DOI 10.1002/mnfr.200500273 Mol. Nutr. Food Res. 2006, *50*, 610 – 619

Review

Introduction to bioinformatics

Pinar Kondu Akalın

Iontek, Meridyen Is Merkezi Ali Riza Gurcan Cad. Cirpici Yolu, Istanbul 34010, Turkey

This article introduces the field of bioinformatics and describes bioinformatic approaches and their application to the study of protein allergens. The predominant bioinformatics tools and resources are listed and discussed.

Keywords: Bioinformatics / Databases / Genome / Protein / Proteome

Received: June 15, 2005; revised: February 17, 2006; accepted: April 8, 2006

1 Introduction

1.1 Description

Bioinformatics is predominantly a discipline for handling genetic information. Bioinformatics involves research, development, or application of computational tools and approaches to acquire, store, visualize, and interpret medical or biological data.

Easy access to bioinformatics tools and efficient analyses of bioinformatics data are important factors for integrating distributed research findings and for suggesting new hypotheses. Needless to say, the contribution of the Internet for this integration is immense, enabling data collection and data access on a global scale.

1.2 History

Although the term "bioinformatics" has been part of the scientific lexicon for only two decades or so, when the origins of many bioinformatics concepts are carefully investigated, we learn that the underlying principles of bioinformatics come from early studies that combined computational and experimental approaches in molecular and cell biology (Tables 1 and 2) [1].

Correspondence: Dr. Pinar Kondu Akalin, Iontek, Meridyen Is Merkezi Ali Riza Gurcan Cad. Cirpici Yolu No. 1/410, Merter, Istanbul 34010, Turkey
E-mail: pkondu@iontek.com.tr
Fax: +90-212-481-5516

Abbreviations: FAO/WHO, Food and Agriculture Organization/ World Health Organization; **HSP,** high-scoring segment pair; **NCBI,** National Center for Biotechnology Information; **RNAi**, RNA interference; **SDAP**, Structural Database of Allergenic Proteins

Table 1. Ten institutions that pioneered and fostered computation in biology [1]

Institutions	Country
Birkbeck College, University of London	UK
Boston University	USA
European Molecular Biology Laboratory (EMBL)	DE and EMBL
Institute of Protein Research, Academy of Sciences	Former USSR
Laboratory of Molecular Biology (LMB), MRC	UK
Los Alamos National Laboratory (LANL)	USA
National Biomedical Research Foundation (NBRF)	USA
Stanford University	USA
University of California San Francisco (UCSF)	USA
University College, University of London (UCL)	UK

With the advancement of powerful methodologies in molecular, structural, and chemical biology, such as genome-scale sequencing, microarray gene expression analysis, RNA interference (RNAi), high-throughput crystallization and more, all of which produce massive amounts of biological data, it has become obvious that "biological information processing" is as important as data production, and merits becoming a separate discipline with its own customs.

1.3 Funding

National health organizations in most developed nations allocate a good share of their public research funding to bioinformatics, often establishing self-contained bioinformatics institutions (Table 3) (Frith, M., Worldwide Bioinformatics Centers, http://zlab.bu.edu/~mfrith/BioinfoCenters.html).

The well-known sequence collaborations between NCBI/ GenBank-GenPept in the USA [2], European Bioinfor-

Table 2. Early publications that influenced our view of bioinformatics [1]

Publication	Comments
Zuckerkandl and Pauling, 1965	First use of molecular sequences for evolutionary studies
Fitch and Margoliash, 1967	Use of molecular sequences to build trees
Needleman and Wunsch, 1970	First implementation of dynamic programming for protein sequence comparison
Lee and Richards, 1971	Calculation of accessibility on protein structures
Chou and Fasman, 1974	First secondary structure prediction method
Tanaka and Scheraga, 1975	Simulation of protein folding
Dayhoff, 1978	First collection of protein sequences
Hagler and Honig, 1978	One of the first explicit attempts to simulate protein folding
Doolittle, 1981	Seminal paper examining divergence and convergence in protein evolution
Felsenstein, 1981	One of the first statistical treatments of evolutionary tree construction
Richardson, 1981	The most comprehensive description of protein structure to that date
Kabsch and Sander, 1984	Discovery with profound implications for model building by homology and structure prediction
Novotny *et al*., 1984	The inability of distinguishing correct from incorrect structures threw back structure prediction approaches for a long while
Chothia and Lesk, 1986	Examination of divergence between sequence and structure
Doolittle, 1986	Influential book on sequence analysis
Feng and Doolittle, 1987	The first approach for an efficient multiple sequence alignment procedure, later implemented in CLUSTAL
Lathrop *et al*., 1987	One of the first applications of Artificial Intelligence in protein structure analysis and prediction
Ponder and Richards, 1987	The very first threading approach, using sequence enumeration
Altschul *et al*., 1990	The implementation of a sequence matching algorithm based on Karlin's statistical work
Bowie *et al*., 1991	The first implementation of protein structure prediction using threading

