Bioinformatics

Gerald Benoît ©2003 Benoit
Simmons College
300 The Fenway
Boston, MA 02115-5898
617-521-2879, benoit@simmons.edu

Introduction

This is the first review of bioinformatics to appear in *ARIST* and so is written primarily for computer- and information scientists (CS/IS) interested in exploring how the typical work of CS/IS, such as information retrieval, visualization, data modeling, algorithms, and web-based resources might contribute to biology. It is written secondarily for biologists curious about the computer-based technologies that have been harnessed and what data resources have been created to support their research.

The first goal of this review, then, is to encourage in its readers an appreciation of the very large amounts and specific properties of detailed data about humans and other species. It should lay before the reader the range of bioinformatics applications to molecular biology, clinical medicine, pharmacology, biotechnology and many other associated disciplines. Finally, the reviewer wishes to suggest how a knowledge of computer science and information science techniques applied to Internet-based databases can further the research of many areas of biology.

For computer- and information science, the work of molecular biology is puzzling although the products that have been created so far to address the needs of biology is staggering. While simultaneously researching increasingly focused areas, biologists, in concert with computer science, are developing tools able to integrate a wide range of data types. For example, data models, such as the NCBI and XML, described below, are highly developed; programming applications and tools for Perl, Corba, and Java have evolved very quickly; specialized visualization techniques, too, are available for a range of data processing needs, from interactive information retrieval, information visualization, data mining, to sequence data and protein folds. The impetus for collaboration between biology and CS/IS lay in managing the quantities of data to reveal biological relationships and to achieve this goal through innovative techniques to locate, aggregate, manipulate, and present the data, through user-friendly, cross-platform applications.

Wide-ranging as these efforts are, they can be categorized into two groups: data management and biological analysis. The first, managing the data, involves manipulating files and strings of data, here specifically strings of DNA or proteins, making bioinformatics seem very much like an IR and data mining activity. Similarly, there are IR systems to aggregate and provide annotations for chemical sequences of DNA, and make the results available through browsers. The second task, biological analysis, is to make the data representing the structure and function of genomic sequences more comprehensible to research biologists. There are many tools already available for data management and manipulation, which form the body of this review. But the challenge to biology, and the opportunity for computer- and information science, is in exploiting data to provide researchers the tools for exploring increasingly sophisticated questions.

This review follows the recommendation (Altman and Koza, 1996, 73; Denn and MacMullan, 2002; Bayat, 2002) that the fusion of biology and computing – bioinformatics – is served by providing computer- and information

scientists with the fundamentals of biology and simultaneously offering biologists enough technical know-how to see how computer technology can facilitate their work. However, the field evolves daily, making defining the field and demonstrating the work of bioinformatics a challenge that may not delight all. Nevertheless, the benefits that bioinformatics offer society through a deeper understanding of genes, disease, and drug treatments, encourage a review that will encompass as many researchers as possible from both domains. The intersection of skills create a foundation for both domains: facility with using the Internet, knowledge of databases of sequence data and of structure data, specific biological knowledge of sequence analysis, sequence alignment, phylogenetic analysis, data models, information retrieval, and predictive methods (based on Lesk, 2002). A good starting point is a data-centric synopsis of how some records are created and manipulated by biologists.

The basics of biology and the computer files produced through research

In brief, bioinformatics is gathering data about DNA (deoxyribonucleic acid), protein sequences and structures, genomes and proteomes (all defined below) and storing these data in local, commercial, and freely-accessible, Internet-enabled databases. Molecular biology generates tremendous amounts of data from DNA or protein sequences, macromolecular structures and the results of functional genomics that need to be made useful – scientifically-sound and able to be interpreted by subject specialists – in a computationally efficient manner. Researchers need to know the nature of individual genomes, or the genetic material in the chromosomes of an organism, and their relationships. Early research efforts resulted in data that were stored in flat-files and queried using tools like Fasta and PSI-Blast for comparing protein sequences (Altshul, Madden, Schaffen, Zhang & Zhang 1997; Korf, Yandell & Bedell, 2003). Knowledge of how and why the data are gathered and modeled as they are, and how the research literature is being integrated into the biologists' toolkit, will suggest opportunities for CS/IS to further the information and processing needs of the field.

