

Database Mining in the Human Genome Initiative

John L. Houle,^a Wanda Cadigan,^b Sylvain Henry,^b Anu Pinnamaneni^b and Sonny Lundahl^c

^aScientific Author, ^bScientific Reviewer, ^cSenior Manager

Bio-databases.com, Amita Corporation, 1420 Blair Place, Suite 500, Ottawa, Ontario, Canada, K1J 9L8

info@bio-databases.com

Abstract

The Human Genome Initiative is an international research program for the creation of detailed genetic and physical maps of the human genome. Genome research projects generate enormous quantities of data. Database mining is the process of finding and extracting useful information from raw datasets. Computational genomics has identified a classification of three successive levels for the management and analysis of genetic data in scientific databases:

1. Genomics.
2. Gene expression.
3. Proteomics.

Genome database mining is the identification of the protein-encoding regions of a genome and the assignment of functions to these genes on the basis of sequence similarity homologies against other genes of known function. Gene expression database mining is the identification of intrinsic patterns and relationships in transcriptional expression data generated by large-scale gene expression experiments. Proteome database mining is the identification of intrinsic patterns and relationships in translational expression data generated by large-scale proteomics experiments. Improvements in genome, gene expression and proteome database mining algorithms will enable the prediction of protein function in the context of higher order processes such as the regulation of gene expression, metabolic pathways and signalling cascades. Thus, the final objective of such higher-level functional analysis will be the elucidation of high-resolution structural and functional maps of the human genome.

Contents

1. Human Genome Initiative 1
 2. Computational Molecular Biology and Scientific Databases 2
 3. Genomics 3
 1. Genome Databases 3
 2. Genome Database Mining 4
 1. Computational Gene Discovery 4
 2. Sequence Similarity Searching 5
 4. Gene Expression 7
 - 4.1. Gene Expression Databases 7
 - 4.2. Gene Expression Database Mining 8
 5. Proteomics 8
 1. Proteome Databases 9
 2. Proteome Database Mining 9
 6. Conclusions 10
- References 10

1. Human Genome Initiative

The Human Genome Initiative is an international research program for the creation of detailed genetic and physical maps for each of the twenty four different human chromosomes and the elucidation of the complete deoxyribonucleic acid (DNA) sequence of the human genome. A genetic map depicts the linear arrangement of genes or genetic marker sites along a chromosome. Two types of genetic maps are identified: genetic linkage maps and physical maps. Genetic linkage maps are based on the frequency with which genetic markers are coinherited. Physical maps determine actual distances between genes on a chromosome.

As described in the survey of Pearson and Soll,⁽¹⁾ the Human Genome Initiative has six scientific objectives:

1. Construction of a high-resolution genetic map of the human genome.
2. Production of a variety of physical maps of the human genome.
3. Determination of the complete sequence of human DNA.
4. Parallel analysis of the genomes of a selected number of well-characterized nonhuman model organisms.
5. Creation of instrumentation technologies to automate genetic mapping, physical mapping and DNA sequencing for the large-scale analysis of complete genomes.
6. Development of computational tools such as algorithms, software and databases for the collection, interpretation and dissemination of the vast quantities of complex mapping and sequencing data that are generated by human genome research.

Genetic maps serve as resources in the search for genes responsible for genetically-mediated diseases as well as for the further study of gene structure, function and expression. Thus, the advent of a high-resolution genetic map of the human genome will generate advances in six areas of medicine:

1. Genetic counseling.
2. Prediction of genetic disease susceptibility.
3. Diagnostic tests.
4. Gene therapy.
5. Rational drug design.
6. Pharmacogenomic drug customization.

The Human Genome Initiative has been reviewed.⁽²⁻¹⁵⁾

2. Computational Molecular Biology and Scientific Databases

Genome research projects generate enormous quantities of data. Genbank is the National Institutes of Health (NIH) molecular database which is composed of an annotated collection of all publicly available DNA sequences.⁽¹⁶⁾ The February 2000 release of the Genbank molecular database contained 5,691,000 DNA sequences which are further composed of approximately 5,805,000,000 deoxyribonucleotides.⁽¹⁷⁾

A major objective of the Human Genome Initiative is the development of more advanced DNA sequencing technologies. Concerted genome sequencing using these advanced DNA sequencing technologies will result in even further increases in DNA sequence generation rates. Genbank statistics on DNA sequence curation demonstrate exponential growth rates.⁽¹⁸⁾

Computational molecular biology is defined as the mathematical and computational analysis of biological macromolecules.⁽¹⁹⁻²¹⁾ Computational genomics refers to the applications of computational molecular biology in large-scale genome research.⁽²²⁻²⁸⁾ On the basis of the central dogma of molecular biology, computational genomics has identified a classification of three successive levels for the management and analysis of genetic data in scientific databases:

1. Genomics.
2. Gene expression.
3. Proteomics.

These application domains will be subsequently discussed. The objective of human genome database analysis is the

elucidation of structural and functional maps of the human genome. Database mining is defined as the process of finding and extracting useful information from raw datasets.⁽²⁹⁻³³⁾ Large-scale genome database mining is an open research problem that could be addressed by the application of supercomputer technologies. Thus, human genome mapping has been identified as a *Grand Challenge* problem in medical supercomputing.⁽³⁴⁻³⁶⁾

3. Genomics

Genomics is defined as the scientific discipline which focuses on the systematic investigation of genomes, i.e. the complete set of chromosomes and genes of an organism. Genomics consists of two component areas:

1. Structural genomics.
2. Functional genomics.

Structural genomics refers to the large-scale determination of DNA sequences and gene mapping. Functional genomics refers to the attachment of information concerning functional activity to existing structural knowledge about DNA sequences. As the determination of the DNA sequences comprising the human genome nears completion, the Human Genome Initiative is undergoing a paradigm shift from static structural genomics to dynamic functional genomics. The current section will focus on structural genomics. Genomics has been reviewed.^(8,11,26,37-45)

3.1. Genome Databases

As described in the survey of Pearson and Soll,⁽¹⁾ genome databases are used for the storage and analysis of genetic and physical maps. Chromosome genetic linkage maps represent distances between markers based on meiotic recombination frequencies. Chromosome physical maps represent distances between markers based on numbers of nucleotides.

Genome databases should define four data types:

1. Sequence.
2. Physical.
3. Genetic.
4. Bibliographic.

Sequence data should include annotated molecular sequences.

Physical data should include eight data fields:

1. Sequence-tagged sites.
2. Coding regions.
3. Noncoding regions.
4. Control regions.
5. Telomeres.
6. Centromeres.
7. Repeats.
8. Metaphase chromosome bands.

Genetic data should include seven data fields:

1. Locus name.
2. Location.
3. Recombination distance.
4. Polymorphisms.
5. Breakpoints.
6. Rearrangements.

7. Disease association.

Bibliographic references should cite primary scientific and medical literature.