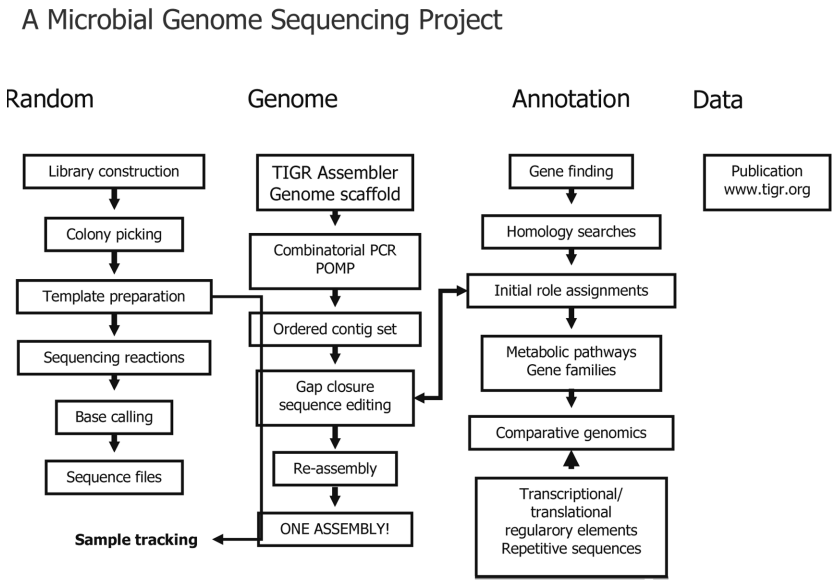


Figure 1. Steps of a microbial genome sequencing project at The Institute of Genetic Research (by Steven Salzberg).

matics Institute/European Molecular Biology Laboratory in Europe [3], and the National Institute of Genetics/DNA Database of Japan in Japan [4] function as invaluable public resources and facilitate access to up-to-date annotated DNA and protein data on three continents.

Main areas of interest in bioinformatics include: genome annotation, particularly gene structure prediction and genome comparison (genomics and comparative genomics), expression data analysis (functional genomics), protein structure prediction (structural biology), and biological networks (systems biology).

2 Genome sequencing, annotation and comparison

During a genome sequencing project (Fig. 1) [5], computational tools are used at different stages, from calling bases in the raw sequences to contig assembly and gap filling.

Table 3. Worldwide Bioinformatics Centers [2]

North America
- Portal to bioinformatics activities in Canada
- University of Washington Department of Genome Sciences
- LBNL Genome Sciences Department
- Bioinformatics at Stanford University
- UCSC Center for Biomolecular Science & Engineering
- UCSD Bioinformatics Program
- Washington U., St. Louis Center for Computational Biology
- Whitehead Institute/MIT Center for Genome Research
- Boston University Bioinformatics
- Harvard-Lipper Center for Computational Genetics
- Bioinformatics at Rensselaer and the Wadsworth Center
- Columbia Genome Center
- Genome Resources at Cold Spring Harbor Laboratory
- Penn. State Center for Comparative Genomics and Bioinformatics
- U. Pennsylvania Center for Bioinformatics
- The Institute for Genomic Research
- National Center for Biotechnology Information
- Nitrogen Fixation Research Center, National Autonomous University of Mexico

Australia/Oceania
- U. Queensland Institute for Molecular Bioscience

Africa
- South African National Bioinformatics Institute

Asia
- Weizmann Institute Bioinformatics & Biological Computing
- Institute of Cytology and Genetics, Siberia
- University of Pune Bioinformatics Center
- National University of Singapore Bioinformatics Centre
- Genome Institute of Singapore
- Keio University Institute for Advanced Biosciences
- University of Tokyo Human Genome Center
- Computational Biology Research Center, Tokyo
- RIKEN Genomic Sciences Center, Yokohama
- Center for Information Biology and DNA Data Bank of Japan
- Kyoto University Bioinformatics Center

Europe
- Virtual Institute of Bioinformatics (Éire)
- European Bioinformatics Institute
- Sanger Institute
- Research Group on Biomedical Informatics, IMIM, Barcelona
- Karolinska Institute Center for Genomics and Bioinformatics
- Technical University of Denmark Center for Biological Sequence Analysis
- Helix Group, Lyon
- Structural and Genetic Information Laboratory, Marseille
- Bielefeld University Center for Biotechnology
- EMBL Heidelberg
- GSF, München
- Max Planck Institute for Molecular Genetics
- Swiss Institute of Bioinformatics

When assembling a new genome for a species closely related to another species with an already completed genome, it is common to use comparative genomics, mapping a newly sequenced genome onto a reference genome [6]. After the primary sequence has been determined, gene identification is performed using a gene-finding algorithm. Several gene-finding algorithms have been developed that work with varying degrees of success. Leading ones are GRAIL [7], GeneParser [8], Genie [9], FGENESH [10], VEIL [11], Glimmer [5], Twinscan [6], and Genscan [12]. These algorithms apply a variety of advanced computational techniques, including neural networks, dynamic programming, decision trees, probabilistic reasoning, and Hidden Markov Models.