DNA Structure and Sequencing

The genetic material of all living organisms, the substance of heredity, is "DNA", or deoxyribonucleic acid. In 1953, British scientists Watson and Crick used x-ray crystallography to discover that DNA is in the form of a double-helix of molecules in pairs of chemical bases held together by weak bonds (Watson & Crick, 1953). The four chemical bases are purines (adenine, abbreviated "A", and guanine, "G") and pyrimidines (cytosine (C) and thymine (T)). The bases form pairs only between A and T and between G and C so one can deduce the base sequence of each single strand from its partner. Each base differs from others by the combination of oxygen, carbon, nitrogen, and hydrogen. Every base is attached to a "deoxyribose," or sugar molecule, and to a phosphate molecule, creating a nucleotide. Each nucleotide is linked in a certain order, or sequence, through the phosphate group. The precise order and linking of the bases within the DNA (i.e., the genotype) determines what proteins that gene produces, and ultimately the phenotype of the organism.

In the simplest sense, a gene is a linear sequence of nucleotides that encode information for the corresponding linear sequence of amino acids that form a protein. The information in DNA is first transcribed to a messenger RNA (mRNA), which is then decoded by ribosomes and other factors, translating the nucleotide code into amino acid code. The linear amino acid sequence folds into a three-dimensional conformation for a functional protein.

Colinearity of DNA and protein code is not exact, because the nucleotide code is often interrupted by introns, segments that are removed from the mRNA. Thus, in most eukaryotic genes, the final protein code is created by the

juxtaposition of exons (expressed segments). [Eukaryotic refers to non-bacterial, non-viral organisms; introns are DNA sequence s that interrupt the protein-coding sequence of a gene and is transcribed into RNA but is cut out of the message before it is translated into protein; exons are the protein-coding DNA sequence of a gene.]

The transcriptome refers to all the mRNA that are present in a cell (genes that are turned on), while the proteome comprises all the proteins that are made. There may not be a correspondence because gene expression can be regulated at various levels.

Biologists may look for functional clustering (e.g., based on metabolic pathways or sequence segments), or relationships of proteins, such as homologous (structurally and sequentially similar), or analogous proteins (related folds). The volume and heterogeneity of the data require the creation of algorithms for basic analysis, such as protein sequence analysis (Miller, Gurd, & Brass 1999) and uncovering introns and exons (Zhang 1999; Boguski 1999).

Some genome projects try to identify the small regions of DNA that vary between individuals. These differences may underlie disease susceptibility and drug responsiveness, particularly the most common variations that are called SNPs (single nucleotide polymorphism) [HGPI, 2002]. In addition, other genome projects focus on non-human DNA sequences, such as plant genetics (Lim, 2002) and biotechnology (Chawla, 2002).

To explain how some biologists generate the data, Table 1 [NCBI] details the sequencing process:

Mapping	Identify set of clones that span the region of the genome to be sequenced
Library Creation	Make sets of smaller clones from mapped clones
Template Preparation	Purify the DNA and perform sequence chemistries
Gel Electrophoresis	Determine sequences from smaller clones
Pre-finishing & Finishing	Apply special techniques to produce high quality sequences
Data Editing/Annotation	Quality control, verification, biological annotation, submission to public databases

Table 1: Sequencing Process

Sequencing and Analysis

First chromosomes (of up to a couple of hundred million bases) are divided into smaller pieces, or "subcloned." A template is created from the shorter piece to generate fragments, each differing only by a single base; this changing of the base is the mutegenesis. That single base is used as an identifier during template preparation of sequence reaction. Using florescent dyes, the fragments can be identified by color when the fragments are separated by a process called "gel electrophoresis." The base at the end of each fragment is now identified ("base-calling") to help recreate the original sequences of A, T, C, and G for each subcloned piece. A four-color histogram (a "chromotograph") is created to show the presence and location for each of the bases. Finally, the short sequences in blocks of about 500 bases (called the "read length") are assembled by computer into long, continuous stretches for analysis of errors, gene-coding regions, and other distinctions.