Genome databases are classified into four categories based on their contents:

1. Molecular.
2. Genetic.
3. Organism.
4. Bibliographic.

Molecular databases include four representative implementations:

1. European Molecular Biology Laboratory Nucleotide Sequence Data Library (EMBL).⁽⁴⁶⁾ <http://www.embl-heidelberg.de/>
2. DNA Database of Japan (DDBJ).⁽⁴⁷⁾ <http://www.ddbj.nig.ac.jp/>
3. Genbank.⁽¹⁶⁾ <http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html>
4. Swiss-Prot.⁽⁴⁸⁾ <http://www.expasy.ch/sprot/sprot-top.html>

Genetic databases include two representative implementations:

1. Genome Database (GDB).⁽⁴⁹⁾ <http://gdbwww.gdb.org>
2. Online Mendelian Inheritance in Man (OMIM).⁽⁵⁰⁾ <http://www3.ncbi.nlm.nih.gov/Omim/>

Organism databases include three representative implementations:

1. Bacterium *Escherichia coli*.⁽⁵¹⁾
2. Mouse *Mus musculus*.⁽⁵²⁾
3. Mustard plant *Arabidopsis thaliana*.⁽⁵³⁾

Bibliographic databases include four representative implementations:

1. Biological Abstracts.
2. CancerLit.
3. Excerpta Medica (Embase).
4. Medline.

Genome databases have been reviewed.^(9,24,54-63)

3.2. Genome Database Mining

Genome database mining is an emerging technology. The process of genome database mining is referred to as computational genome annotation. Computational genome annotation is defined as the process by which an uncharacterized DNA sequence is documented by the location along the DNA sequence of all the genes that are involved in genome functionality. Computational genome annotation consists of two sequential processes:

1. Structural annotation.
2. Functional annotation.

Structural annotation refers to the identification of hypothetical genes termed open reading frames (ORFs) in a DNA sequence using computational gene discovery algorithms. Functional annotation refers to the assignment of functions to the predicted genes using sequence similarity searches against other genes of known function. Computational genome annotation has been reviewed.⁽⁶⁴⁻⁶⁶⁾

3.2.1. Computational Gene Discovery

Functionally-significant sites in DNA sequences have been studied and partially characterized using pattern recognition algorithms. DNA functional sites are sequences recognized and bound to by specific proteins, e.g. promoter elements. Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives). The identification of intron-exon boundaries and splice sites where ribonucleic acid (RNA) is transcribed from genomic DNA is of further importance. The ability to accurately predict introns would greatly facilitate the translation of genomic DNA into the amino acid sequence of the gene product. The comparative analysis of DNA sequences is an important technique in detecting biologically-significant relationships. Multiple sequence alignment is a useful technique in analyzing sequence-structure relationships. The DNA sequence of an unknown gene often exhibits structural homology with a known gene. Multiple sequence alignment is important for the recognition of patterns or motifs common to a set of functionally-related DNA sequences and is of assistance in structure prediction and molecular modeling. Multiple sequence alignment algorithms use variations of the dynamic programming method. Dynamic programming methods use an explicit measure of alignment quality, consisting of defined costs for aligned pairs of residues or residues with gaps and use an algorithm for finding an alignment with minimum total cost. Multiple sequence alignment has been reviewed.⁽⁶⁷⁻⁶⁸⁾

Computational gene discovery algorithms include twenty eight representative implementations:

1. Aat.⁽⁶⁹⁾ <http://genome.cs.mtu.edu/aat.html>
2. Banbury Cross. <http://igs-server.cnrs-mrs.fr/igs/banbury/>
3. EcoParse.⁽⁷⁰⁾ (Not available on the World-Wide Web.)
4. Fex.⁽⁷¹⁾ <http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html>
5. Gap 3.⁽⁷²⁾ (Not available on the World-Wide Web.)
6. GeneID.⁽⁷³⁾ <http://apolo.imim.es/geneid.html>
7. GeneMark.⁽⁷⁴⁾ <http://genemark.biology.gatech.edu/GeneMark/>
8. GeneModeler.⁽⁷⁵⁾ (Not available on the World-Wide Web.)
9. GeneParser.⁽⁷⁶⁾ <http://beagle.colorado.edu/~eesnyder/GeneParser.html>
10. GeneParser2.⁽⁷⁷⁾ (Not available on the World-Wide Web.)
11. GeneParser3.⁽⁷⁷⁾ (Not available on the World-Wide Web.)
12. Genie.⁽⁷⁸⁾ http://www.fruitfly.org/seq_tools/genie.html
13. GenLang.⁽⁷⁹⁾ http://www.cbil.upenn.edu/genlang/genlang_home.html
14. Genscan.⁽⁸⁰⁾ <http://ccr-081.mit.edu/GENSCAN.html>
15. GenViewer.⁽⁸¹⁾ <http://www.itba.mi.cnr.it/webgene/>
16. Glimmer.⁽⁸²⁾ <http://www.cs.jhu.edu/labs/compbio/glimmer.html#get>
17. Grail.⁽⁸³⁾ <http://compbio.ornl.gov/gallery.html>
18. Grail 2.⁽⁸⁴⁾ <http://compbio.ornl.gov/gallery.html>
19. Great.⁽⁸⁵⁾ (Not available on the World-Wide Web.)
20. Hexon / Fgeneh.⁽⁸⁶⁾ <http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html>
21. Morgan.⁽⁸⁷⁾ <http://www.cs.jhu.edu/labs/compbio/morgan.html>
22. Mzef.⁽⁸⁸⁾ <http://www.cshl.org/genefinder/>
23. ORFgene.⁽⁸⁹⁾ <http://www.itba.mi.cnr.it/webgene/>
24. Procrustes.⁽⁹⁰⁾ <http://www-hto.usc.edu/software/procrustes/index.html>
25. Sorfind.⁽⁹¹⁾ <http://www.rabbithutch.com>

26. Veil.⁽⁷⁰⁾ <http://www.cs.jhu.edu/labs/compbio/veil.html>
27. Xgrail.⁽⁷²⁾ <http://www.hgmp.embnet.org/Registered/Option/xgrail.html>
28. Xpound.⁽⁹²⁾ (Not available on the World-Wide Web.)

Computational gene discovery algorithms demonstrate limited performance accuracy in the prediction of eukaryotic genes.⁽⁹³⁾ Computational gene discovery algorithms have been reviewed.⁽⁹³⁻¹⁰⁶⁾

3.2.2. Sequence Similarity Searching

Sequence similarity searching is an important methodology in computational molecular biology. Initial clues to understanding the structure or function of a molecular sequence arise from homologies to other molecules that have been previously studied. Genome database searches reveal biologically-significant sequence relationships and suggest future investigation strategies. As described in the survey of Altschul et al.,⁽¹⁰⁷⁾ molecular sequence database homology is affected by five factors:

1. Algorithms.
2. Scoring systems.
3. Alignment statistics.
4. Database updates.
5. Database sequence bias.

Algorithms. Database search algorithms are based upon measures of local sequence similarity. Algorithms must balance the competing factors of speed, hardware requirements and sensitivity to biological relationships.

Scoring systems. Alignments are ranked by scores whose calculations are dependent upon the particular scoring systems used. The appropriate scoring system to use is largely dependant upon the problem under consideration.

