Rice et al. 2000,

Table 1. A compilation of databases related to DNA sequences.

Name	Description	Available
EMBL	Collection of a large number of databases related to genomic sequences	http://www.ebi.ac.uk
GenBank	Comprehensive sequence database that contains publicly available DNA sequences for more than 119 000 different organisms.	http://www.ncbi.nlm.nih.gov. (Benson et al. 2004).
DDBJ	Data submitted mainly by Japanese genome projects and sequencing teams.	http://www.ddbj.nig.ac.jp (Miyazaki et al. 2004).
GOLD	A comprehensive resource for accessing information related to completed and ongoing genome projects worldwide. The database currently provides information on 350 genome projects, of which 48 have been completed and their analyses published.	http://igweb.integratedgenomics.com/GOLD/ (Bernal et al. 2004).
TIGR	A database of Gene Indices	http://www.tigr.org/tdb/tdb.html (Quakenbush et al. 2001).
HIVDB	Contain data on HIV genetic sequences	http://www.hiv.lanl.gov/content/hiv-db/mainpage.html (Kuiken et al. 2001).
GOBASE	Contains integrated sequence, RNA secondary structure and biochemical and taxonomic information about mitochondria and chloroplast.	http://megasun.bch.umontreal.ca/gobase/gobase.html (Shimko et al. 2001).
SNP	Contains information about human SNP.	http://www.broad.mit.edu/snp/human

Fig. 1. A diagrammatic view of important bioinformatics databases and tools for DNA, RNA and protein analysis. In figure Db# and Ptm$ refer to database and post-translational modification respectively.



Hall, unpublished, Raghava, 1994; Raghava, 1995; Raghava and Sahni 1994). Pprobable gene functions and families of novel sequences can be assigned by

performing similarity searches against annotated sequences in the major databases using software like BLAST and FASTA (Altschul et al. 1990; Pearson and Lipman 1988; Issac and Raghava 2002). Conserved patterns in sequences can be identified using software tools for multiple alignments like PILEUP and CLUSTALX (Thompson et al. 1994). The evolutionary history of a sequence can be traced through PHYLIP and TREECON programs as shown in Table 2 (J. Felsenstein, unpublished; Van de Peer 1994).

Table 2. A list of major bioinformatic tools for the analysis of DNA sequences.

Major Area	Program	Application or Function	URL (Reference)
Restriction sites	GMAP	A program for locating potential restriction sites.	http://www.imtech.res.in/raghava/progs/gmap (Raghava and Sahni 1994)
	WEBCUTTER	Online software for restriction mapping of nucleotide sequences.	http://www.firstmarket.com/cutter/cut2.html
Primer Design	GENERUNNER	Allow the designing of primer from nucleotide sequence.	http://www.generunner.com/
	NETPRIMER	Software for analysis of primers.	http://www.premierbiosoft.com/netprimer/
	PRIMER3	Picks primers for PCR reactions depending on oligonucleotide melting temperature, size, GC content, and primer-dimer possibilities and PCR product size.	http://www.broad.mit.edu/genome_software/other/primer3.html (Rozen and Skaletsky 2000)
DNA Analysis	BIOEDIT	Contains a large number of software programs which can be used for DNA or protein analysis.	http://www.mbio.ncsu.edu/BioEdit/bioedit.html (Hall, unpublished)
	EMBOSS	A suite of software programs that is useful for the analysis of DNA or protein sequences.	http://www.hgmp.mrc.ac.uk/Software/EMBOSS/ (Rice et al. 2000).
	DNAOPT	A program for optimizing gel conditions.	http://www.imtech.res.in/raghava/progs/danopt (Raghava 1995)
	DNASIZE	Computation of DNA fragment sizes.	http://www.imtech.res.in/raghava/progs/dnasize/ (Raghava 1994).
Visualization Tools	ARTEMIS	Genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation	http://www.sanger.ac.uk/Software/Artemis/ (Rutherford et al. 2000).
	RASMOL	Software for looking at macromolecular structure and its relation to function.	http://www.umass.edu/microbio/rasmol/
Phylogeny	PHYLIP	A program for inferring phylogenies	http://evolution.genetics.washington.edu/phylip.html
	DELTA	A flexible method for encoding taxonomic descriptions for computer processing	http://biodiversity.uno.edu/delta/ (Askevold et al. 1994).
	TREECON	A package for the construction of phylogenetic trees	http://iubio.bio.indiana.edu:7780/archive/00000138/ (Van de Peer 1994).

The analysis of genomic sequences to identify regions that code for RNA is also very useful for molecular biologists. There are three major classes of RNA (mRNA, rRNA and tRNA) generated during the process of transcription. Information related

to RNA is available from many repositories or databases as shown in Table 3. A brief description about each database and the corresponding URL are noted in Table 3. Since RNA is a very important molecule involved in transcription, splicing and other biochemical processes, many bioinformatics tools have been devised to analyze RNA sequences as depicted in Table 4. These tools are mostly available for prediction of RNA secondary structure (folded form) from the nucleotide sequence (Chen et al. 2000). (See Table 4). The prediction of structure can be performed using the tools based on energy minimization (e.g. MULFOLD) or based on conserved patterns in the sequence (RNADRAW) (Mtazura and Wenborg et al. 1996). Tools for finding conserved motifs and patterns from RNA sequences are also available from online resources (e.g. PATSEARCH, RNA MOTIF, and ERPIN) (Grillo et al. 2003).

In addition to the analysis of genomic and RNA sequences, analysis of amino acid sequences is also important for biologists. The analysis of primary amino acid sequence is useful to predict the secondary and tertiary structure of a protein and thereby its function. There are many databases, which contain information about proteins, as shown in Table 4. The primary sequences of proteins can be retrieved from databases like SWISS-PROT, PIR and NCBI (Boeckmann et al. 2003; Barker et al. 2000; Pruitt et al. 2003). Databases containing three-dimensional structures of proteins are also available online (e.g. PDB) (Robinson et al. 2003). There are many other protein databases which have specific functions such as MHCBN, IMGT (Robinson et al. 2003; Bhasin and Raghava 2003). The analysis of protein sequences in terms of structure and function is important from the perspective of a molecular biologist. In the arena of bioinformatics, a large number of tools are available for the analysis of protein sequences as shown in Table 5. Computational programs for the analysis of protein properties, specifically immunological properties, are also available on the web. Assignment of post-translational modification sites and protein topologies are also important from the biologist's point of view. In the last decade, highly accurate tools have been developed for predicting the post-transnational modification sites (NETOGLYC, LIPOP and SIGNALP) and topology or sub cellular localization (Nielsen et al. 1997; Juncker et al. 2003; Hansen et al. 1998; Nakai and Kanehisa 1991; Tusnady and Simon 2001). These tools can provide important insights about the biological functions of proteins. A summary of various bioinformatics tools or programs related to proteins is shown in Table 5.

3. GENOME ANNOTATION

The first step in genome annotation involves the integration of features revealed by the DNA and protein sequences into a systematic view of the organism's molecular machinery. Annotation can be divided into two processes; i) direct analysis of DNA sequences to locate coding regions and repeated elements, and ii) prediction of function and structure of the proteins encoded in the genome. In all organisms, coding regions are differentiated from neighboring non-coding regions by specific features. Detecting these features is essential to transforming sequence data into a fully annotated genome. Once a gene is identified or predicted, the next step is to assign a putative function, identify possible homologs in other organisms

Table 3. Databases related to RNA and protein sequences.