Microarrays

"Sequencing" is the determining of the order of nucleotides (the base sequences) in a DNA or RNA (ribonucleic acid) or the order of amino acids in a protein. Analyzing these sequences provides the functional identification of genes. The analysis of whole genomes can be performed through "DNA microarrays." This technique, described in Kohane, Kho and Butte (2003) and Bowtell & Sambrook (2003), is applied to many endeavors, such as high-throughput genotyping, comparative genomic hybridization, monitoring of gene expression, and detection of single nucleotide polymorphisms. Such procedures may determine the effect of gene expression, map disease loci, demonstrate chromosomal aberrations, and categorize tumor expression patterns.

Protein structure and sequencing

Proteins are made up of sequences of amino acids, and to date approximately 400,000 protein sequences are known. There are, 20 different amino acids and depending on their arrangement create larger macromolecular structures, such as the 51 amino acids that form insulin. Various techniques, such as x-ray crystallography and nuclear magnetic resonance (NMR), generate three-dimensional coordinate data (*x-y-z*), and stored in the appropriately themed database, such as the Protein Data Bank (PDB) (Berman, 2000; Bernstein et al., 1977), for manipulation by programs (e.g., Orengo, 1999; Orengo & Taylor 1996).

Examples of projects generating data and the publishing of the data

Some examples illustrate. One, the Human Genome Project (HGP), is an example of a well-known molecular biology project that generates data stored in publicly-accessible, Internet-enabled databases. The project's original goals was to reveal the human genes, once estimated to be as many as 100,000. ~~This project's original goal was to "reveal the estimated 100,000 human genes within our DNA as well as the regions controlling them."~~ The human genome is built from almost 3 billion bases (Lander, 2001; Venter, 2001). Since the completion of the sequencing of the human genome (or the determining of the exact order of the genes in the chromosomes, described by the total number of base pairs), though still genetically complex, it turns out humans have about only, 21,000 genes (Goodman, 2003, p. 12): the number of possible sequences of the pairs of DNA (about 3 billion) and twenty amino acids lead to a huge number of possible combinations in DNA and proteins, suggesting to the reader one reason why the data sets are so large.

Other examples include research into specific organisms' or cell types' gene expression (measuring mRNA produced in cells under different conditions) (Eisen & Brown, 1999; Cheung et al.,1999; Duggan , 1999; Lipshutz, Fodor, Gingeras & Lockhard, 1999). Still others concentrate on systems, such as metabolic pathways, regulatory networks, and protein-protein interaction data from, 2-hybrid experiments.

Finally, the results of sequencing and expression research are published in publicly accessible databases, such as the National Center for Biotechnology Information's (NCBI) GenBank, which holds over 12 billion bases in 11.5 million entries (Benson et al., 2000). NCBI, a part of the National Library of Medicine (NLM) and National Institutes of Health (NIH), hosts PubMed, GenBank, molecular sequencing databases, and research literature databases (e.g., PubMed, Online Mendelian Inheritance of Man (OMIM) and are detailed below). Researchers might also store the sequence data in other very large, multidimensional databases, such as the Wisconsin Package or GCG sequencers. Ultimately results appear in professional communication forums, intended to teach bioinformatics and to keep the scientist up to date, through online means (e.g., Medline, SCI), printed and online journals, and conferences.

The variety of uses and formats of the data, coupled with the desire for integrating all resources through a single interface begs the question of data integration (Gerstein, 2000; Gerstein & Jansen, 2000). Molecular biology's information needs (e.g., as described in Jiang, Xu, & Zhang, 2002) might be classified into three groups: raw data, data integrated with bibliographic systems and other data reflecting biological processes, and data used predictively (Wilson, Kreychman, & Gerstein, 2000) to solve specific questions.