Alignment statistics. Given a specific query, database search algorithms produce an ordered list of imperfectly-matched database similarities. An important question is defining the critical point of statistical significance.

Database updates. The use of a current comprehensive sequence database is essential to any similarity search.

Database sequence bias. There are biases in the molecules chosen to be included in molecular sequence databases.

Database search algorithms are used to compute pairwise comparisons between a candidate query sequence and each of the sequences stored within a database in order to find all the pairs of sequences that have a similarity above a defined threshold. There are three principal database search algorithms:

1. Smith-Waterman algorithm.
2. FASTA.
3. BLAST.

The Smith-Waterman algorithm uses dynamic programming to compute the most sensitive pairwise similarity alignments. However, these optimal computations require execution in order quadratic time.⁽¹⁰⁸⁾ The Smith-Waterman algorithm has been implemented. http://decypher2.stanford.edu/algo-sw/SW_nn.html-ssi

The FASTA algorithm is an approximate heuristic algorithm used to compute suboptimal pairwise similarity comparisons. Dynamic programming is used to compute a series of subsequence alignments called *hotspots* which are combined to approximate a larger sequence alignment and global similarity score. Although not as optimal as the Smith-Waterman algorithm, the FASTA algorithm nevertheless executes in more rapid time and thus offers a tradeoff between comparison accuracy versus execution time.^(20,109-110) The FASTA algorithm has been implemented. <http://www-nbrf.georgetown.edu/pirwww/search/fasta.html>

The BLAST (basic local alignment search tool) algorithm is another approximate heuristic algorithm used to compute suboptimal pairwise similarity comparisons. The BLAST algorithm uses the hotspot strategy of employing more stringent

rules to locate fewer and better alignment hotspots. This strategy concentrates on finding regions of high local similarity in alignments without gaps although alignments with some gaps can be created by chaining together several locally similar regions. Hotspot extensions are attempted into the surrounding regions. The BLAST algorithm is an improvement over the similar FASTA algorithm by offering three advantages:

1. More rapid execution time.
2. Output includes a range of solutions.
3. Each reported match is accompanied by an estimate of statistical significance.

Thus, the BLAST algorithm has become the dominant search engine for biological sequence databases.^(20,111) The BLAST algorithm has been implemented. <http://www.ncbi.nlm.nih.gov/BLAST/>

Sequence similarity searching has been reviewed.^(20,107,112-115)

4. Gene Expression

Gene expression is defined as the use of quantitative messenger RNA (mRNA)-level measurements of gene expression in order to characterize biological processes and elucidate the mechanisms of gene transcription. The objective of gene expression is the quantitative measurement of mRNA expression particularly under the influence of drug or disease perturbations.

As described in the survey of Carulli et al.,⁽¹¹⁶⁾ the identification of differential gene expression associated with biological processes is a central research problem in molecular genetics. High throughput analysis of differential gene expression incorporates five technologies:

1. Expressed sequence tags (ESTs).
2. DNA microarrays.
3. Subtractive cloning.
4. Differential display.
5. Serial analysis of gene expression (SAGE).

High throughput gene expression experiments are used for four purposes:

1. Identification of novel genes.
2. Identification of molecular markers for pathological processes.
3. Identification of potential drug targets.
4. Elucidation of molecular events associated with drug treatment in pharmacogenomics.

High throughput gene expression assays enable the simultaneous monitoring of thousands of genes in parallel and generate vast amounts of gene expression data. The large-scale investigation of gene expression attaches functional activity to structural genetic maps and therefore is an essential milestone in the paradigm shift from static structural genomics to dynamic functional genomics. High throughput gene expression technologies have been reviewed.⁽¹¹⁶⁻¹¹⁷⁾

4.1. Gene Expression Databases

Gene expression databases provide integrated data management and analysis systems for the transcriptional expression data generated by large-scale gene expression experiments. As described in the survey of Baldock and Davidson,⁽¹¹⁸⁾ gene expression databases should include fourteen data fields:

1. Gene expression assays.
2. Database scope.
3. Gene expression data.
 - a. Gene name.
 - b. Method or assay.

- c. Temporal information.
- d. Spatial information.
- e. Quantification.
- f. Gene products.
- g. User annotation of existing data.
- h. Linked entries.
- i. Links to other databases.
1. Internet access.
2. Internet submission.

Gene expression databases have not established defined standards for the collection, storage, retrieval and querying of gene expression data derived from libraries of gene expression experiments.

Human gene expression databases include eight representative implementations:

1. Cellular Response Database.⁽¹¹⁹⁾ <http://LHI5.umbc.edu/crd>
2. dbEST.⁽¹²⁰⁾ <http://www.ncbi.nlm.nih.gov/dbEST/index.html>
3. GeneCards.⁽¹²¹⁾ <http://bioinformatics.weizmann.ac.il/cards/>
4. Globin Gene Server.⁽¹²²⁾ <http://globin.cse.psu.edu>
5. Human Developmental Anatomy. <http://www.ana.ed.ac.uk/anatomy/database/humat/>
6. Kidney Development Database.⁽¹²³⁾
<http://www.ana.ed.ac.uk/anatomy/database/kidbase//kidhome.html>
7. Merck Gene Index.⁽¹²⁴⁾ http://www.merck.com/mrl/merck_gene_index.2.html
8. Tooth Gene Expression Database.⁽¹²⁵⁾ <http://bite-it.helsinki.fi/>

Gene expression databases have been reviewed.^(9,118-119,126-132)

4.2. Gene Expression Database Mining

Gene expression database mining is an emerging technology. Gene expression database mining is used to identify intrinsic patterns and relationships in gene expression data. The identification of patterns in complex gene expression datasets provides two benefits:

1. Generation of insight into gene transcription conditions.
2. Characterization of multiple gene expression profiles in complex biological processes, e.g. pathological states.

As described in the survey of Bassett et al.,⁽¹³¹⁾ gene expression data analysis uses two approaches:

1. Hypothesis testing.
2. Knowledge discovery.

Hypothesis testing investigates whether the induction or perturbation of a biological process leads to predicted results. Knowledge discovery detects internal structure in biological data. Knowledge discovery in gene expression data analysis employs two methodologies:

1. Statistics, e.g. cluster analysis.
2. Visualization.

Data visualization is used to display snapshots of cluster analysis results generated from large gene expression datasets.

Gene expression database mining has been reviewed.^(131,133-139)

5. Proteomics

Proteomics is defined as the use of quantitative protein-level measurements of gene expression in order to characterize biological processes and elucidate the mechanisms of gene translation. The objective of proteomics is the quantitative measurement of protein expression particularly under the influence of drug or disease perturbations.⁽¹⁴⁰⁾

Proteomics analysis of a mixture of proteins incorporates three procedures:

1. Protein resolution.
2. Protein identification.
3. Protein quantitation.

Protein resolution is performed using two-dimensional polyacrylamide gel electrophoresis. Protein identification is accomplished using Edman degradation, mass spectrometry and Western immunoblotting. Protein quantitation is achieved using scanners and phosphorimagers.