Once initial gene identification is over (though rarely completed, because gene finding is enormously difficult in eukaryotes), it is necessary to assign functional roles to those genes and to delineate the non-coding genetic elements. A variety of methods are used, usually in combination, looking at the translated nucleotide sequences, identifying protein domains, comparing the unknown genome to a previously characterized one, measuring sequence similarity to existing cDNA sequences in the nucleotide databases, and determining synteny.

Numerous genomes have been sequenced in the last decade, including the model organisms budding yeast *Saccharomyces cerevisiae*, fruit fly *Drosophila melanogaster,* fission yeast *Schizosaccharomyces pombe*, worm *Caenorhabditis elegans*, plant *Arabidopsis thaliana*, mouse *Mus musculus,* and rat *Rattus norvegicus*, many viral and microbial organisms, human *H. sapiens*, and most recently chicken *Gallus gallus*.

Visualizing the chromosomal locations of genes, and the annotations contributed by scientists all over the world, is made possible by genome browsers. Two notable ones are the University of California, Santa Cruz (UCSC Genome Browser, http://genome.ucsc.edu/) and Ensemble (Ensemble Genome Browser, http://www.ensembl.org/) genome browsers. The sites hosting the genome browsers offer up-to-date biological information, ranging from chromosomal coordinates to expressed sequence tag data and more, integrating many disparate types of information about the completed genomes.

Annotation transfer between genomes, based on protein similarity, gene orthologs, interologs, and regulogs, is also commonly used to predict the function of experimentally uncharacterized proteins. Figure 2 shows use of the BioKnowledge transfer method to identify a conserved protein and to annotate the proteome of a newly sequenced genome [13]. In 2000, when the draft genome sequencing project of worm was completed, analysis of the worm proteome, which had a large fraction of uncharacterized genes at the time, benefited from the annotation transfer method based on sequence similarity to heavily studied yeast genes. Later findings, accelerated by the discovery of RNAi for gene

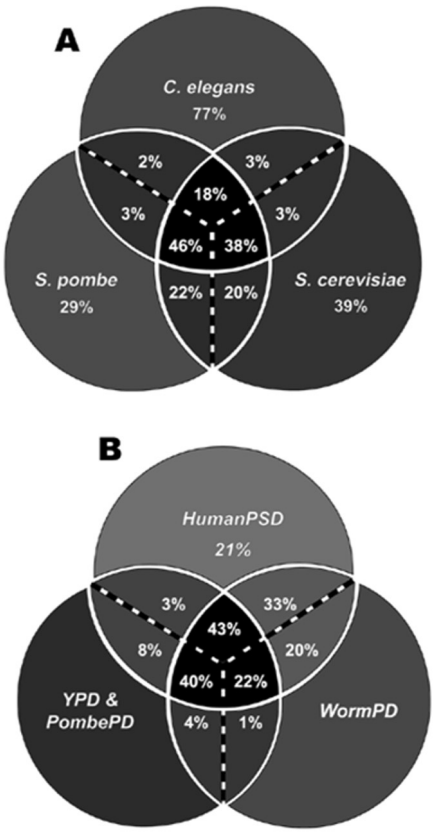


Figure 2. Venn diagrams of proteome relationships for model organisms. (A) Proportion of proteins conserved among the model organisms. (B) Comparison of predicted proteins of the model yeasts (*S. cerevisiae*, 6145 proteins and *S. pombe*, 4837 proteins), *C. elegans* (18546 proteins) and mammals (10229 from human, 5966 from mouse and 3188 from rat). In the intersection of all three circles (colored in black) are proteins that find a match in all three species, represented as the percentage of total proteins in that species, as represented in the indicated volumes of the BioKnowledge Library as of September 2000 [13].

knock-down analysis, showed that many of those predictions were correct, indicating that refined knowledge transfer actually leads to realistic predictions. A similar approach was used to annotate a select number of human genes, contributing to the initial RefSeq human database of NCBI [14].

"Nothing in biology makes sense except in the light of evolution!" [15]. By comparing the genes of related species, a tremendous amount can be learned about how genomes are organized and how major evolutionary changes take place (Fig. 3) (Ponting, C., *The Human Genome – Mammalian Genes I: Conservation and Slow Evolution,* http://www.stats.ox.ac.uk/~hein/HumanGenome/).