Database	Description	Availability (Reference)
Small RNA Database	Compilation of small RNA sequences including nuclear, nucleolar, cytoplasmic and mitochondrial small RNAs from eukaryotic organisms and small RNAs from prokaryotic cells and viruses.	http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html (Perumal et al. 1999).
SRPDB	Signal recognition particle database. Provides aligned, annotated and phylogenetically ordered sequences related to the structure and function of SRPs.	http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html (Rosenbald et al, 2003).
European rRNA Database	Curated database that contains complete or nearly complete LSU rRNA sequences in aligned form. Incorporates secondary structure information for each rRNA sequence.	http://rrna.uia.ac.be/lsu/ (Wuyts et al. 2004).
tRNA Sequence Database	Contains 3279 sequences of tRNA genes and tRNAs.	http://www.uni-bayreuth.de/departments/biochemie/trna/ (Sprinzl et al. 1998).
SWISS-PROT	A curated database of protein sequences.	http://www.ebi.ac.uk/swissprot/ (Boeckmann et al. 2003).
PROSITE	Consists of biologically significant patterns and profiles designed in such a way that with appropriate computational tools it can rapidly and reliably help to determine to which known family of proteins (if any) a new sequence belongs, or which known domain(s) it contains	http://www.expasy.org/prosite/ (Hulo et al. 2004).
MHCBN	MHC-binding, non-binding, and TAP-binding peptides.	http://www.imtech.res.in/raghava/mhcbn (Bhasin et al. 2003).
PIR	A comprehensive, quality controlled and well-organized protein sequence information resource.	http://pir.georgetown.edu/ (Barker et al. 2000).
PRINTS	Collection of protein fingerprints.	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/ (Attwood et al. 2003).

gene and within the ome, and to postulate its role in the biology of the organism. By comparing the genetic complement and genome organization of related organisms, novel insights may be realized regarding their evolutionary relationships.

Table 4: A compilation of important computational tools related to RNA

Major Area	Software	Description	Availability (Reference)
Patterns & Motifs	PATSEARCH	Pattern-matching tool that can find a well-defined pattern in a given sequence(s) or database (primary or specialized) divisions.	http://bighost.area.ba.cnr.it/BIG/PatSearch/ (Grillo et al. 2003).
	RNA MOTIF	A program for finding RNA motifs.	ftp://ftp.scripps.edu/pub/macke/
Fold Prediction	LOOP VIEWER 1.0	Graphical representation of RNA folding.	http://softwareseek.progenote.net/downloads/loopviewer.hqx
	SFOLD	Predicts RNA secondary structures, assesses target accessibility, and provides tools for the rational design of RNA-targeting nucleic acids.	http://sfold.wadsworth.org/index.pl
Secondary Structure Prediction	MULFOLD 2.0	Software for prediction of RNA secondary structure by free energy minimization.	http://softwareseek.progenote.net/downloads/mulfold.hqx
	MFOLD	RNA/DNA secondary structure prediction.	http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html
	RNAGA	Prediction of common secondary structures of RNAs.	http://bioweb.pasteur.fr/seqanal/interfaces/rnaga.html (Chen et al. 2000).
Structure Prediction	RNADRAW	An integrated program for RNA secondary structure calculation and analysis.	http://iubio.bio.indiana.edu/soft/molbio/ibmpc/rnadraw-readme.html (Matzura and Wenborg et al. 1996).
	STAR	Structure analysis of RNA using three different algorithms.	http://wwwbio.leidenuniv.nl/~Batenburg/STAR.html
3D Visualization	3DNA	A package for analyzing and rebuilding 3-dimensional nucleic acid structures.	http://rutchem.rutgers.edu/%7Exiangjun/3DNA/ (Lu and Olson 2003)
Molecular Modeling	NAMOT	A useful nucleic acid modeling tool.	http://namot.lanl.gov/
2D view of RNA	RNAMLVIEW	Generates 2-dimensional displays of RNA/DNA secondary structures with tertiary interactions.	http://ndbserver.rutgers.edu/servlet/RNAView.Frame2DMgr (Waugh et al. 2000).

estimating complete annotation of a genome includes information regarding gene location and organization, transcripts and products of those genes, as well as regulation and control of expression, translation and degradation. This process included boundaries between coding and non-coding sequence, identification of DNA features associated with gene structures, and translation of protein coding

Table 5. A list of computational tools for analysis of protein sequences.

Major Area	Software	Description	Availability
Feature Analysis	PROTPARAM	Allows the computation of various physical and chemical parameters for a given protein.	http://us.expasy.org/tools/protparam.html (Gasteiger et al. 2003).
	REP	Searches a protein sequence for repeats.	http://www.embl-heidelberg.de/~andrade/papers/rep/search.html (Andrade et al. 2000).
	PSA	Analysis of protein properties.	http://www.imtech.res.in/raghava/psa
Protein Motifs	MOTIFSCAN	Scans a sequence against protein profile databases.	http://hits.isb-sib.ch/cgi-bin/PFSCAN?
	ELM	Eukaryotic linear motif resource for functional sites in proteins.	http://elm.eu.org/ (Puntervoll et al. 2003).
Protein Patterns	INTERPRO	Assists in finding domains and family assignment by performing an integrated search in PROSITE, PFAM, PRINTS databases.	http://www.ebi.ac.uk/interpro/scan.html (Mulder et al. 2003).
	SMART	Allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.	http://smart.embl-heidelberg.de/ (Schultz et al. 1998).
	PRATT	Allows the user to search for conserved patterns in sets of unaligned protein sequences.	http://www.ebi.ac.uk/pratt/ (Jonassen et al. 1995).
Visualization tools	SPDBV	A user-friendly program that allows visualization and analysis of 3D structures of proteins.	http://us.expasy.org/spdbv/ (Guex and Peitsch 1997).
	Cn3D	A macromolecular structure viewer.	http://www.biosino.org/mirror/www.ncbi.nlm.nih.gov/Structure/cn3d/
	CHIME	A free program to show molecular structure in three dimensions.	http://www.umass.edu/microbio/chime/
Post-transnational Modifications	SIGNALP	Prediction of signal peptide cleavage sites.	http://www.cbs.dtu.dk/services/SignalP/ (Nielsen et al. 1997).
	LIPOP	Prediction of lipoproteins and signal peptides in Gram negative bacteria.	http://www.cbs.dtu.dk/services/LipoP/ (Juncker et al. 2003).
	NETOGLYC	Prediction of N-glycosylation sites in human proteins.	http://www.cbs.dtu.dk/services/NetNGlyc / (Hansen et al. 1998).

genes into protein sequence. The following subsections describe two of the major challenges in genome annotation; repeat prediction and gene prediction.

3.1. Repeat Prediction

The genomes of all organisms, particularly eukaryotic organisms, contain repetitive elements of varying lengths that can occupy a significant fraction of the total DNA content. For example, the human genome consists of more than 50% repeated sequences of various types (Lander et al. 2001). Repeats play a vital role in a number of regulatory functions and are responsible for instability of genomes. Many tandem repeats like the trinucleotide motifs, (e.g. CCG; CAG; AAG; CTG; GCG etc.) are associated with diseases such as fragile X, myotonic dystrophy, Huntington's, ataxia and others. Thus, identification of repeat elements is an important task in annotating a genome. Genomic repeat elements can be divided in two categories; i) tandem repeats which are usually confined to specific chromosomal regions, and ii) interspersed repeats mainly represented by inactive (pseudogenes) copies of historically or contemporarily active transposable elements (Strachan and Read 1999). Tandem repeats are grouped into three major subclasses; satellites, minisatellites and microsatellites (Strachan and Read 1999). Satellite repeats are composed of very long tandem arrays of short units usually present at centromeres. Minisatellites consists of tandem repeats of short units with lengths of about 7 to 64 bp located near telomeres, while microsatellite repeats are highly repetitive sequences consisting of 1 to 6 bp segments that are repeated up to 5 times the unit length as tandem arrays dispersed throughout all the chromosomes. Similarly, interspersed repeats can also be sub grouped into 5 types: SINEs (Short Interspersed Nuclear Elements) of 80-300 bp long units, LINEs (Long Interspersed Nuclear Elements) that are 6000-8000 bp long, LTRs (Long Terminal Repeats) that are 300 - 1000 bp long, and DNA transposons of variable lengths with two short inverted repeats flanking the element (Smit, 1996). Several repeat-finding algorithms have been developed to detect repeats, and these programs can be divided into two groups based on the type of repetitive DNA they identify; i) Tandem repeat finders and ii) interspersed repeat finders. Table 6 lists the major repeat finder programs available.

3.2. Gene Prediction

Correct predictions of gene location and structure are major challenges in the post genomic era, particularly for eukaryotic genomes. In the last decade a large number of computer programs have been developed for scanning genomic sequences to locate DNA segments that encode proteins. Prokaryotic genes may be predicted with considerable accuracy if one knows the codon usage pattern of the organism in question. A simple, long ORF (open reading frame) in a prokaryotic DNA sequence can be predicted as protein coding. The problem with gene prediction in prokaryotes lies in identifying the promoter and regulatory region.