Gene expression monitors gene transcription whereas proteomics monitors gene translation. Because of the additional requirements of the secondary translation stage, proteomics has more restrictive expression and post-translational modification conditions than gene expression. Proteomics poses greater stringency conditions than gene expression for the phenotypic expression of a candidate gene. Thus, proteomics provides a more direct response in functional genomics than the indirect approach provided by gene expression. Proteomics has been reviewed.^(64,140-161)

5.1. Proteome Databases

Proteome databases provide integrated data management and analysis systems for the translational expression data generated by large-scale proteomics experiments. Proteome databases integrate the expression levels and properties of thousands of proteins with the thousands of genes identified on genetic maps and offer a global approach to the study of gene expression.

As described in the survey of Celis et al.,⁽¹⁶²⁾ proteome databases address five research problems that cannot be resolved by DNA analysis:

1. Relative abundance of protein products.
2. Post-translational modifications.
3. Subcellular localizations.
4. Molecular turnover.
5. Protein interactions.

The creation of comprehensive databases of genes and gene products will lay the foundation for the further construction of comprehensive databases of higher-level mechanisms, e.g. regulation of gene expression, metabolic pathways and signalling cascades.⁽¹⁴⁰⁾

Human proteome databases include eleven representative implementations:

1. ANL Human Breast Epithelial Cell Protein 2DE Database.⁽¹⁶³⁾

http://www.anl.gov/BIO/PMG/projects/index_hbreast.html

2. FindMod Tool.⁽¹⁶⁴⁾ <http://www.expasy.ch/tools/findmod>

1. Heart High-Performance 2-DE Database.⁽¹⁶⁵⁾ <http://www.mdc-berlin.de/~emu/heart/>
2. HEART-2DPAGE.⁽¹⁶⁶⁾ <http://userpage.chemie.fu-berlin.de/~pleiss/dhzb.html>
3. HSC-2DPAGE.⁽¹⁶⁷⁾ <http://www.harefield.nthames.nhs.uk/nhli/protein/>

4. Human 2-D Page Databases.⁽¹⁶⁸⁾ <http://biobase.dk/cgi-bin/celis>
5. Joint Protein Structure Laboratory 2D-Gel.⁽¹⁶⁹⁾ <http://www.ludwig.edu.au/jpsl/jpslhome.html>
6. NCI 2DWG Image Meta-Database.⁽¹⁷⁰⁾ <http://www-lecb.ncifcrf.gov/2dwgDB/>
7. Prostate Expression Database.⁽¹⁷¹⁾ <http://chroma.mbt.washington.edu/PEDB/>

10. SWISS-2DPAGE.⁽¹⁷²⁾ <http://www.expasy.ch/ch2d/>
11. WORLD-2DPAGE.⁽¹⁷³⁾ <http://www.expasy.ch/ch2d/2d-index.html>

Proteome databases have been reviewed.^(140,162,174-175)

5.2. Proteome Database Mining

Proteome database mining is an emerging technology. Proteome database mining is used to identify intrinsic patterns and relationships in proteomics data. The identification of patterns in complex proteomic datasets provides two benefits:

1. Generation of insight into gene translation and post-translational modification conditions.
2. Characterization of multiple protein expression profiles in complex biological processes, e.g. pathological states.

Proteome database mining has been conducted on three experimental systems:

1. *Escherichia coli* K-12 proteome.⁽¹⁷⁶⁾
2. Human lymphoid proteins.⁽¹⁷⁷⁾
3. Toxicity evaluation of drug candidates.⁽¹³⁸⁾

6. Conclusions

The Human Genome Initiative is an international research program for the creation of high-resolution structural and functional maps of the human genome. Genome research projects generate enormous quantities of data. Database mining is the process of finding and extracting useful information from raw datasets. On the basis of the central dogma of molecular biology, computational genomics has identified a classification of three successive levels for the management and analysis of genetic data in scientific databases:

1. Genomics.
2. Gene expression.
3. Proteomics.

Genome database mining is referred to as computational genome annotation. Computational genome annotation is the identification of the protein-encoding regions of a genome and the assignment of functions to these genes on the basis of sequence similarity homologies against other genes of known function. Gene expression database mining is the identification of intrinsic patterns and relationships in transcriptional expression data generated by large-scale gene expression experiments. Proteome database mining is the identification of intrinsic patterns and relationships in translational expression data generated by large-scale proteomics experiments. As the determination of the DNA sequences comprising the human genome nears completion, the Human Genome Initiative is undergoing a paradigm shift from static structural genomics to dynamic functional genomics. Thus, gene expression and proteomics are emerging as the major intellectual challenges of database mining research in the postsequencing phase of the Human Genome Initiative.

Genome, gene expression and proteome database mining are complementary emerging technologies with much scope being available for improvements in data analysis. Improvements in genome, gene expression and proteome database mining algorithms will enable the prediction of protein function in the context of higher order processes such as the regulation of gene expression, metabolic pathways and signalling cascades. The final objective of such higher-level functional analysis will be the elucidation of integrated mapping between genotype and phenotype.⁽⁶⁴⁾ Future research

directions in genome database technologies have been reviewed.^(57,178-179)

References

1. Pearson, M.L. and Soll, D. (1991). The Human Genome Project: a paradigm for information management in the life sciences. *FASEB J.* **5**, **1**, 35-39.
2. Dizikes, G.J. (1995). Update on the Human Genome Project. *Clin. Lab. Med.* **15**, **4**, 973-988.
3. Gibbs, R.A. (1995). Pressing ahead with human genome sequencing. *Nat. Genet.* **11**, **2**, 121-125.
4. Guyer, M.S. and Collins, F.S. (1995). How is the Human Genome Project doing and what have we learned so far? *Proc. Natl. Acad. Sci. USA* **92**, **24**, 10841-10848.
5. Schlessinger, D. (1995). Genome sequencing projects. *Nat. Med.* **1**, **9**, 866-888.
6. Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A., Castle, A.B., Chiannikulchai, N., Chu, A., Clee, C., Cowles, S., Day, P.J.R., Dibling, T., East, C., Drouot, N., Dunham, I., Duprat, S., Edwards, C., Fan, J.B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T.C., McKusick, K.B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H.C., Page, D.C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D.K., Soderlund, C., Sun, W.L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M.D., Auffray, C., Walter, N.A.R., Brandon, R., Dehejia, A., Goodfellow, P.N., Houlgatte, R., Hudson, J.R., Ide, S.E., Iorio, K.R., Lee, W.Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M.H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J.C., Sikela, J.M., Beckmann, J.S., Weissenbach, J., Myers, R.M., Cox, D.R., James, M.R., Bentley, D., Deloukas, P., Lander, E.S. and Hudson, T.J. (1996). A gene map of the human genome. *Science* **274**, **5287**, 540-546.
7. Collins, F.S. (1997). Sequencing the human genome. *Hosp. Pract.* **32**, **1**, 35-54.
8. McKusick, V.A. (1997). Genomics: structural and functional studies of genomes. *Genomics* **45**, **2**, 244-249.
9. Strachan, T., Abitbol, M., Davidson, D. and Beckmann, J.S. (1997). A new dimension for the human genome project: towards comprehensive expression maps. *Nat. Genet.* **16**, **2**, 126-132.
10. Beck, S. and Sterk, P. (1998). Genome-scale DNA sequencing: where are we? *Curr. Opin. Biotechnol.* **9**, **1**, 116-121.
11. Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., Fearon, E., Hartwell, L., Langley, C.H., Mathies, R.A., Olson, M., Pawson, A.J., Pollard, T., Williamson, A., Wold, B., Buetow, K., Branscomb, E., Capecchi, M., Church, G., Garner, H., Gibbs, R.A., Hawkins, T., Hodgson, K., Knotek, M., Meisler, M., Rubin, G.M., Smith, L.M., Smith, R.F., Westerfield, M., Clayton, E.W., Fisher, N.L., Lerman, C.E., McInerney, J.D., Nebo, W., Press, N. and Valle, D. (1998). New goals for the U.S. Human Genome Project. *Science* **282**, **5389**, 682-689.
12. Hudson, T.J. (1998). The human genome project: tools for the identification of disease genes. *Clin. Invest. Med.* **21**, **6**, 267-276.
13. Kelavkar, U. and Shah, K. (1998). Advances in the human genome project: a review. *Mol. Biol. Rep.* **25**, **1**, 27-43.
14. Uddhav, K. and Ketan, S. (1998). Advances in the Human Genome Project: a review. *Mol. Biol. Rep.* **25**, **1**, 27-43.
15. van Ommen, G.J.B., Bakker, E. and den Dunnen, J.T. (1999). The human genome project and the future of diagnostics, treatment and prevention. *Lancet* **354**, **Supplement 1**, SI5-SI10.
16. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2000). Genbank. *Nucleic Acids Res.* **28**, **1**, 15-18.
17. National Center for Biotechnology Information (2000). Genbank overview.