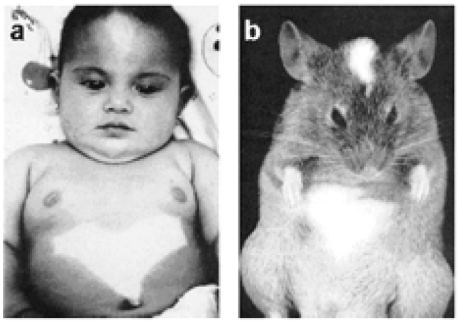


Figure 3. Human and mouse c-kit mutations show similar phenotypes. The utility of mouse as a biomedical model for human disease is enhanced when mutations in orthologous genes give similar phenotypes in both organisms. In a visually striking example of this, the same pattern of hypo-pigmentation is seen in (a) a patient with the piebald trait and (b) a mouse with dominant spotting, both resulting from heterozygous mutations of the c-*kit* proto-oncogene (by Chris Ponting; Ponting, C., *The Human Genome – Mammalian Genes I: Conservation and Slow Evolution,* http://www.stats.ox.ac.uk/~hein/HumanGenome/).

All these studies have yielded significant insights into how genes function, how cellular processes work, and how species have evolved.

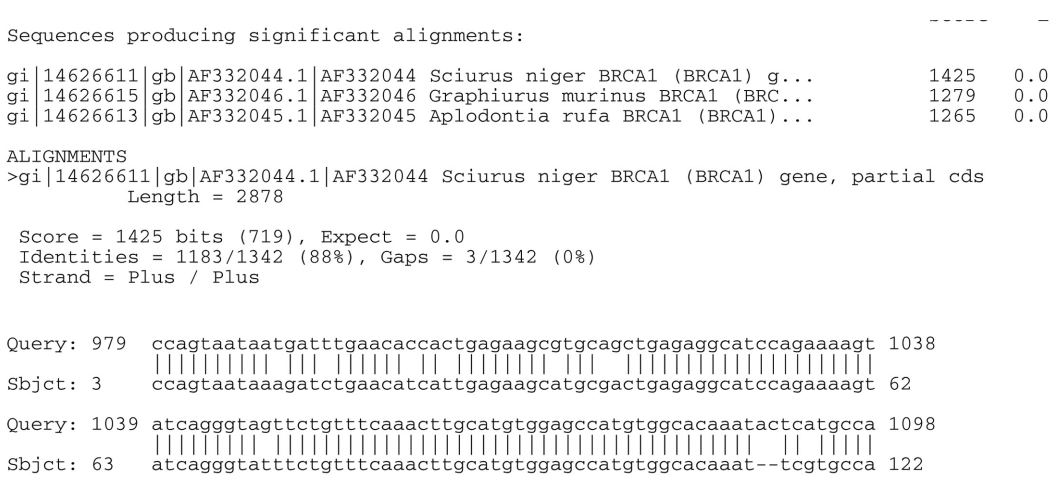
3 Sequence analysis

Large-scale DNA sequencing efforts for genome projects, which are producing an immense amount of sequence data, necessitated sensitive and robust methods for data analysis and evaluation. Since only a fraction of the genes in any of the large genomes such as worm, human, mouse, rat, and chicken have been studied experimentally, computational inference of biological information is required to stimulate thought and to increase understanding of these genomes. Many of the most powerful sequence analysis methods are based on probabilistic modeling and comparing sequences within and across species. Examples include the use of probabilistically derived score matrices to determine the significance of sequence alignments, the use of Hidden Markov Models as the basis for profile searches to identify distant members of sequence families, and the inference of phylogenetic trees using maximum likelihood approaches. Well-known heuristic algorithms BLAST and FASTA are used widely where high similarity and regional matches are the expected outcome.

The basic aim of a molecular biologist is to find out whether a newly identified sequence has any similarity to other sequences in the target databases. The BLAST package [16], the most popular one, is used both for nucleic and amino acid sequences, and provides alignments of short

Table 4. BLAST program selection for nucleotide and peptide sequence queries (BLAST Program Selection Guide, by blast-help group, NCBI User Service. http://www.ncbi.nlm.nih.gov/BLAST/producttable.shtml#blastp#blastp)

Length	Database	Purpose	Program
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	discontiguous megablast, megablast, or blastn
		Find sequences similar to query sequence	discontiguous megablast or blastn
		Find similar sequence from the Trace archive	Trace megablast, or Trace discontiguous megablast
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)
7 – 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches
5 – 15 Residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches
15 Residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST
		Find proteins similar to the query around a given pattern	PHI-BLAST
		Find conserved domains in the query	CD-search (RPS-BLAST)
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)



```
                                                                          -----    -
Sequences producing significant alignments:

gi|14626611|gb|AF332044.1|AF332044 Sciurus niger BRCA1 (BRCA1) g...          1425   0.0
gi|14626615|gb|AF332046.1|AF332046 Graphiurus murinus BRCA1 (BRC...          1279   0.0
gi|14626613|gb|AF332045.1|AF332045 Aplodontia rufa BRCA1 (BRCA1)...          1265   0.0

ALIGNMENTS
>gi|14626611|gb|AF332044.1|AF332044 Sciurus niger BRCA1 (BRCA1) gene, partial cds
          Length = 2878

 Score = 1425 bits (719), Expect = 0.0
 Identities = 1183/1342 (88%), Gaps = 3/1342 (0%)
 Strand = Plus / Plus


Query: 979  ccagtaataatgatttgaacaccactgagaagcgtgcagctgagaggcatccagaaaagt 1038
            |||||||||| ||| |||||| || |||||||| |||  ||||||||||||||||||||
Sbjct: 3    ccagtaataaagatctgaacatcattgagaagcatgcgactgagaggcatccagaaaagt 62

Query: 1039 atcagggtagttctgtttcaaacttgcatgtggagccatgtggcacaaatactcatgcca 1098
            ||||||||| ||||||||||||||||||||||||||||||||||||  || |||||
Sbjct: 63   atcagggtatttctgtttcaaacttgcatgtggagccatgtggcacaaat--tcgtgcca 122
```