Unlike prokaryotic genes, the eukaryotic genes are neither continuous nor contiguous. They are separated by long stretches of intergenic DNA and their coding sequences are interrupted by non-coding introns. Coding sequences occupy just a small

fraction of a typical higher eukaryotic genome. Additionally, some eukaryotic genes are alternatively spliced -- i.e. they have more than one possible exon assembly. The arrangement of genes in genomes is also prone to exceptions. Some genes are nested (overlapping) within each other (Dunham et al. 1999). The presence of pseudogenes further complicates the identification of protein coding regions. Regulatory sequences usually located upstream of coding sequences can sometimes be found downstream and within the introns of genes. In prokaryotic systems, genes are simple in structure where introns do not split protein-coding regions and they are comparatively easy to identify. However, finding genes in eukaryotic genomic sequences is far from being a

Table 6. List of major gene finders and repeat finder software.

Name	Description	URL/Reference
DOTTER	Finds repeats without prior knowledge using dot plot.	Sonnhammer and Durbin 1995
SPUTNIK	Finds small repeats using recursive algorithm	abajian.net/sputnik
TANDYMAN	Finds all exact repeats in an entire genome sequence	www.stdgen.lanl.gov/tandyman/index.html
TROLL	Tandem Repeat occurrence locator based on slight modification of the Aho-Corasick algorithm.	Castelo et al. 2002
FORREPEAT	FORRepeats: detects repeats on entire chromosomes and between genomes using novel data structure called factor oracle	Lefebyre et al. 2003
REPUTER	Applications of repeat analysis on a genomic scale	Kurtz et al. 2001
SRF	Identification of repeat sequences using Fourier transformation.	Sharma et al. 2004
GLIMMER	Primary microbial gene finder at TIGR, and has been used to annotate the complete genomes	www.tigr.org/software/glimmer/ (Majoros et al. 2003).
EGPRED	Similarity aided ab initio method for gene prediction	http://www.imtech.res.in/raghava/egpred (Issac and Raghava 2004).
GENSCAN	Identification of complete gene structures in genomic DNA.	http://genes.mit.edu/GENSCAN.html (Burg and Karlin 1997).
HMMGENE	Prediction of vertebrate and *C. elegans* genes	http://www.cbs.dtu.dk/services/HMMgene/ (Krog 1997).
FTG	Prediction of protein coding regions using Fourier transform	http://www.imtech.res.in/raghava/ftg (Issac et al. 2002).

trivial problem. Unlike prokaryotic genomes, the coding regions in eukaryotes represent only a small proportion of the eukaryotic genome and are mostly found to lie in non-repetitive regions of the genome. The major existing methods used for gene prediction are listed in Table 6.

4. COMPARATIVE GENOMICS

Comparative genomics is playing major role in extracting useful information from biological sequences. One important aspect of comparative genomics is the comparison of proteomes (the complete protein set) of two or more organisms. In addition, it involves the comparison of gene locations, relative gene order, and regulation. It also involves an examination of such events such as gene loss, duplications, and horizontal gene transfer. Such analyses aim to go beyond mere descriptions of similarities and differences, and they are directed toward the development of models and rules that might explain such events (Tatusov et al. 1997). What can we expect comparative genomics to reveal? One of the major goals of comparative genomics is to attempt prediction of gene function. Even for well studied bacteria such as *E. coli* (~ 4600 genes) and the well studied yeast, *S. cerevisiae* (~ 6500 genes), only 60-70% of the genes have known or predicted functions. An important goal is to understand the role of the remaining 30-40% of the genes. The field of comparative genomics has led to the development of novel tools and resources as well as new terminologies and vocabularies. A few important terminologies are defined here: Homology is the relationship of any two characters (such as two proteins that have similar sequences) that have descended, usually through divergence, from a common ancestral character. Homologs are genes/proteins with similar sequences that can be attributed to a common ancestor of the two organisms during evolution. Homologs can either be orthologs,paralogs, or xenologs. Orthologs are homologs that have evolved from a common ancestral gene by speciation. They usually have similar functions. Paralogs are homologous genes/proteins that are related or produced by duplication within a genome followed by subsequent divergence. They often have different functions. Xenologs are homologs that are related by an interspecies (horizontal transfer) of the genetic material for one of the homologs. The functions of the xenologs are quite often similar. Analogues are non-homologous genes/proteins that have descended convergently from unrelated ancestors. They have similar functions although they are unrelated.

Comparative genomics is a powerful approach for deciphering function through sequence comparisons, gene order, and regulation. These studies can also reveal insights into the recruitment of enzymes in a pathway. Specialized software tools can help to reveal how enzymes and domains are recruited and how enzymes are specifically lost in some lineages. In other words, comparative genomics may be useful to help us understand the genetic basis of diversity in organisms, both speciation and variation, events that are important aspects of evolutionary biology (Snel et al. 2000). Comparative genomic studies will also shed important light on the pathogenesis of organisms, as well as help in understanding and identifying human disease genes. Another important benefit of such analyses is the identification and development of novel drug targets (Frishman et al. 2003). These may be either virulence genes, uncharacterized essential genes, or species-specific genes. There are number of programs and databases which allow comparative analysis, and they are listed in Table 7.

Table 7. Software used for comparative genomics.

Software	Description	URL (Reference)
BLASTN	Method for rapid searching of nucleotide and protein databases. Since the BLAST algorithm detects local as well as global alignments, regions of similarity embedded in otherwise unrelated proteins can be detected.	http://www.ncbi.nih.gov/BLAST/ (Altschul et al. 1990).
GWFASTA	Compares a DNA sequence to another DNA sequence.	http://www.imtech.res.in/raghava/gwfasta (Issac and Raghava 2002).
BLAST	Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes/proteins. Both functional and evolutionary information can be inferred from well designed queries and alignments.	http://www.ncbi.nlm.nih.gov/BLAST/ (Altschul et al. 1990).
GWBLAST	A genome wide BLAST server.	http://www.imtech.res.in/raghava/gwblast
MPSRCH	Smith/Waterman sequence comparison at EBI.	http://www.ebi.ac.uk/MPsrch/
TREEALIGN	Phylogenetic alignment of homologous sequences.	http://bioweb.pasteur.fr/seqanal/interfaces/treealign-simple.html (Hein, 1990).
MBGD	Facilitates comparative genomics from various points of view such as ortholog identification, paralog clustering, motif analysis and gene order comparison.	http://mbgd.genome.ad.jp/
STRING	A tool for the retrieval of interacting genes/proteins.	http://string.embl.de/ (Snel et al. 2000).
PEDANT	It allows protein extraction, description and tools for analysis.	http://pedant.gsf.de/ (Frishman et al. 2003).
GENECENSUS	Tools for analysis of genomic data.	http://bioinfo.mbb.yale.edu/genome/

5. PROTEIN STRUCTURE PREDICTION

Knowledge of protein three-dimensional structure or tertiary structure (3D) is a basic prerequisite for understanding the function of a protein. Currently, the main techniques used to determine protein 3D structure are X-ray crystallography and nuclear magnetic resonance (NMR). In X-ray crystallography the protein is crystallized and then using X-ray diffraction the structure of protein is determined. Determination of 3D structure by X-ray crystallography is not always straightforward and sometimes takes as much as three to five years. NMR is another useful technique to determine the protein structure. The advantage of NMR over X-ray crystallography is that the protein can be studied in an aqueous environment that may resemble its actual physiological state more closely. The main limitation of NMR is that it is only suitable for small proteins that have less than 150 amino acids. The gap between known protein

sequences and the known protein structure is increasing exponentially. Thus, there is a need to develop the computational techniques to predict protein structures. Computer-aided protein conformation/tertiary structure prediction could facilitate i) the prediction of tertiary structures for proteins with known sequences and unknown structures, ii) understanding of protein folding, iii) engineering of proteins so that new functions may be incorporated, and iv) drug designing.