Successfully describing, retrieving, and presenting these data from these sources is one part of bioinformatics; another is the manipulation of these data in biologically-meaningful ways. The problem facing scientists in both fields is how to address data redundancy and multiplicity in these very large databases and how to integrate

multiple sources of data where different nomenclature and file formats are common. To understand how some of the differences arose, we review the many definitions of bioinformatics. After defining them, the review considers specific genome projects and the myriad databases to store and techniques to retrieve and present the data.

Bioinformatics defined

Bayat accurately and broadly defines the discipline as "the application of tools of computation and analysis to the capture and interpretation of biological data" and, operationally, that "The main tools of a bioinformatician are computer software programs and the internet. A fundamental activity is sequence analysis of DNA and proteins using various programs and databases available on the world wide web" (2002, p. 1019).

NCBI also defines bioinformatics as a single broad discipline, but with three "important subdisciplines":

- "The development of new algorithms and statistics with which to assess relationships among members of large data sets;
- The analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and
- The development and implementation of tools that enable efficient access and management of different types of information" (http://www.ncbi.nlm.nih.gov/Education/)

However, Ellis (2003a) notes 40 published operational definitions between, 2000-2001 and another 37 (2003b) in, 2003, suggesting the definitions vary by subdiscipline. It appears that several specialties were working out for themselves their roles in molecular biology (Ibba, 2002) and the legacy of computational biology's influence on their work. For example, the question of bioinformatics appears as basic scientific work, education (Brass, 2000; Ouzounis, 2002; Sander, 2002; Pearson, 2001), employment and retooling (Henry, 2001; Zauhar, 2001; Gardiner, 2001; Bass, 2000), computing (NIH, 2000; Watkins, 2001; Sansom & Smith, 2000; Miller & Attwood, 2001); genomics and proteomics (Zimmerman, 2003) and sub-domains, as well as overviews of the field (McDonald, 2001; Bernstein, 2001; Adler & Conklin, 2000; Cottle, 2001). Some of the reviews are biologists speaking across domain barriers to computer & information science, wondering aloud how to manage the data (Watkins, 2001; Butte, 2001; Roos, 2001; Attwood, 2000).

A second reason for the many definitions is because gene research is expanding from its base of gene sequences towards function prediction (Tsigelny, 2002), among other topics, and so raises different research investigations, and the role of other fields, specifically computer science, information science, statistics (Ewens & Grant, 2001), and mathematics.

There are also more questions about how to make the results of retrieval more intelligible to biologists (information retrieval, visualization, and data mining). The literature reflects the maturing practice of bioinformatics as more evidently two distinct work roles (biology + information technology) collaborating on specific biological questions: e.g., drug discovery (Gatto, 2003; Dougherty & Projan, 2003; Hillisch & Hilgenfeld, 2003), pharmaceutics (Fagen & Swindells, 2000), pharmacogenomics (Jain, 2003; Kalow, Meyer & Tyndale, 2001), neurosurgery (Taylor, Mainprize, & Rutka, 2003), and medical practice (Grant, Moshyk, Kuskniruk, & Moehr, 2003; Breski, 2002).

Turning the tables, there are also bioinformatics reviews of biology and the emerging bioinformatics industry (Benaim Jalfon, 2001). The question of employment and grants in bioinformatics also appear in the literature

(Van Haren, 2002; Kolatkar, 2002; Jenson, 2002; Basi, Clum, & Modi, 2003; Henry, 2002; Calandra, 2002; Schachter, 2002).

Another source of definitions is found where the concept of "informatics" has invaded. There are several fusions of information science (informatics) with biomedical fields and this prompts some researchers to try to define themselves and their relationship to bioinformatics (Altman & Dugan, 2003; Bayat, 2003; NCBI, 2003; Ouzounis, 2002; Fuchs, 2001). Altman (2000, 2003) questions the relationship with medical informatics (Altman, 2000; Altman, 2003); Grant, Moshyk, Kushmiruk & Moehr (2003) with health informatics; Andersson, Larsson, Larsson & Jacob (1999) with mathematics, and others with traditional genomics (Rost, Honig & Valencia, 2002; Valencia, 2002; Altman, 2003; Chicurel, 2002), and even specifically proteins (Zimmerman, 2003).