Figure 4. The top part of a BLAST output showing an HSP.

stretch of identities and with high scoring matches at reasonable speed. A list of BLAST programs and their intended usage is shown in Table 4, taken from the NCBI's BLAST website (BLAST Program Selection Guide, By blast-help group, NCBI User Service, http://www.ncbi.nlm.nih.gov/BLAST/producttable.-shtml#blastp#blastp).

The fundamental unit of the BLAST algorithm output is the high-scoring segment pair (HSP). An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cutoff score (Fig. 4).

In the BLAST output shown in Fig. 4, the sequences producing significant alignments are listed in descending order of their Score (S), which represents the sum of values for each aligned pair of residues plus values for each gap. The (E) value, the Expect value, is a parameter that describes the number of hits one can "expect" to see just by chance

when searching a database of a particular size. It decreases exponentially as the Score (S) that is assigned to a match between two sequences increases. Essentially, the E value describes the random background noise that exists for matches between sequences. The closer the E value is to "0," the more "significant" the match.

In the alignment section, matching sequences are connected with a "|" symbol, and mismatches are connected with a space. A gap is represented with the symbol "–".

4 Expression data analysis

Microarray technology has made possible studies at a genome-wide scale, contributing to the functional annotation of genomes and shedding light on underlying mechanisms of many human diseases. The basic technology starts with probes, which may be cDNA (500 – 5000 bases long), oligonucleotides (20 – 80'mer oligonucleotides) or peptide nucleic acid. Probes are complementary to genes of interest, and are immobilized in microscopic quantities on a solid surface, typically a glass slide, in a spatially defined array. When fluorescently labeled DNA from tissue or cell samples is washed over the surface, complementary DNA binds. The bound DNA is detected by fluorescence following laser excitation. Probes can be laid down by spotting (Stanford technology) or photolithography (Affymetrix technology). Gene expression monitoring and single nucleotide polymorphism detection are the most common applications of microarrays.

A number of strategies (Leming, S., *DNA Microarray (Genome Chip) – Monitoring the Genome on a Chip Web site*, http://www.gene-chips.com/) can be used at each step in the design and implementation of a DNA microarray experiment (Fig. 5).

From probe design to analysis of gene expression there is a good variety of both public and commercial software and on-line tools available. An introductory list of these tools can be found on the Internet (Leung, Y. F., *My Microarray Software Comparison*, http://ihome.cuhk.edu.hk/~b400559/arraysoft.html; SMD, Microarray Resources, http://genome-www5.stanford.edu/resources/), and some are also listed in Table 5.

The Microarray Gene Expression Database consortium (MGED, Microarray Gene Expression Data Society, http://www.mged.org/), supported by the microarray community, sets standards for experimental controls and for data normalization in microarray experiments, so as to ensure the interpretability of the results and to allow potential verification by third parties.

5 Protein structure prediction

Structural information about a protein is perhaps the most useful data for understanding protein function. Experimentally determined 3-D structures of proteins are deposited in the Protein Data Bank [17]. For the majority of proteins whose structures have not been determined experimentally, computational structure prediction methods can provide valuable information.

One class of protein structure prediction methods, which includes threading and comparative modeling, relies on detecting similarity between a modeled sequence and at least one known structure. A second class of methods, the *de novo* or *ab initio* methods, predict the structure from protein sequence alone without relying on similarity between the modeled sequence and any known structure [18]. The *ab initio* prediction methods first model the energetics of protein folding, and then try to find the structure with lowest thermodynamic state. This approach utilizes the thermodynamic hypothesis, which states that the native structure of a protein is the one for which the free energy achieves a global minimum. A brief list of protein structure prediction methods (CMGM, Bioinformatics Resources, http://cmgm.stanford.edu/) is given in Table 6.

1) **Probe** (cDNA/oligo with known identity)	2) **Chip fabrication** (Putting probes on the chip)	3) **Target** (fluorescently labeled sample)	4) **Assay**	5) **Readout**	6) **Informatics**
Small oligos, cDNAs, chromosome, ... (whole organism on a chip?)	Photolithography, pipette, drop-touch, piezoelectric (ink-jet), electric, ...	RNA, (mRNA==>) cDNA	Hybridization, long, short, ligase, base addition, electric, MS, electrophoresis, fluocytometry, PCR-DIRECT, TaqMan, ...	Fluorescence, probeless (conductance, MS, electrophoresis), electronic, ...	Robotics control, Image processing, DBMS, WWW, bioinformatics, data mining and visualization

Figure 5. Steps in the design and implementation of a DNA microarray experiment (by Leming Shi; Leming, S., *DNA Microarray (Genome Chip) – Monitoring the Genome on a Chip Web site,* http://www.gene-chips.com/).