The problem of protein structure prediction has been approached through three main routes: 1) computer simulation based on empirical energy calculations, 2) knowledge based approaches using information derived from structure-sequence relationships from experimentally determined protein 3-D structures; and iii) hierarchical methods. Each approach has its merits and limitations.

5.1. Energy Minimization Based Methods

Protein structure predictions based on energy minimization methods are rooted in observations that native protein structures correspond to a system at thermodynamic equilibrium with a minimum free energy. Energy-based methods do not make *a priori* assumptions about the coding properties of amino acids. Rather attempts to locate the global minimum in surface free energy of the protein molecule is assumed to correspond with the native conformation of the molecule. Methods based on the principle of energy minimization can be classified broadly in two categories; i) static minimization methods and ii) dynamical minimization methods. The major software packages based on energy minimizations are AMBER; CHARMS; ECEPP; and GROMOS (Pearlman et al. 1995; van Gunsteren and Berendsen 1990; Brooks et al. 1990). Energy calculations offer the advantage of being based on physicochemical principles but are hampered by the large number of degrees of freedom to be considered and the limited performance of energy functions. There are essentially two major problems with methods based on energy calculations. First, the computations required for assigning protein structure based on energy minimization are beyond the reach of presently available computers. Secondly, the interaction potentials used for such calculations are not good enough to model the native structure of a protein at atomic detail (Somorjai 1990).

5.2. Knowledge Based Approaches

5.2.1. Homology modeling

Presently, homology modeling is the most powerful method for predicting the tertiary structure of proteins in cases where a query protein has sequence similarity to a protein with known atomic structure. (Blundell et al. 1987; Sali et al. 1990; Sutcliffe et al. 1987). These methods are based on the observation that structures are more conserved than sequences. Therefore, an accurate molecular model of a protein may be constructed by assigning a conformation that is based on sequence alignment, followed by model building and energy minimization. Due to the availability of plentiful genome sequence data, the number of protein sequences is increasing at an exponential rate, and the gap between the number of sequences and their corresponding structures is

widening. Therefore, construction of protein models is becoming an increasingly important technique (Orengo et al. 1992). The first crucial step in homology modeling involves generation of a structure-based alignment between the query protein and the sequence with known three-dimensional structure (Pascarella and Argos 1992). For cases of low homology (less than 20 % identity) the quality of the optimal alignments produced by automatic methods is often poor. A conceptually different approach to homology modeling is based on distance geometry. In this prospective, the tertiary template restrictions are translated into distance restraints that are used as input for distance geometry programs (Havel and Snow 1991; Sali and Blundell 1993). Homology-based modeling approaches fail in the absence of homologous structures.

5.2.2. Threading Approach

The concept of threading protein sequences through alternative folding motifs involves the construction of misfolded model structures, where an incorrect sequence is deliberately built onto the backbone of another protein. Threading a sequence through a fold requires a specific alignment between the amino acid sequence of the protein under consideration and the corresponding amino acid residue positions of the folding motif. The known structure establishes a set of possible amino acid positions in three-dimensional space. The query sequence is made similar to the known structure by placing its amino acids into their aligned positions. The primary aim of these methods is to select the most probable fold for a given sequence or to recognize suitable sequences that might fold into a given structure. The threading method is normally applied only on proteins whose amino acid sequences accept one of the protein folds previously studied by experimental techniques. The success of threading depends on the number of available folds whose structures are known at a level of atomic detail. In cases the atomic structure of folds are known then a query protein sequence can fitted with the known fold.

5.3. Hierarchical Approach

An alternate strategy for prediction of protein structures from their amino acid sequences uses the hierarchy of protein structure from primary to secondary and secondary to tertiary. An intermediate step in understanding the relationship between amino acid sequence and tertiary structure is to predict an intermediate state such as the secondary structure of a protein. This procedure involves constructing a model for the secondary structure from amino acid sequence data and use of the secondary structure model to build a tertiary structure prediction. There are a number of algorithms that have been developed for secondary modeling of proteins. Presently available methods can be classified into i) statistical methods, ii) physiochemical methods, (iii) artificial intelligence (AI) based methods, vi) evolutionary information based methods, and v) combinatorial methods (Rost 1996; Mcguffin et al. 2000; Cuff et al. 1998). Unfortunately, the prediction accuracy of secondary structures from sequence information is only about 80%. In using secondary structure models to predict tertiary structures attempts have been made to predict tight-turns and super secondary

structures in addition to helices, turns, sheets and strands (Kaur and Raghava 2003a; Kaur and Raghava 2003b; Kaur and Raghava 2004).

Table 8: A list of major software packages for protein structure prediction.

Software Program	*Use or Function*	*URL (Reference)*
PHD	A method for sequence analysis and structure prediction	http://www.embl-heidelberg.de/predictprotein/predictprotein.html (Rost 1996).
APSSP2	Advanced protein secondary structure prediction server.	http://www.imtech.res.in/raghava/apssp2/
PSIPRED	Allows prediction of protein secondary structure, topology of transmembrane domains and fold prediction.	http://bioinf.cs.ucl.ac.uk/psipred/ (Mcguffin et al. 2000).
JPRED	A consensus method for predicting protein secondary structure.	http://www.compbio.dundee.ac.uk/~www-jpred/ (Cuff et al. 1998)
BETATPRED2	Predicts beta turns in proteins from multiple alignments using neural networks.	http://www.imtech.res.in/raghava/betatpred2 (Kaur and Raghva 2003a).
GAMMAPRED	Predicts gamma turns in proteins from multiple alignments using neural networks.	http://www.imtech.res.in/raghava/gammmapred (Kaur and Raghava 2003b).
ALPHAPRED	Predicts alpha turns in proteins from multiple alignments using neural networks.	http://www.imtech.res.in/raghava/alphapred (Kaur and Raghava 2004).
SWISS-MODEL	An automated comparative protein modeling server.	http://www.expasy.org/swissmod/SWISS-MODEL.html (Peitsch et al. 1995).
GENO3D	Automatic modeling of protein three-dimensional structures.	http://geno3d-pbil.ibcp.fr/ (Combet et al. 2002).
CPHMODELS	Fold recognition/homology modeling.	http://www.cbs.dtu.dk/services/CPHmodels/
Meta Fold Recognition Server	Allows submission to multiple servers.	http://bioinfo.pl/Meta/ (Ginalski et al. 2003).
HMMSTR	Predicts the secondary, local, super secondary, and tertiary structures of proteins from sequences.	http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php (Bystroff and Shao 2002).
AMBER	A set of molecular mechanics force fields for the simulation of biomolecules.	http://amber.scripps.edu/ (Pearlman et al. 1995).
CHARMS	A set of programs for molecular simulation.	(Gunsteren and Berendsen 1990).

5.4. Benchmarking of Structure Prediction Methods

A major problem in the field of protein structure prediction is to assess the performance of existing methods. Methods have been developed using different sets of proteins and using different criteria for evaluation. In order to assist the developers and users, an open world wide experiment was initiated in 1994 called the Critical

Assessment of Techniques for Protein Structure Prediction (CASP). CASP experiments aim to establish the current state of the art in protein structure prediction by identifying what progress has been made and highlighting where future efforts may be most productively focused. These activities are held in alternate years, and the sixth CASP was initiated in December 2004 (http://PredictionCenter.llnl.gov/casp6). In addition to CASP, a number of other experiments were initiated to assess the performance of structure prediction methods such as the Critical Assessment of Fully Automatic Structure Prediction Servers (CAFASP), and the Evaluation of Automatic protein structure predictions (EVA). These experiments allow evaluation of online web servers for protein structure prediction. Table 8 lists major software and web servers for protein structure prediction.