<http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>

18. National Center for Biotechnology Information (1999). Genbank statistics.

<http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>

19. Waterman, M.S. (1995). *Introduction to Computational Biology*. Chapman and Hall, New York.

20. Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, New York.

21. Hunter, L. (1999). Progress in computational molecular biology. *Sigbio News*. **19**, **3**, 9-12.

22. Frenkel, K.A. (1991). The human genome project and informatics. *CACM* **34**, **11**, 40-51.

23. Robbins, R.J., Benton, D. and Snoddy, J. (1995). Informatics and the human genome project. *IEEE Eng. Med. Biol. Mag.* **14**, **6**, 694-701.

24. Ashburner, M. and Goodman, N. (1997). Informatics: genome and genetic databases. *Curr. Opin. Genet. Dev.* **7**, **6**, 750-756.

25. Miyano, S. (1997). Genome informatics: new frontiers of computer science and biosciences. In Y. Kambayashi and K. Yokota (Eds.), *International Symposium on Cooperative Database Systems for Advanced Applications*. World Scientific, Singapore, pp. 12-21.

26. Brutlag, D.L. (1998). Genomics and computational molecular biology. *Curr. Opin. Microbiol.* **1**, **3**, 340-345.

27. Saier, M.H. (1998). Genome sequencing and informatics: new tools for biochemical discoveries. *Plant Physiol.* **117**, **4**, 1129-1133.

28. Boland, M.V. and Murphy, R.F. (1999). Engineering in genomics. *IEEE Eng. Med. Biol. Mag.* **18**, **5**, 115-119.

29. Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery: an overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, California, pp. 1-34.

30. Kloggen, W. (1996). Knowledge discovery in databases and data mining. In Z.W. Ras and M. Michalewicz (Eds.), *Foundations of Intelligent Systems: 9th International Symposium, ISMIS '96*. Springer, Berlin, pp. 623-632.

31. Ming-Syan, C., Jiawei, H. and Yu, P.S. (1996). Data mining: an overview from a database perspective. *IEEE Trans. Knowl. Data Eng.* **8**, **6**, 866-883.

32. Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., Kloggen, W. and Simoudis, E. (1996). An overview of issues in developing industrial data mining and knowledge discovery applications. In E. Simoudis, J. Han and U. Fayyad (Eds.), *KDD-96: the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, California, pp. 89-95.

33. Rainsford, C.P. and Roddick, J.F. (1999). Database issues in knowledge discovery and data mining. *Aust. J. Inf. Syst.* **6**, **2**, 101-128.

34. Bohm, K. (1995). High performance computing for one of the Grand Challenges. In B. Hertzberger and G. Serazzi (Eds.), *High-Performance Computing and Networking*. Springer, Berlin, pp. 496-501.

35. Bohm, K. (1995). High performance computing for the human genome project. *Comput. Methods Programs Biomed.* **46**, **2**, 107-112.

36. Kanehisa, M. (1998). Grand challenges in bioinformatics. *Bioinformatics* **14**, **4**, 309.

37. Lander, E.S. (1996). The new genomics: global views of biology. *Science* **274**, **5287**, 536-539.

38. Kennedy, G.C. (1997). Impact of genomics on therapeutic drug development. *Drug Dev. Res.* **41**, **3-4**, 112-119.

39. Clark, M.S. (1999). Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* **21**, **2**, 121-130.
40. Grausz, J.D. (1998). Redefining genomics. *Drug Discov. Today* **3**, **1**, 11-18.
41. Wiley, S.R. (1998). Genomics in the real world. *Curr. Pharm. Des.* **4**, **5**, 417-422.
42. Cowley, A.W. (1999). The emergence of physiological genomics. *J. Vasc. Res.* **36**, **2**, 83-90.
43. Jordan, B.R. (1999). 'Genomics': buzzword or reality? *J. Biomed. Sci.* **6**, **3**, 145-150.
44. Kao, C.M. (1999). Functional genomic technologies: creating new paradigms for fundamental and applied biology. *Biotechnol. Prog.* **15**, **3**, 304-311.
45. Shapiro, L. and Harris, T. (2000). Finding function through structural genomics. *Current Opin. Biotechnol.* **11**, **1**, 31-25.
46. Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M.A. (2000). The EMBL nucleotide sequence database. *Nucleic Acids Res.* **28**, **1**, 19-23.
47. Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. and Gojobori, T. (1998). DNA Data Bank of Japan at work on genome sequence data. *Nucleic Acids Res.* **26**, **1**, 16-20.
48. Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, **1**, 45-48.
49. Letovsky, S.I., Cottingham, R.W., Porter, C.J. and Li, P.W.D. (1998). GDB: the Human Genome Database. *Nucleic Acids Res.* **26**, **1**, 94-99.
50. Hamosh, A., Scott, A.F., Amberger, J., Valle, D. and McKusick, V.A. (2000). Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15**, **1**, 57-61.
51. Rudd, K.E. (2000). EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res.* **28**, **1**, 60-64.
52. Blake, J.A., Eppig, J.T., Richardson, J.E. and Davisson, M.T. (2000). The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res.* **28**, **1**, 108-111.
53. Palm, C.J., Federspiel, N.A. and Davis, R.W. (2000). DAtA: database of Arabidopsis thaliana annotation. *Nucleic Acids Res.* **28**, **1**, 102-103.
54. Fuchs, R. and Cameron, G.N. (1991). Molecular biological databases: the challenge of the genome era. *Prog. Biophys. Mol. Biol.* **56**, **3**, 215-245.
55. Pearson, P.L. (1991). Genome mapping databases: data acquisition, storage and access. *Curr. Opin. Genet. Dev.* **1**, **1**, 119-123.
56. Fields, C. (1992). Data exchange and inter-database communication in genome projects. *Trends Biotechnol.* **10**, **1-2**, 58-61.
57. Fuchs, R., Rice, P. and Cameron, G.N. (1992). Molecular biological databases: present and future. *Trends Biotechnol.* **10**, **1-2**, 61-66.
58. Boguski, M.S. (1994). Bioinformatics. *Curr. Opin. Genet. Dev.* **4**, **3**, 383-388.
59. Karp, P.D. (1996). Database links are a foundation for interoperability. *Trends Biotechnol.* **14**, **8**, 273-279.
60. Borsani, G., Ballabio, A. and Banfi, S. (1998). A practical guide to orient yourself in the labyrinth of genome databases. *Hum. Mol. Genet.* **7**, **10**, 1641-1648.