Table 5. Commonly used microarray software and tools (SMD: Microarray Resources, http://genome-www5.stanford.edu/resources/)

Program	Description	Provider	Platform
Array Designer	Tool assisting in primer design for microarray construction	Premier Biosoft International	JAVA
ArrayMiner®	Set of analysis tools using advanced algorithms to reveal the true structure of gene expression data.	Optimal Design, Sprl.	Windows MacOS
ArrayViewer	Identification of statistically significant hybridization signals	National Human Genome Research Institute	JAVA
ArrayVision	Automated analysis of macro & micro-arrays	Imaging Research Inc.	Windows
BAGEL	Bayesian Analysis of Gene Expression Levels: a program for the statistical analysis of spotted microarray data.	University of Connecticut	MacOS, Windows, Linux
BAMarray	Bayesian Analysis of Variance for Microarrays – detect differentially expressed genes from multigroup micro-array data	Case University	Windows, Linux
BASE	Microarray database and analysis platform	Lund University	Web
Cluster 3.0	An enhanced version of Mike Eisen's Cluster	University of Tokyo, Japan	UNIX Linux MacOS Windows
DNA arrays analysis tools	Hierarchical clustering, Self Organizing Maps and more	Spanish National Cancer Center, Bioinformatics Unit	Web
Expression Profiler	Analysis & clustering of gene expression data	European Bioinformatics Institute (EBI)	Web
Expression Sieve	Microarray data analysis package, strong in linking biological significance to expression patterns, data and analysis process management, signature gene discovery and class discovery & prediction.	BioSieve	Java
GEDA	Gene expression data analysis and simulation tools, offering a variety of options for processing and analyzing results.	University of Pittsburgh and UPMC	Web
Gene Cluster	Self-organizing maps	Whitehead Institute/MIT Center for Genome Research	JAVA Windows NT
Gen MAPP	Tools for visualizing data from gene expression experiments in the context of biological pathways.	Conklin lab; Gladstone Institute & the UCSF	Windows
Gene X	Gene Expression Database : integrated toolset for data analysis and comparison	National Center for Genome Resources	Windows Linux SunOS/Solaris
Gene Xplorer	Web-visualization of microarray datasets	Rees/Demeter/Matese Stanford	Perl
GenMaths	Analysis of high density microarrays and gene chips	Applied Maths	Windows
Java Treeview	View Results of Clustering	Alok Saldhana Stanford	Java
MicroArray Suite	Extracting and visualizing DNA micro-array data	Scanalytics Inc.	MacOS
Partek Pattern Recognition	Extracting and visualizing patterns in large multivariate data	Partek Incorporated	Linux, Unix, Windows
ScanAlyze	Processes fluorescent images of micro-arrays	Michael Eisen Stanford	Windows
TIGR Multi Experiment Viewer	Analysis and Visualization of Microarray Data	TIGR	JAVA
TreeArrange and Treeps	Software for displaying and manipulating hierarchical clustered data	University of Waterloo, Canada	Linux, Unix, Windows

Table 6. Protein structure prediction tools: threading, homology modeling, secondary structure determination and ab-initio methods (CMGM: Bioinformatics Resources. http://cmgm.stanford.edu/)

Name	Accuracy	Comments
Experimentally determined 3-D structures		
PDB http://www.rcsb.org/pdb/		Repository of crystal structures of macromolecules
Threading	PHDthreader: http://www.embl-heidelberg.de/predictprotein/	< 30%, less than 30% of the predicted first hits are true remote homologs. Evaluated by cross-validation on 89 unique protein structures. Prediction-based threading detecting the fold type and aligning a protein of unknown structure and a protein of known structure for low levels of sequence identity (< 25%).
T3P2: http://www.mbi.ucla.edu/people/frsvr/frsvr.html		Prediction-based threading detecting the fold type and aligning a protein of unknown structure and a protein of known structure for low levels of sequence identity (< 25%).
Homology-Modeling		
SWISS-MODEL: http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html		An automated knowledge-based protein modeling server; first approach and optimize (Peitsch M.C. Protein Modeling by E-mail. Bio/Technology 13:658 – 660. (1995)
Secondary Structure Prediction		
PSIPRED http://bioinf.cs.ucl.ac.uk/psipred/psiform.html	> 78%	Multiple alignment-based two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST
PHDsec: http://www.embl-heidelberg.de/predictprotein/	> 72% (+/– 10%, one standard deviation)	Multiple alignment-based neural network system.
NSSP: http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html	> 71%. Evaluated on > 200 unique proteins.	Multiple alignment-based nearest-neighbor method.
SOPM: http://www.ibcp.fr/predict.html	> 70%.	Multiple alignment-based method combining various other prediction programs.
DSC: http://bonsai.lif.icnet.uk/bmm/dsc/dsc_read_align.html	70%	Multiple alignment-based program using statistics.
SSPRED: http://www.embl-heidelberg.de/sspred/ssp_mul.html	> 70%.	Multiple alignment-based program using statistics.
MultiPredict: http://kestrel.ludwig.ucl.ac.uk/zpred.html	> 65%	Multiple alignment-based method using physicochemical information from a set of aligned sequences and statistical secondary structure decision constants.
PSA: http://bmerc-www.bu.edu/psa/		The PSA server analyzes amino acid sequences to predict secondary structures and folding classes.
NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html	> 65%	Single-sequence based neural network prediction.
Ab-Initio Methods		
POSETTA Package http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/Rosetta/		Monte Carlo simulated annealing procedure