6. FUNCTIONAL ANNOTATION & CLASSIFICATION OF PROTEINS

6.1. Subcellular Localization

Information concerning the subcellular localization of a protein may provide an important clue to elucidate its function, because it must be in the proper subcellular compartment to perform its biological function (Eisenhaber and Bork 1998). Knowledge about subcellular localization is sometimes useful in understanding disease mechanisms and for developing novel drugs. Therefore, the experimental determination of the subcellular localization of a protein constitutes one step on the long way to determine its biological function (Chou 2001). A number of methods have been developed for the prediction of subcellular localization in prokaryotes as well as eukaryotes. Such predictions for prokaryotic proteins are easy in comparison to eukaryotes due to the complex organization of eukaryotic cells. Similarity searches using BLAST and FASTA are commonly used to obtain evidence that a protein may be localized to a specific cellular compartment, however, these methods often fail in the absence of sequence similarity between query and target proteins (Eisenhaber and Bork 1998). Another way to predict subcellular localization is to identify local sequence motifs such as signal peptides or nuclear localization signals. Proteins designated for the secretory pathway, the mitochondria and the chloroplast contain N-terminal targeting peptides that are recognized by translocation machinery. Thus, these prediction methods will only analyze the N-terminus of the peptide. The reliability of methods based on sorting signals is strongly dependent on the quality of the gene sequence in the 5'-region or in the protein N-terminal sequence (Hua and Sun 2001). The major problem for methods that detect N-terminal sorting signals is that start codons are predicted with less than 70% accuracy by various genome projects and gene prediction methods. Prediction methods based on N-terminal sorting signals will be inaccurate when the signals are missing or only partially included (Reinhardt and Hubbard 1998). In addition, known signals are not general enough to cover the resident proteins of each compartment. To overcome these limitations, a number of methods based on amino acid and dipeptide composition have been developed (Table 9).

on various approaches such as hidden Markov models (Jaakkota et al. 2000), hierarchical assignments (Attwood et al. 2002), amino acid composition (Karchin et al. 2002), and dipeptide composition (Bhasin and Raghava 2004c).

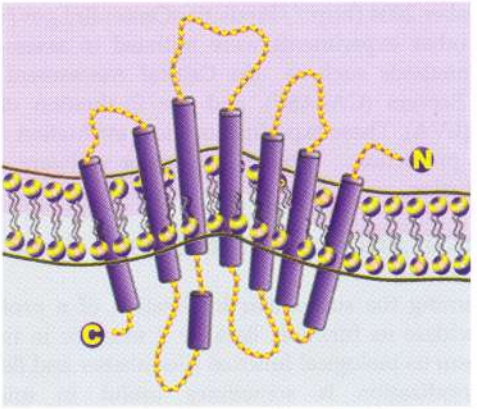


Fig. 2. GPCRs structure and topology in the cell membrane.

6.3. Nuclear Receptors

Nuclear receptors are key transcription factors that regulate crucial gene networks responsible for cell growth, differentiation and homeostasis. Recognition of nuclear receptors is crucial, because many of them are potential targets for developing therapeutic strategies for diseases like breast cancer and diabetes (Robinson-Rechavi and Laude 2003). All nuclear receptors consist of six distinct regions or domains (Figure 3). The N-terminal region (A/B) is highly variable, and contains one constitutionally active transactivation region (AF-1) and several autonomous transactivation domains. The A and B domains are variable in length from less than 50 to more than 500 amino acids. Recently, Bhasin and Raghava (2004) developed a method for predicting nuclear receptors (Bhasin and Raghava 2004a).

7. IDENTIFICATION OF VACCINE TARGETS

Traditionally vaccinations are achieved by injecting patients with a preparation of killed or seriously weakened (attenuated) virus or pathogen. Vaccines based on this approach can lead to potentially catastrophic results if, for some reason, the virus "catched" and the patient actually developed disease (Goldsby et al. 2000). Nevertheless, this approach has achieved limited success; however, the immunity raised is usually sufficient to provide only protection against individual isolates of a virus, and not for all isolates obtained. This is primarily due to the changing nature of viruses. In order to overcome the limitations of traditional vaccine design, there has been a significant change in the strategy for vaccine development in last few years. Presently, subunit vaccines are now employed in which vaccine candidates are derived from immunogenic peptides/regions in proteins instead of the complete antigenic protein. Most subunit vaccines are based on T cell epitopes. Therefore, identification of immunologically active regions/

7. IDENTIFICATION OF VACCINE TARGETS

Traditionally vaccinations are achieved by injecting patients with a preparation of killed or seriously weakened (attenuated) virus or pathogen. Vaccines based on this approach can lead to potentially catastrophic results if, for some reason, the virus "catched" and the patient actually developed disease (Goldsby et al. 2000).

Table 9: Prediction methods for subcellular localization of proteins and classification of proteins.

Program and Reference	Description	URL
PSORT (Nakai and Horton 1999)	For proteins of gram-negative bacteria. Based on rules derived from experimental data.	http://psort.nibb.ac.jp/form.html
iPSORT (Nakai and Kanehisa 1991)	For eukaryotic proteins. Based on amino acid sequences and features such as hydrophobicity and hydrophilicity.	http://www.hypothesiscreator.net/iPSORT
PSORT-B (Gardy et al. 2003).	For improved prediction of subcellular localization of proteins in gram-negative bacteria. Based on amino acid composition, presence of signal peptide, transmembrane alpha helices, motifs and similarity searches.	http://www.psort.org/psortb/
NNPSL (Reinhardt and Hubbard 1998)	For prokaryotic and eukaryotic proteins. Based on amino acid composition using ANN.	http://www.doe-mbi.ucla.edu/%7Eastrid/astrid.html
SUBLOC (Hua and Sun 2001)	For prokaryotic and eukaryotic proteins. Based on amino acid composition using SVM.	http://www.bioinfo.tsinghua.edu.cn/SubLoc
ESLPRED (Bhasin and Raghava 2004b)	For eukaryotic proteins. Based on amino acid, dipeptide composition, physiochemical properties using SVM.	http://www.imtech.res.in/raghava/eslpred/
GPCRPRED (Bhasin and Raghava 2004c)	For classification of GPCRs using SVM	http://www.imtech.res.in/raghava/gpcrpred
NRPred (Bhasin and Raghava, 2004a)	For classification of Nuclear receptors.	http://www.imtech.res.in/raghava/nrpred/

Nevertheless, this approach has achieved limited success; however, the immunity raised is usually sufficient to provide only protection against individual isolates of a virus, and not for all isolates obtained. This is primarily due to the changing nature of viruses. In order to overcome the limitations of traditional vaccine design, there has been a significant change in the strategy for vaccine development in last few years. Presently, subunit vaccines are now employed in which vaccine candidates are derived from immunogenic peptides/regions in proteins instead of the complete antigenic protein. Most subunit vaccines are based on T cell epitopes. Therefore, identification of immunologically active regions/epitopes recognized by T cells plays a crucial role in

subunit vaccine design (Masignani et al. 2002; Singh and Raghava 2001, Rappuoli 2000). Experimental methods for the identification of such regions are costly and time-consuming. Therefore, computational methods for prediction of such sites are of great value. In the last decade, a large number of computational methods were developed to predict T/B cell epitopes for potential vaccine candidates as described in the following sections.

7.1. B-cell Epitopes

The antigenic regions of proteins that are recognized by binding sites or paratopes of immunoglobulin molecules are called B-cell epitopes. These epitopes may be linear (continuous) or conformational (discontinuous). These epitopes play an important role in peptide-based vaccines design, disease diagnosis and allergy research. The development of computational methods for prediction of B-cell epitopes remains a vital and challenging task due to the inherent complexity of antigen recognition. It is nearly impossible to predict conformational epitopes as it require knowledge of the tertiary structure for both antigens and antibodies. In the past, a number of algorithms were developed for predicting continuous/linear B-cell epitopes based on physicochemical properties such as hydrophilicity, flexibility, accessibility, and turns (Hopp and woods 1983; Kyte and Doolittle 1982; Karplus and Schulz 1985; Kolaskar and Tongaonkar 1990; Alix 2000; Odorico and Pellequer 2003). Recently, Saha and Raghava (2004). (http://www.imtech.res.in/raghava/bcepred/) studied the performance of several computational methods on a clean and large data set of B-cell epitopes (Saha and Raghava unpublished; http://www.imtech.res.in/raghava/bcipred/). The performance of algorithms based on physicochemical properties varied from 52.9% to 57.5%, whereas, combined methods showed 58.7% accuracy. Recently our group has used artificial neural networks to predict linear B cell epitopes (http://www.imtech.res.in/raghava/abcpred/).