61. Bishop, M.J. (1999). *Genetics Databases*. Academic Press, San Diego.
62. Letovsky, S.I. (1999). *Bioinformatics: Databases and Systems*. Kluwer Academic Publishers, Boston.
63. Pandey, A. and Lewitter, F. (1999). Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem. Sci.* **24**, **7**, 276-280.
64. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, **4**, 707-725.
65. Rouze, P., Pavy, N. and Rombauts, S. (1999). Genome annotation: which tools do we have for it? *Curr. Opin. Plant Biol.* **2**, **2**, 90-95.
66. Schilling, C.H., Schuster, S., Palsson, B.O. and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications. *Biotechnol. Prog.* **15**, **3**, 296-303.
67. Chan, S.C., Wong, A.K.C. and Chiu, D.K.Y. (1992). A survey of multiple sequence comparison methods. *Bull. Math. Biol.* **54**, **4**, 563-598.
68. Gotoh, O. (1999). Multiple sequence alignment: algorithms and applications. *Adv. Biophys.* **36**, 159-206.
69. Huang, X., Adams, M.D., Zhou, H. and Kerlavage, A. (1997). A tool for analyzing and annotating genomic sequences. *Genomics* **46**, **1**, 37-45.
70. Krogh, A., Mian, I.S. and Haussler, D. (1994). A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.* **22**, **22**, 4768-4778.
71. Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994). The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *ISMB* **2**, 354-362.
72. Xu, Y., Mural, R.J., and Uberbacher, E.C. (1994). Constructing gene models from accurately predicted exons: an application of dynamic programming. *Comput. Appl. Biosci.* **10**, **6**, 613-623.
73. Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992). Prediction of gene structure. *J. Mol. Biol.* **226**, **1**, 141-157.
74. Borodovsky, M.Y. and McIninch, J.D. (1993). GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, **2**, 123-133.
75. Fields, C.A. and Soderlund, C.A. (1990). gm: a practical tool for automating DNA sequence analysis. *Comput. Appl. Biosci.* **6**, **3**, 263-270.
76. Snyder, E.E. and Stormo, G.D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* **21**, **3**, 607-613.
77. Snyder, E.E. and Stormo, G.D. (1995). Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* **248**, **1**, 1-18.
78. Henderson, J., Salzberg, S. and Fasman, K.H. (1997). Finding genes in DNA with a hidden Markov model. *J. Comput. Biol.* **4**, **2**, 127-142.
79. Dong, S. and Searls, D.B. (1994). Gene structure prediction by linguistic methods. *Genomics* **23**, **3**, 540-551.
80. Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, **1**, 78-94.
81. Milanese, L., Kolchanov, N.A., Rogozin, I.B., Ischenko, I.V., Kel, A.E., Orlov, Y.L., Ponomarenko, M.P. and Vezzoni, P. (1993). GenViewer: a computing tool for protein-coding regions prediction in nucleotide sequences. In H.A. Lim, J.W. Fickett, C.R. Cantor and R.J. Robbins (Eds.), *The Second International Conference on Bioinformatics, Supercomputing*

and *Complex Genome Analysis*. World Scientific, Singapore, pp. 573-587.

82. Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26, 2**, 544-548.
83. Uberbacher, E.C., Einstein, J.R., Guan, X. and Mural, R.J. (1993). Gene recognition and assembly in the GRAIL system: progress and challenges. In H.A. Lim, J.W. Fickett, C.R. Cantor and R.J. Robbins (Eds.), *The Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*. World Scientific, Singapore, pp. 465-476.
84. Xu, Y., Einstein, J.R., Mural, R.J., Shah, M. and Uberbacher, E.C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. *ISMB* **2**, 376-384.
85. Gelfand, M.S. and Roytberg, M.A. (1993). Prediction of the exon-intron structure by a dynamic programming approach. *Biosystems* **30, 1-3**, 173-182.
86. Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22, 24**, 5156-5163.
87. Salzberg, S., Delcher, A.L., Fasman, K.H. and Henderson, J. (1998). A decision tree system for finding genes in DNA. *J. Comput. Biol.* **5, 4**, 667-680.
88. Zhang, M.Q. (1997). Identification of protein coding regions in the human genome based on quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94, 2**, 565-568.
89. Rogozin, I.B., Milanese, L. and Kolchanov, N.A. (1996). Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci.* **12, 3**, 161-170.
90. Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996). Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA* **93, 17**, 9061-9066.
91. Hutchinson, G.B. and Hayden, M.R. (1992). The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res.* **20, 13**, 3453-3462.
92. Thomas, A. and Skolnick, M.H. (1994). A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.* **11, 3**, 149-160.
93. Burset M. and Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34, 3**, 353-367.
94. Gelfand, M.S. (1995). Prediction of function in DNA sequence analysis. *J. Comput. Biol.* **2, 1**, 87-115.
95. Fickett, J.W. (1996). Finding genes by computer: the state of the art. *Trends Genet.* **12, 8**, 316-320.
96. Fickett, J.W. (1996). The gene identification problem: an overview for developers. *Comput. Chem.* **20, 1**, 103-118.
97. Tiwari, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1996). Gene identification in silico. *Curr. Sci.* **71, 1**, 12-24.
98. Claverie, J.M. (1997). Computational methods for the identification of genes in vertebrate genome sequences. *Hum. Mol. Genet.* **6, 10**, 1735-1744.
99. Guigo, R. (1997). Computational gene identification. *J. Mol. Med.* **75, 6**, 389-393.
100. Guigo, R. (1997). Computational gene identification: an open problem. *Comput. Chem.* **21, 4**, 215-222.
101. Rawlings, C.J. and Searls, D.B. (1997). Computational gene discovery and human disease. *Curr. Opin. Genet. Dev.* **7, 3**, 416-423.
102. Batzoglou, S., Berger, B., Kleitman, D.J., Lander, E.S. and Pachter, L. (1998). Recent developments in computational