6 Modeling biological networks – systems biology

With so much information in the literature about DNA, RNA, protein, interactions, gene expression, regulatory elements, metabolic pathways, integration of these diverse types of data represents a huge challenge. The interrelationships of the data sets must be determined, in order to begin to model the cellular system as a whole. Systems biology is an emerging field that aims to generate a model of a cell or organism using integrated analytical and computational tools. An increasing number of research groups are using systems biology approaches, including dedicated institutes such as the Institute for Systems Biology (http://www.systemsbiology.org/).

7 Bioinformatics applied to protein allergenicity

Allergy is one of the most common chronic health problems worldwide, and methods that provide more than sympto-

matic relief for allergic conditions are being sought. The introduction of recombinant proteins into food and other products has raised public awareness about allergy in recent years. Bioinformatics analyses of genomic, functional and structural data can provide significant insight into relationships between allergens, and these analyses can be complemented by clinical observations from the medical literature.

Allergy databases have been created as a repository for a variety of information on allergens, including sequence, epitope, structure, and cross-reactivity. A review of allergen databases has been published recently [19] (Table 7). It is important that these databases provide reliable and up-to-date data in order for them to be useful for systematic bioinformatics analyses. These databases are also an invaluable resource for clinicians and researchers in this field.

With the aim of preventing the marketing of potentially IgE-inducing or cross-reactive food products, the Food and Agriculture Organization and World Health Organization (FAO/WHO) established a set of recommendations for assessing the allergenic potential of proteins, which were further refined by the Codex Alimentarius Commission (Food and Agriculture Organization of the United Nations, FAO Corporate Document Repository, http://www.fao.org/documents/show_cdr.asp?url_file=/docrep/007/y0820e/y0820e00.htm).

Among the initial screening recommendations is a sequence similarity comparison. A 35% sequence identity over a length of 80 amino acids, or a scientifically justified match of contiguous amino acids, is recommended to evaluate cross-reactivity. In addition to sequence similarity, the report warns against structural similarities to protein families that have allergenic members, such as napines, lipocalins, and others.

Brusic *et al*. [20], upon reviewing the computational tools for the study of allergens, suggested that searching for common structural motifs may improve the quality of assessments of cross-reactivity and allergenicity.

Li *et al.* [21] at the Bioinformatics Institute of Singapore used the wavelet approach for detecting proteins with high allergenicity potential to address exactly this suggestion.

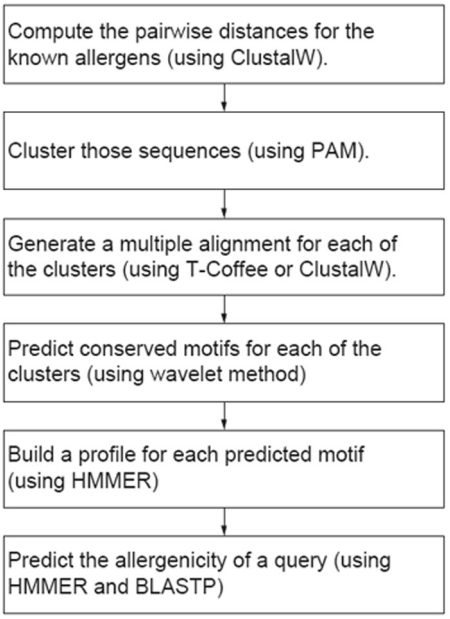


Figure 6. Wavelet transform method for prediction of allergenic proteins. Flowchart of the entire prediction system [21].

Here, training sequences are partitioned into clusters; then the sequences in each cluster are aligned and wavelet transformed into levels with different resolution. Conserved motifs are identified using multi-resolution plots. A higher recall was achieved using this method, despite decreased precision, demonstrating a potential for detecting cross-reactivity between distantly related allergens. Figure 6 shows the steps of the wavelet approach.