7.2. T-cell Epitopes

Extracellular antigens are processed via an exogenous pathway and recognized by helper T (T_h) cells, whereas, intracellular antigens processed via the endogenous path are recognized by cytotoxic T-lymphocytes (Watts and Powis 1999). Earlier methods for prediction of T cell epitopes were based on analysis of experimentally determined T cell epitopes and are known as direct methods of T-cell epitope prediction (Table 10). These methods are based on the assumption that the conformation of a peptide is responsible for its recognition by T cells. These methods were superseded after the analysis of MHC peptide complexes by X- ray crystallography, which demonstrated that a peptide bound in the MHC groove has an extended conformation (Stern et al. 1994). It was also observed that binding of a peptide to an MHC allele requires more specificity than its recognition by T-cells. This started a new era of predictive methods called indirect methods of T-cell epitope prediction where the predictor identifies the MHC binding regions in an antigen rather than T-cell epitopes.

linearity in the data. Therefore, machine learning techniques like artificial neural networks (ANN) and support vector machines (SVM) have been introduced for MHC class II binder prediction. Machine learning

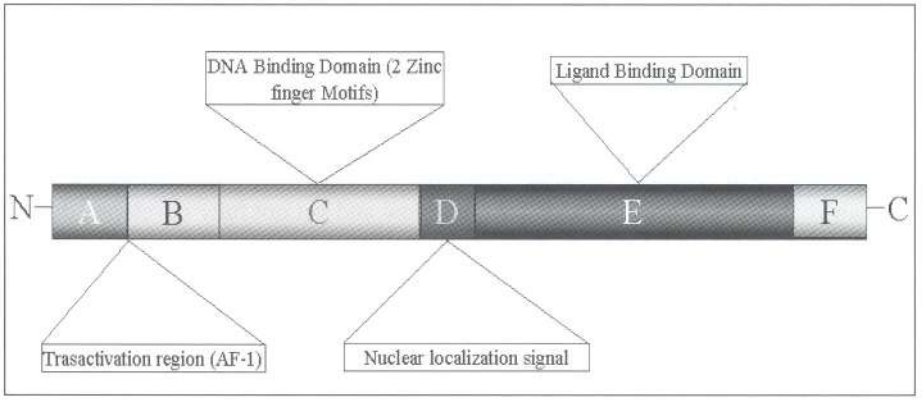


Fig. 3. Schematic representation of nuclear receptors.

based methods have achieved better accuracy compared to matrices and motif-based methods (Brusic et al. 1998; Bhasin and Raghava 2004d). The major existing methods for MHC class II binder prediction are summarized in Table 10.

7.2.2. Prediction of MHC class I binders

Several methods have been developed for prediction of MHC class I binding peptides from antigenic sequences (Table 10). Designing methods for MHC class I binders prediction is easy compared to predicting class II binders since the length of binding peptides is nearly fixed. Preliminary methods were based on motifs (Rammensee et al. 1995). SYFPEITHI is the most successful method based on refined motifs derived from pooled sequences and single peptide analysis exclusively of natural ligands. The motif-based methods have low accuracy, because all MHC binders do not contain exact motifs. Quantitative matrix-based methods consider all positions and residues in peptides to determine its MHC binding potential. These methods fail in handling non-linear data. In order to handle non-linearity in data and to adapt self-learning, machine learning techniques like artificial neural networks (ANN) and support vector machines have been introduced for prediction of MHC class I binders. All of the above approaches are knowledge-based where rules are derived from known binders and non-binders. Another alternative to the knowledge-based approach is a structure-based prediction in which the conformations of peptides to fit in the MHC groove are studied. Hanan Marglit and co-workers (2000). devised a method for prediction of MHC binding peptides on the basis of structural information (Schueler-Furman et al. 2000). These methods are quite slow and yet not fully developed due to limited information from the MHC and peptides.

ligands. The motif-based methods have low accuracy, because all MHC binders do not contain exact motifs. Quantitative matrix-based methods consider all positions and residues in peptides to determine its MHC binding potential. These methods fail in handling non-linear data. In order to handle non-linearity in data and to adapt self-learning, machine learning techniques like artificial neural networks (ANN) and support vector machines have been introduced for prediction of MHC class I binders. All of the above approaches are knowledge-based where rules are derived from known binders and non-binders. Another alternative to the knowledge-based approach is a structure-based prediction in which the conformations of peptides to fit in the MHC groove are studied. Hanan Marglit and co-workers (2000). devised a method for prediction of MHC binding peptides on the basis of structural information (Schueler-Furman et al. 2000). These methods are quite slow and yet not fully developed due to limited information from the MHC and peptides.

8. RESOURCES

8.1. Software at EMBL

EMBL is a major source of free biological software, and more than 300 software packages are available (Stoehr and Omond 1989). These software packages are also available from the European Bioinformatics Institute (http://www.ebi.ac.uk/), an outstation of EMBL. These programs are divided in four categories based on operating systems: i) MS-DOS or Windows, ii) Apple Macintosh, iii) UNIX, and iv) VAX-VMS (Fuchs 1990). The software available at EBI is kindly provided by its authors. A major advantage of this repository is that software can be obtained by email, ftp or http.

Email Server: To obtain information about software, users should send email to netserv@ebi.ac.uk with command "help software" in the body of the message.

- **FTP Server:** Software is obtained by anonymous ftp from EBI (ftp.ebi.ac.uk).
- **Web Server:** Software may be downloaded via the internet from http://www.ebi.ac.uk/

All of the files at EBI are converted into printable ASCII format, so they can be distributed via standard email. This repository encourages authors to make their software available to the scientific community.

8.2. Freeware at Indiana University

Indiana University offers a large collection of software packages for biology (Gilbert 2000). This repository allows users to browse, search and download available software packages from http://iubio.bio.indiana.edu/.

8.3. BioCatalog

The BioCatalog is a database that contains information about biology and genetics software (Rodriguez-Tome, 1998). It is different from other software repositories, because it maintains information about software rather than software itself. It can be accessed at http://www.ebi.ac.uk/biocat/. The catalog is freely distributed as an ASCII

file. It categorizes software packages based on their functions. Users can download the catalog from EBI at ftp://ftp.ebi.ac.uk/databases/biocat/.

Table 10: Summary of methods used for prediction of potential vaccine candidates.

Program	Description	URL
MHC Class II Binder Prediction methods		
SYFPEITHI (Rammensee et al. 1999)	Motif based prediction of large numbers of MHC class I and class II alleles.	http://www.syfpeithi.de/
PROPRED (Singh and Raghava 2001)	Prediction of promiscuous binders for 51 HLA class II alleles using virtual matrices.	http://www.imtech.res.in/raghava/propred/ or http://bioinformatics.uams.edu/mirror/propred/
TEPITOPE (Sturniolo et al. 1999)	Prediction of promiscuous binders for 25 HLA class II alleles using virtual matrices.	PC program can be downloaded from (http://www.vaccinome.com/)
HLADR4PRED (Bhasin and Raghava 2004d)	Prediction of binders for HLA-DRB1*0401 using SVM and ANN.	http://www.imtech.res.in/raghava/hladr4pred/
T cell epitope Prediction		
CTLPRED (Bhasin and Raghava 2004e)	A direct method for CTL epitope prediction.	http://www.imtech.res.in/raghava/ctlpred/
EPIMER & OPTIMER (Meister et al. 1995)	Based on density of MHC binding motifs and their conformation.	---
AMPHI (Spouge et al. 1987; Margalit et al. 1987)	Based on the assumption that T cell epitopes have amphipathic alpha-helices.	---
MHC Class I Binder Prediction		
PROPRED1 (Singh and Raghava 2003)	Matrix based prediction of promiscuous binders of 47 MHC class I alleles.	http://www.imtech.res.in/raghava/propred1/ or http://bioinformatics.uams.edu/mirror/propred1/
nHLAPRED	Prediction of promiscuous binders for 67 MHC class I alleles using ANN and QM techniques.	http://www.imtech.res.in/raghava/nhlapred/ or http://bioinformatics.uams.edu/mirror/nhlapred/
BIMAS	Ranks potential peptides based on predicted half-time of dissociation to HLA class I molecules.	http://www-bimas.cit.nih.gov/molbio/hla_bind/

8.4. RFSB: Repository of Free Software in Biology

In order to promote free software resources in biology, the Bioinformatics Centre (BIC) at the Institute of Microbial Technology (IMTECH) in India has initiated a project called Public Domain Resources in Biology (PDRB). The major goal of this project is to collect, manage and distribute free biological software to the academic community via the World Wide Web. Under this project, a repository of free software has been

developed (Raghava 2001a), and the latest version of the database contains more than 800 biological software packages. This is largest repository of free software in biology. The RFSB was created using POSTGRESSQL, a free RDBMS program. The database stores the following information about each software package: i) program name; ii) category of software based on function; iii) operating system requirements; iv) main function; v) a brief description of the software; vi) reference (if published); vii) authors; viii) hardware requirements; ix) software requirements; and x) original ftp/http site for obtaining information and downloading the software. In addition to offering users the ability to download software packages, the database also allows the submission of new software via the internet.