- gene recognition. In G. Fischer and U. Rehmann (Eds.), Proceedings of the International Congress of Mathematicians, Vol I (Berlin, 1998). *Doc. Math. Extra Volume ICM I*, 649-658.
103. Burge, C.B. and Karlin, S. (1998). Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8, 3**, 346-354.
104. Claverie, J.M. (1998). Computational methods for exon detection. *Mol. Biotechnol.* **10, 1**, 27-48.
105. Mural, R.J. (1999). Current status of computational gene finding: a perspective. *Methods Enzymol.* **303**, 77-83.
106. Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome* **10, 2**, 168-175.
107. Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994). Issues in searching molecular sequence databases. *Nat. Genet.* **6, 2**, 119-129.
108. Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147, 1**, 195-197.
109. Lipman, D.J. and Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science* **227, 4693**, 1435-1441.
110. Pearson, W.R. and Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85, 8**, 2444-2448.
111. Altschul, S., Gish, W., Miller, W., Myers, E.W. and Lipman, D. (1990). A basic local alignment search tool. *J. Mol. Biol.* **215, 3**, 403-310.
112. Taylor, W.R. (1994). Protein structure modelling from remote sequence similarity. *J. Biotechnol.* **35, 2-3**, 281-291.
113. Waterman, M.S. and Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. USA* **91, 11**, 4625-4628.
114. Bucher, P. and Hofmann, K. (1996). A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. *ISMB* **4**, 44-51.
115. Pearson, W.R. (1997). Identifying distantly related protein sequences. *Comput. Appl. Biosci.* **13, 4**, 325-335.
116. Carulli, J.P., Artinger, M., Swain, P.M., Root, C.D., Chee, L., Tulig, C., Guerin, J., Osborne, M., Stein, G., Lian, J. and Lomedico, P.T. (1998). High throughput analysis of differential gene expression. *J. Cell Biochem.* **30-31, Supplement 0**, 286-296.
117. Lennon, G.G. (2000). High-throughput gene expression analysis for drug discovery. *Drug Discov. Today* **5, 2**, 59-66.
118. Baldock, R. and Davidson, D. (1999). Gene expression databases. In M.J. Bishop (Ed.), *Genetics Databases*. Academic Press. San Diego, pp. 247-268.
119. Sorace, J.M. and Canfield, K. (1998). Collaborative bioinformatics: data warehouses for targeted experimental results. *J. Interferon Cytokine Res.* **18, 9**, 799-802.
120. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993). dbEST – database for "expressed sequence tags". *Nat. Genet.* **4, 4**, 332-333.
121. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1998). GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* **14, 8**, 656-664.
122. Reimer, C., ElSherbini, A., Stojanovic, N., Schwartz, S., Kwitkin, P.B., Miller, W. and Hardison, R. (1998). A database of experimental results on globin gene expression. *Genomics* **53, 3**, 325-337.
123. Davies, J.A. (1999). The Kidney Development Database. *Dev. Genet.* **24, 3-4**, 194-198.

124. Eckman, B.A., Aaronson, J.S., Borkowski, J.A., Bailey, W.J., Elliston, K.O., Williamson, A.R. and Blevins, R.A. (1998). The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics* **14**, **1**, 2-13.
125. Nieminen, P., Pekkanen, M., Aberg, T. and Thesleff, I. (1998). A graphical WWW-database on gene expression in tooth. *Eur. J. Oral Sci.* **106**, **Supplement 1**, 7-11.
126. Matsubara, K. and Okubo, K. (1993). Identification of new genes by systematic analysis of cDNAs and database expression. *Curr. Opin. Biotechnol.* **4**, **6**, 672-677.
127. Fields, C. (1994). Analysis of gene expression by tissue and developmental stage. *Curr. Opin. Biotechnol.* **5**, **6**, 595-598.
128. Matsubara, K. and Okubo, K. (1995). Recent progress in human molecular biology and expression profiling of active genes in the body. *Jpn. J. Pharmacol.* **69**, **3**, 181-185.
129. Fannon, M.R. (1996). Gene expression in normal and disease states: identification of therapeutic targets. *Trends Biotechnol.* **14**, **8**, 294-298.
130. Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M. and Boguski, M.S. (1998). Data management and analysis for gene expression arrays. *Nat. Genet.* **20**, **1**, 19-23.
131. Bassett, D.E., Eisen, M.B. and Boguski, M.S. (1999). Gene expression informatics: it's all in your mine. *Nat. Genet.* **21**, **Supplement 1**, 51-55.
132. Jones, D.A. and Fitzpatrick, F.A. (1999). Genomics and the discovery of new drug targets. *Curr. Opin. Chem. Biol.* **3**, **1**, 71-76.
133. Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. and Davis, R.W. (1998). Microassays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16**, **7**, 301-306.
134. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999). Expression profiling using cDNA microarrays. *Nat. Genet.* **21**, **Supplement 1**, 10-14.
135. Going, J.J. and Gusterson, B.A. (1999). Molecular pathology and future developments. *Eur. J. Cancer* **35**, **14**, 1895-1904.
136. Nakamura, R.M. (1999). Technology that will initiate future revolutionary changes in healthcare and the clinical laboratory. *J. Clin. Lab. Anal.* **13**, **2**, 49-52.
137. Stratowa, C. and Wilgenbus, K.K. (1999). Gene expression profiling in drug discovery and development. *Curr. Opin. Mol. Ther.* **1**, **6**, 671-679.
138. Todd, M.D. and Ulrich, R.G. (1999). Emerging technologies for accelerated toxicity evaluation of potential drug candidates. *Curr. Opin. Drug Discov. Dev.* **2**, **1**, 58-68.
139. Zweiger, G. (1999). Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. *Trends Biotechnol.* **17**, **11**, 429-436.
140. Anderson, N.L. and Anderson, N.G. (1998). Proteome and proteomics: new technologies, new concepts and new words. *Electrophoresis* **19**, **11**, 1853-1861.
141. Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphrey-Smith, I., Hochstrasser, D.F. and Williams, K.L. (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **13**, 19-50.
142. Humphrey-Smith, I. and Blackstock, W. (1997). Proteome analysis: genomics via the output rather than the input code. *J. Protein Chem.* **16**, **5**, 537-544.