Ivanciuc *et al.* [22] at the University of Texas offer web-based computational tools for structural studies of allergens at the Structural Database of Allergenic Proteins (SDAP) website. These computational tools use an algorithm based on conserved properties of amino acid side chains to identify regions of known allergens similar to user-supplied peptides or IgE epitopes in the SDAP database. Amino acid descriptors, E_1–E_5, are used to locate sequences with similar chemical properties, and for two sequences A and B, each one consisting of N residues, to compute the property distance function PD (Eq. 1)

Table 7. Prominent allergen databases

Database	URL	Content highlights
IUIS	http://www.allergen.org	Nomenclature
Allergome [17]	http://www.allergome.org	Comprehensive information on allergens
SDAP [18]	http://fermi.utmb.edu/SDAP	Sequence alignment and conformation prediction tools
FARRP	http://www.allergenonline.com	Searchable by categories, concise information
PROTALL	http://www.ifr.bbsrc.ac.uk/Protall	Biochemical and clinical data
BIFS	http://www.iit.edu/~sgendel/fa.htm	Links to PIR, SP, Entrez
Swiss-Prot's Allergen Index	http://ca.expasy.org/cgi-bin/lists?allergen.txt	

$$PD(A,B) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{5}\lambda_j(E_j(A_i) - E_j(B_i))^2\right]^{1/2} \quad (1)$$

where λ_j is the eigenvalue of the *j*th *E* component, $E_j(A_i)$ is the E_j value for the amino acid in the *i*th position from sequence *A*, and $E_j(B)_i$ is the E_j value for the amino acid in the *i*th position from sequence *B*. The tool can be used to identify identical or more distantly related epitope sequences to assess cross-reactivity.

To date it appears that most of the bioinformatics methods used to study protein allergens have focused on sequence similarity between known allergens and IgE-binding epitopes. Since conformational specificity between IgE antibodies and epitopes is gaining more significance, it may be reasonable to investigate algorithms that create substitute sequences that preserve the conformation of epitopes with known 3-D structure; subsequently, searches can be conducted for potential allergenic protein sequences against a database containing those "conformational substitute stretches." In addition, while there is considerable interest in using structural motifs as a basis for evaluating novel proteins, further investigation is needed to determine the usefulness of structural motifs for evaluating protein allergenicity.

8 References

[1] Ouzounis, C. A., Valencia, A., *Bioinformatics* 2003, *19*, 2176 – 2190.

[2] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. *et al*., *Nucleic Acids Res*. 2002, *30*, 17 – 20.

[3] Stoesser, G., Baker, W., Van Den Broek, A., Camon, E. *et al*., *Nucleic Acids Res.* 2002, *30*, 21 – 26.

[4] Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K. *et al*., *Nucleic Acids Res.* 2002, *30*, 27 – 30.

[5] Delcher, A. L., Harmon, D., Kasif, S., White, O. *et al*., *Nucleic Acids Res.* 1999, *27*, 4636 – 4641.

[6] Korf, I., Flicek, P., Duan, D., Brent, M. R., *Bioinformatics* 2001, *17*, S140 – S148.

[7] Uberbacher, E., Mural, R., *Proc. Natl. Acad. Sci. USA* 1991, *88*, 11261 – 11265.

[8] Snyder, E. E., Stormo, G. D., *Nucleic Acids Res.* 1993, *21*, 607 – 614.

[9] Reese, M., Eeckman, F., Kulp, D., Haussler, D., *Improved splice site detection in genie. RECOMB 97*, 1997, 232 – 240.

[10] Solovyev, V., Salamov, A., Lawrence, C., *Nucleic Acids Res*. 1994, *22*, 5156 – 5163.

[11] Henderson, J., Salzberg, S., Fasman, K. H., *J. Comput. Biol.* 1997, *4*, 127 – 141.

[12] Burge, C. Karlin, S., *J. Mol. Biol.* 1997, *268*, 78 – 94.

[13] Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E. *et al*., *Nucleic Acids Res.* 2001, *29*, 75 – 79.

[14] Hodges, P. E., Carrico, P. M., Hogan, J. D., O'Neill, K. E. *et al*., *Nucleic Acids Res.* 2002, *30*, 137 – 141.

[15] Dobzhansky, T., *The American Biology Teacher* 1973, *35*, 125 – 129.

[16] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. *et al*., *J. Mol. Biol.* 1990, *215*, 403 – 410.

[17] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G. *et al*., *Nucleic Acids Res.* 2000, *28*, 235 – 242.

[18] Baker, D., Sali, A., *Science* 2001, *294*, 93 – 96.

[19] Brusic, V., Millot, M., Petrovsky, N., Gendel, S. M. *et al*., *Allergy* 2003, *58*, 1093 – 1100.

[20] Brusic, V., Petrovsky, N., Gendel, S. M., Millot, M. *et al*., *Allergy* 2003, *58*, 1083 – 1092.

[21] Li, K. B., Issac, P., Krishnan, A., *Bioinformatics* 2004, *20*, 2572 – 2578.

[22] Ivanciuc, O., Schein, C. H. Braun, W., *Nucleic Acids Res.* 2003, *31*, 359 – 362.