9. CONCLUSION

Improvements in sequencing technologies has lead to the elucidation of complete genomes for a large number of organisms. Therefore, annotation of these genomes and assignment of functions for the corresponding genes and proteins is a major challenge in the field of genome research. Fortunately, a number of software packages and web servers have been developed which facilitate genome analysis and functional annotation. Within the scope of this chapter it is not possible to describe all the computer programs that have been developed for genomics, so we have focused primarily on those that are most popular. We have made an attempt to provide an overview of various computational tools that can assist in biological research.

REFERENCES

Alix AJP (2000). Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 18: 311-314.

Altschul S F, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. J Mol Biol 215: 403-410.

Andrade MA, Ponting C, Gibson T and Bork P (2000). Identification of protein repeats and statistical significance of sequence comparisons. J Mol Biol 298: 521-537.

Askevold IS and O'Brien CW (1994). DELTA, an invaluable computer program for generation of taxonomic monographs. Ann Entomol Soc Am 87: 1-16.

Attwood TK, Croning MD and Gaulton A (2002). Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. Protein Eng 15: 7-12.

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL,Moulton G, Nordle A, Paine K, Taylor P, Uddin A and Zygouri C (2003). PRINT and its automatic supplement, prePRINTS. Nucleic Acids Res 31: 400-402.

Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A and Wu C (2000). The protein information resource (PIR). Nucleic Acids Res 28: 41-44.

Baxevanis AD (2003). The Molecular Biology Database Collection: 2003 update. Nucleic Acids Res 31:1-12.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL (2004). GenBank: update. Nucleic Acids Res 32: D23-D26.

Bernal A, Ear U and Kyrpides N (2001). Genomes Online Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res 29: 126-127.

Bhasin M and Raghava GPS (2004a). Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279:23262-23266.

Bhasin M and Raghava GPS (2004b). ESLpred: SVM based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32:W414-W419.
Bhasin M and Raghava GPS (2004c). GPCRpred: An SVM based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res 32:W383-W389.
Bhasin M and Raghava GPS (2004d). SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinf 20: 421-423.
Bhasin M and Raghava GPS (2004e). Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine (in press).
Bhasin M, Singh H and Raghava GPS (2003). MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinf 19: 665-666.
Blundell TL, Sibanda BL, Sternberg MJ and Thornton JM (1987). Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326: 347-352.
Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E,Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370.
Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, and Karplus M (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 4:187-217.
Brusic V, Rudy G, Honeyman G, Hammer J and Harrison L (1998b). Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinf 14: 121-130.
Burge C and Karlin S (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78-94.
Bystroff C and Shao Y (2002). Fully automated *ab initio* protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinf 18:S54-S61.
Castelo AT, Martins W and Gao GR (2002). TROL-Tandem Repeat Occurrence Locator. Bioinf 18:634-636.
Chen J, Le S and Maize J (2000). Prediction of common secondary structures of RNAs: A genetic algorithm approach. Nucleic Acids Res 28:991-999.
Chou KC (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246-255.
Combet C, Jambon M, Deléage G and Geourjon C (2002). Geno3D an automated protein modeling web server. Bioinf 18: 213-214.
Cuff JA, Clamp ME, Siddiqui AS, Finlay M and Barton GJ (1998). JPred: a consensus secondary structure prediction server. Bioinf 14:892-893.
Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO and Alizadeh AA (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 31:219-223.
Dieterich, C., Wang, H., Rateitschak, K., Luz, H. and Vingron, M. (2003). CORG: a database for Comparative Regulatory Genomics. Nucleic Acids Res 31: 55-57.
Dunham A, Matthews LH, Burton J, Ashurst JL, Howe KL,Ashcroft KJ, Beare DM, Burford DC, Hunt SE, Griffiths-Jones S et al. (2004). The DNA sequence and analysis of human chromosome 13. Nature 428: 522-528.
Eisenhaber F and Bork P (1998). Wanted: subcellular localization of proteins based on sequence. Trends Cell Biol 8: 69-70.
Felsenstein J PHYLIP: Phylogeny Inference Package (unpublished).
Fichant GA and Burks C (1991). Identifying potential tRNA genes in genomic DNA sequences. J Mol Bio 220: 659-671.
Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR,Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.
FlyBase Consortium (2003). The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 31: 172-175.

Frishman D, Mokrejs M, Kosykh D, Karstenmuller G, Kolesov G, Zubrzycki I, Gruber C, Geier B, Kaps A, Volz A, Wagner C, Fellenberg M, Heumann K and Mewes HW (2003). The Pedant genome database. Nucleic Acids Res 31: 207-211.

Fuchs R (1990). Free molecular biological software available from the EMBL file server. Comput Appl Biosci 6:120-121.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD and Bairoch A (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis Nucleic Acids Res 31:3784-3788.

Gilbert D (2000). Free software in molecular biology for Macintosh and MS Windows computers. Methods Mol Biol 132:149-184.

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinf 19:1015-1018.

Goldsby RA, Kindt TJ and Osborne BA (2000). Kuby Immunology, WH Freeman and Company, 4th edition.

Grillo G, Licciulli F, Liuni S, Sbisa E and Pesole G (2003). PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. Nucleic Acids Res 31: 3608-3612

Guex N and Peitsch MC (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18:2714-2723.

Hall T BioEdit: Biological Sequence alignment editor for windows 95/98/NT (unpublished).

Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL and Brunak S (1998). NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconj J 15:115-130.

Havel TF and Snow ME (1991). A new method for building protein conformations from sequence alignments with homologs of known structure. J Mol Biol 217: 1-7.

Hein J (1990). Unified approach to alignment and phylogenies. Methods Enzymol 183: 626-645.

Hopp TP and Woods KR (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78: 3824-3828.

Horn F, Vriend G and Cohen FE (2001). Collecting and harvesting biological data: the GPCRDB and nuclear information systems. Nucleic Acids Res 29:346-349.

Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P and Bairoch A (2004). Recent improvements to the PROSITE database. Nucleic Acids Res 32:D134-D137.

Issac B and Raghava GPS (2002). GWFASTA: A server for FASTA search in eukaryotic and microbial genomes. Biotechniques 33: 548-556.

Issac B, Singh H, Kaur H and Raghava GPS (2002). Locating probable genes using Fourier transform. Bioinf 18: 196-7.

Issac B and Raghava GP (2004). EGPred: prediction of eukaryotic genes using ab initio methods after combining with sequence similarity approaches. Genome Res 14:1756-66.

Jaakkota T, Diekhans M and Haussler D (2000). A discriminative framework for detecting remote protein homologies. J Comput Biol 7: 95-114.

Jonassen I, Collins JF and Higgins DG (1995). Finding flexible patterns in unaligned protein sequences. Protein Sci 4: 1587-1595.

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H and Krogh A (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12: 1652-1662.

Karchin R, Karplus K and Haussler D (2002). Classifying G-protein coupled receptors with support vector machines. Bioinf 18:147-159.

Kaur H and Raghava GPS (2003a). Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 12:627-634.

Kaur H and Raghava GPS (2003b). A neural network based method for prediction of gamma-turns in proteins from multiple sequence alignment Protein Sci 12:923-929.

Kaur H and Raghava GPS (2004). Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins: Structure, Function, and Bioinformatics 55:83-90.

Kogelnik AM, Lott MT, Brown MD, Navathe SB and Wallace DC (1998). MITOMAP: a human mitochondrial genome database -- 1998 update. Nucleic Acids Res 26: 112-115.