143. Humphrey-Smith, I., Cordwell, S.J. and Blackstock, W.P. (1997). Proteome research: complementarity and limitations with respect to the RNA and DNA words. *Electrophoresis* **18, 8**, 1217-1242.
144. James, P. (1997). Breakthroughs and views of genomes and proteomes. *Biochem. Biophys. Res. Commun.* **231, 1**, 1-6.
145. James, P. (1997). Of genomes and proteomes. *Biochem. Biophys. Res. Commun.* **231, 1**, 1-6.
146. James, P. (1997). Protein identification in the post-genome era: the rapid rise of proteomics. *Q. Rev. Biophys.* **30, 4**, 279-331.
147. Ashton, C. (1998). Proteomics – extending the molecular understanding of disease processes to the protein level. *Pharm. Technol. Int.* **10, 11**, XLVI-LVI.
148. Haynes, P.A., Gygi, S.P., Figeys, D. and Aebersold, R. (1998). Proteome analysis: biological assay or data archive? *Electrophoresis* **19, 11**, 1862-1871.
149. Hochstrasser, D.F. (1998). Proteome in perspective. *Clin. Chem. Lab. Med.* **36, 11**, 825-836.
150. Mullner, S., Neumann, T. and Lottspeich, F. (1998). Proteomics – a new way for drug target discovery. *Arzneimittel-Forschung* **48, 1**, 93-95.
151. Yates, J.R. (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33, 1**, 1-19.
152. Blackstock, W.P. and Weir, M.P. (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17, 3**, 121-127.
153. Hancock, W., Apffel, A., Chakel, J., Hahnenberger, K., Choudhary, G., Traina, J.A. and Pungor, E. (1999). Integrated genomic/proteomic analysis. *Anal. Chem.* **71, 21**, 742A-748A.
154. Hatzimanikatis, V., Choe, L.H. and Lee, K.H. (1999). Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.* **15, 3**, 312-318.
155. Lopez, M.F. (1999). Proteome analysis: I. Gene products are where the biological action is. *J. Chromatogr. B Biomed. Sci. Appl.* **722, 1-2**, 191-202.
156. Page, M.J., Amess, B., Rohlf, C., Stubberfield, C. and Parekh, R. (1999). Proteomics: a major new technology for the drug discovery process. *Drug Discov. Today* **4, 2**, 55-62.
157. Patton, W.F. (1999). Proteome analysis: II. Protein subcellular redistribution: linking physiology to genomics via the proteome and separation technologies involved. *J. Chromatogr. B Biomed. Sci. Appl.* **722, 1-2**, 203-223.
158. Stubberfield, C.R. and Page, M.J. (1999). Applying proteomics to drug discovery. *Expert Opin. Invest. Drugs* **8, 1**, 65-70.
159. Wang, J.H. and Hewick, R.M. (1999). Proteomics in drug discovery. *Drug Discov. Today* **4, 3**, 129-133.
160. Williams, K.L. (1999). Genomes and proteomes: towards a multidimensional view of biology. *Electrophoresis* **20, 4-5**, 678-688.
161. Yates, J.R. (2000). Mass spectrometry. From genomics to proteomics. *Trends Genet.* **16, 1**, 5-8.
162. Celis, J.E., Ostergaard, M., Jensen, N.A., Gromova, I., Rasmussen, H.H. and Gromov, P. (1998). Human and mouse proteomic databases: novel resources in the protein universe. *FEBS Lett.* **430, 1-2**, 64-72.
163. Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997). A two-dimensional electrophoresis database of human breast epithelial cell proteins. *Electrophoresis* **18, 3-4**, 573-581.
164. Wilkins, M.R., Gasteiger, E., Gooley, A.A., Herbert, B.R., Molloy, M.P., Binz, P.A., Ou, K., Sanchez, J.C., Bairoch, A., Williams, K.L. and Hochstrasser, D.F. (1999). High-throughput mass spectrometric discovery of protein

post-translational modifications. *J. Mol. Biol.* **289**, **3**, 645-657.

165. Muller, E.C., Thiede, B., Zimny-Arndt, U., Scheler, C., Prehm, J., Muller-Werdan, U., Wittmann-Liebold, B., Otto, A. and Jungblut, P. (1996). High-performance human myocardial two-dimensional electrophoresis database: edition 1996. *Electrophoresis* **17**, **11**, 1700-1712.

166. Pleissner, K.P., Sander, S., Oswald, H., Regitz-Zagrosek, V. and Fleck, E. (1996). The construction of the World Wide Web-accessible myocardial two-dimensional gel electrophoresis protein database "HEART-2DPAGE": a practical approach. *Electrophoresis* **17**, **8**, 1386-1392.

167. Evans, G., Wheeler, C.H., Corbett, J.M. and Dunn, M.J. (1997). Construction of HSC-2DPAGE: a two-dimensional gel electrophoresis database of heart proteins. *Electrophoresis* **18**, **3-4**, 471-479.

168. Leffers, H., Dejgaard, K., Honore, B., Madsen, P., Nielsen, M.S. and Celis J.E. (1996). cDNA expression and human two-dimensional gel protein databases: towards integrating DNA and protein information. *Electrophoresis* **17**, **11**, 1713-1719.

169. Ji, H., Reid, G.E., Moritz, R.L., Eddes, J.S., Burgess, A.W. and Simpson, R.J. (1997). A two-dimensional gel database of human colon carcinoma proteins. *Electrophoresis* **18**, **3-4**, 605-613.

170. Lemkin, P.F. (1997). The 2DWG meta-database of two-dimensional electrophoretic gel images on the Internet. *Electrophoresis* **18**, **15**, 2759-2773.

171. Hawkins, V., Doll, D., Bumgarner, R., Smith, T., Abajian, C., Hood, L. and Nelson, P.S. (1999). PEDB: the Prostate Expression Database. *Nucleic Acids Res.* **27**, **1**, 204-208.

172. Hoogland, C., Sanchez, J.C., Tonella, L., Binz, P.A., Bairoch, A., Hochstrasser, D.F. and Appel, R.D. (2000). The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* **28**, **1**, 286-288.

173. Appel, R.D., Bairoch, A., Sanchez, J.C., Vargas, J.R., Golaz, O., Pasquali, C. and Hochstrasser, D.F. (1996). Federated 2-DE database: a simple means of publishing 2-DE data. *Electrophoresis* **17**, **3**, 540-546.

174. Celis, J.E., Gromov, P., Ostergaard, M., Madsen, P., Honore, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D.B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J. and Rasmussen, H.H. (1996). Human 2-D PAGE databases for proteome analysis in health and disease: <http://biophase.dk/cgi-bin/celis>. *FEBS Lett.* **398**, **2-3**, 129-134.

175. Bairoch, A. (1997). Proteome databases. In M.R. Wilkins, K.L. Williams, R.D. Appel and D.F. Hochstrasser (Eds.), *Proteome Research: New Frontiers in Functional Genomics*. Springer, New York, pp. 93-132.

176. Link, A.J., Robison, K. and Church, G.M. (1997). Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18**, **8**, 1259-1313.

177. Hanash, S.M. and Techroew, D. (1998). Mining the human proteome: experience with the human lymphoid protein database. *Electrophoresis* **19**, **11**, 2004-2009.

178. Letovsky, S.I. and Berlyn, M.B. (1994). Issues in the development of complex scientific databases. In L. Hunter (Ed.), *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences. Vol. V: Biotechnology Computing*. IEEE Computer Society Press, Los Alamitos, California, pp. 5-14.

179. Sargent, R., Fuhrman, D., Critchlow, T., Di Sera, T., Mecklenburg, R., Lindstrom, G. and Cartwright, P. (1996). The design and implementation of a database for human genome research. In P. Svenson and J.C. French (Eds.), *Eighth International Conference on Scientific and Statistical Database Management*, IEEE Computer Society Press, Los Alamitos, California, pp. 220-225.