Kolaskar AS and Tongaonkar PC (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett 276:172-174.

Krogh A (1997). Two methods for improving performance of an HMM and their application for gene finding. In Proc Fifth Int Conf on Intelligent Systems for Molecular Biology, ed. (Gaasterland, T. et al., Menlo Park, CA: AAAI Press), pp. 179-186.

Kuiken CL, Foley B, Hahn B, Korber B, Marx PA, McCutchan F, Mellors JW and Wolinksy S (2001). HIV sequence compendium eds. (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 02-2877).

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J and Giegerich R (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633-4642.

Lander ES, Linton LM, Birren B, et al. (2001). Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Lefebvre A, Lecroq T, Dauchel H and Alexandre J (2003). FORRepeats: detects repeats on entire chromosomes and between genomes. Bioinf 19:319-326.

Lu X and Olson WK (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 31: 5108-5121.

Majoros WH, Pertea M, Antonescu C and Salzberg SL (2003). GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic gene finders. Nucleic Acids Res 31: 3601-3604.

Masignani V, Rappuoli R, and Pizza M (2002). Reverse vaccinology: a genome-based approach for vaccine development. Expert Opin Biol The. 2:895-905.

Matzura O and Wennborg A (1996). RNAdraw: an integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. CABIOS 12: 247-249.

Maxam AM and Gilbert W (1977). A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560-564.

McGuffin LJ, Bryson K, and Jones DT (2000). The PSIPRED protein structure prediction server. Bioinf 16: 404-405.

Miyazaki S, Sugawara H, Ikeo K, Gojobori T and Tateno Y (2004). DDBJ in the stream of various biological data. Nucleic Acids Res 32: D31-D34.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R and Zdobnov EM (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31: 315-318.

Nakai K and Kanehisa M (1991). Expert system for predicting protein localization sites in gram-negative bacteria. Proteins 11: 95-110.

Nielsen H, Engelbrecht J, Brunak S and von Heijne G (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1-6.

Odorico M and Pellequer JL (2003). BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. J Mol Recognit 16:20-22.

Orengo CA, Brown NP and Taylor WR (1992). Fast structure alignment for protein databank searching. Proteins 14: 139-167.

Pascarella S and Argos P (1992). A data bank merging related protein structures and sequences. Protein Eng 5:121-137.

Pearlman DA, Case DA, Caldwell JW, Ross WR, Cheatham TE III, DeBolt S, Ferguson D, Seibel G and Kollman P (1995). AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. Comp Phys Commun 91: 1-41.

Pearson WR and DJ Lipman (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444-2448.

Peitsch M, Schwede T, Guex N and Peitsch MC (1995). Protein modeling by e-mail. Bio/Technol 13: 658-660.

Perumal K, Gu J, Chen Y and Reddy R (1999). Small RNA database compiled by the Department of Pharmacology, Baylor College of Medicine. Ed . (C Burks, Molecular Biology Database List, Nucleic Acids Res) pp.1-9.

Puntervoll P, Linding R, Gemünd C. Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A, Ferrè F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Küster B, Helmer-Citterich M, Hunter WN, Aasland R and Gibson TJ (2003). ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31:3625-3630.

Quackenbush, J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29:159-164.

Raghava GPS (1995). DNAOPT: A computer program to aid optimization of gel conditions of DNA gel electrophoresis and SDS-PAGE. Biotechniques 18: 274-281.

Raghava GPS (1994). Improved estimation of DNA fragment lengths from gel electrophoresis. Biotechniques 17: 100-104.

Raghava GPS (2001a). PDSB: public domain software in biology. Biotech Software & Internet Report 2:154-156.

Raghava GPS (2001b). PDWSB: public domain web servers in biology. Biotech Software & Internet Report 2:152-153.

Raghava GPS and Sahni G (1994). GMAP: a multipurpose computer program to aid synthetic gene design, cassette mutagenesis and introduction of potential restriction sites into DNA sequences. Biotechniques 16: 1116-1123.

Rammensee HG, Friede T and Stevanovic S (1995b). MHC ligands and peptide motifs: first listing. Immunogenetics 41: 178-228.

Rappuoli R (2000). Reverse vaccinology. Curr Opin Microbiol 3: 445-450.

Rice P, Longden I and Bleasby A (2000). EMBOSS: The European molecular biology open software suite. Trends Genet 16: 276-277.

Robinson-Rechavi M and Laude V (2003). Bioinformatics of nuclear receptors. Methods Enzymol 364: 95-118.

Rodriguez-Tome P (1998). The BioCatalog. Bioinf 14: 469-470.

Rosenblad MA, Gorodkin J, Knudsen B, Zwieb C and Samuelsson T (2003). SRPDB: signal recognition particle database. Nucleic Acids Res 31: 363-364.

Rost B (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 266:525-539.

Rozen S and Skaletsky HJ (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S and Misener S ed. (*Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ) pp 365-386.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A and Barrell B (2000). ARTEMIS: sequence visualization and annotation. Bioinf 16: 944-945.

Sadowski MI and Parish JH (2003). Automated generation and refinement of protein signatures: case study with G-protein coupled receptors. Bioinf 19: 727-734.

Sali A and Blundell TL (1993). Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234:779-815.

Sali A, Overington JP, Johnson MS, Blundell TL(1990). From comparisons of protein sequences and structures to protein modeling and design. Trends Biochem Sci 15: 235-240.

Sanger F, Nicklen S and Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 74:5463-5467.

Schneider TD, Stormo GD, Haemer JS and Gold L (1982). A design for computer nucleic-acid sequence storage, retrieval and manipulation. Nucleic Acids Res 10:3013-3024.

Schueler-Furman O, Altuvia Y, Sette A and Margalit H (2000). Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. Protein Sci 9:1838-1846.

Schultz J, Milpetz F, Bork P and Ponting CP (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA. 95: 5857-5864.

Sharma D, Issac B, Raghava GP, Ramaswamy R (2004). Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinf 20:1405-1412.

Shimko N, Liu L, Lang BF and Burger G (2001). GOBASE: the organelle genome database. Nucleic Acids Res 29: 128-132.

Singh H and Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics* **17**, 1236-7 .

Smit,A.F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin.*

Snel, B., Lehmann, G., Bork, P. and Huynen, M.A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighborhood of a gene. Nucleic Acids Res. 28, 3442-4.

Somorjai RL. (1990). Theories and simulation of protein folding. Biotechnology. 14, 1-19.

Sonnhammer ELL and Durbin R (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:, GC1-GC10.

Sprinzl M, Horn C, Brown M, Ioudovitch A and Steinberg S (1998). Compilation of tRNA sequences and sequences of tRNA genes. Nucl Acids Res 26:148-153.

Stern LJ, Brown JH, Jardefzky TS, Gogra JC, Urban RG, Strominger JL and Wiley DC (1994). Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 368: 215-221.

Stoehr PJ and Omond RA (1989). The EMBL network file server. Nucleic Acids Res 17: 6763-6764.

Strachan T and Read AP (1999). *Human Molecular Genetics* 2nd Edition. John Wiley

Sutcliffe MJ, Hayes FR and Blundell TL (1987). Knowledge based modeling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. Protein Eng 5: 385-392.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV,Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA, (2003). The COG database: an updated version includes eukaryotes. BMC Bioinf 4: 41.

Tatusov RL, Koonin EV and Lipman DJ (1997). A genomic perspective on protein families. Science 278:631-637.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

Van de Peer Y (1994). A new version (3.0) of TREECON .Department of Biochemistry, University of Antwerp (UIA) Universiteitsplein 1B-2610 .

Van Gunsteren WF and Berendsen HJC (1990). Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry angew. Chem Int Ed Engl 29: 992-1023.

Watts C and Powis S (1999). Pathways of antigen processing and presentation. Rev Immunogenet 1: 60-74.

Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC,Leontis N, Westbrook J, Westhof E, Zuker M and Major F (2000). RNAML: a standard syntax for exchanging RNA information. RNA 8: 707-717.

Wuyts J, Perriere G and Van de Peer Y (2004). The European ribosomal RNA database. Nucleic Acids Res 32: D101-D103.