Most intriguing, perhaps, for readers of *ARIST* are the information metaphors in biology. Nishikawa (2002), for example, proposes an "Island Model" of biology – how given a set of inputs, the proteins will cluster based on similarity in amino acid sequences. His description of clustering uses identifiable fields of molecular processes (such as amino acid sequences) that map directly to concepts in the information retrieval and clustering literature. The description of the behavior of polypeptides under physiological conditions parallel the behavior of a query under different user cognitive conditions.

The critical role, however, of "information" (in the LIS sense) remains (Gywne, 2002; Denn & MacMullen, 2002; Paris, 2003; Luscombe, Greenbaum & Gerstein, 2001) as does the search for resolving data modeling related problems, such as applying XML, or Extensible Markup Language.

Historically some of the work today associated with "bioinformatics" was viewed as genomics and computational biology (Priami, 2003; Fogel & Corne, 2003; Gascuel & Sagot, 2000). Faced with more data than can be efficiently processed and new research questions, biology turned, as many fields do, to statistics and technology to help model phenomena, and to expose interesting patterns and deviations from patterns. With the introduction of computer technology, a greater variety of patterns could be examined more quickly and without the introduction of human error. As applied to biomedical processes, a picture of the invisible world of the molecule became possible. Expanded modeling of these processes made biology more approachable and enabled research simultaneously on a broader descriptive level and on highly focused questions. For instance, some computational molecular biologists (e.g., Leszczynski 1999) moved to numerical simulation as a complement to traditional theoretical and experimental approaches, in order to probe *in silico* theories that cannot otherwise be examined, such as phenomena at the atomic level. However, pursuing this path leads to quantum mechanics and the various simplifications (e.g., the Born-Oppenheimer approximation) employed and outside this review. Nevertheless, it emphasizes that computational molecular biology focuses on computerized and mathematical answers to biological questions.

Metaphorically, genomics is the precursor to bioinformatics. Genomics "is operationally defined as investigations into the structure and function of very large numbers of genes undertaken in a simultaneous fashion" (Univ. of California-Davis, 2003). The primary work effort is comparative genomics and functional genomics. Functional genomics infers the function of gene expression, typically based on eukaryotic homologues (non-viral, non-bacterial organisms with the same origin and function) or other model organisms, not usually tested *in vivo*. Functional genomics also includes the mutagenesis (the production of changes in DNA sequences that affect gene products); the study of genotypes (the specific changes in DNA sequences in a mutant); and the effect of these on phenotypes (the biological consequence of a mutagen's presence).

Functional genomic testing of phenotypes relies heavily upon technology for analysis, such as analytical chemistry, imagery, robotics and process automation. The boundaries between genomics and bioinformatics are porous, the literature suggesting that genomics, like computational molecular biology, focuses on physical biological processes. Bioinformatics emphasizes the storage and retrieval of biological data and the research literature: to organize very large heterogeneous structures and determine algorithms for clustering, retrieving and displaying subsets in meaningful ways, relying heavily upon information visualization, statistics, and integrating appropriate supporting bibliographies. For example, genomics may emphasize the research of specific microbiology and genetics of specific organism, such as the *E. coli* genome or the fruitfly *Drosophila melanogaster*; bioinformatics on the manipulation of the generated data, although some work, such as DNA microarrays, may belong to both (e.g., Tessier, Benoît, Rigby, Hogues, van het Hoog, Thomas & Brousseau, 2000).

Some biologists define the concepts of biology in terms similar to those of information science. Nishikawa (2002) proposes an "Island Model" of biology – how given a set of inputs, the proteins will cluster based on similarity in amino acid sequences. His description of clustering uses identifiable fields of molecular processes (such as amino acid sequences) that map directly to concepts in the information retrieval and clustering literature. The description of the behavior of polypeptides under physiological conditions parallel the behavior of a query under different user cognitive conditions.

There are also recently published monographs that define the field and specific work within it, at times focusing on one of the subdisciplines. For general introductions and text-book type treatments, see especially Bergeron (2003), Baxivanis & Ouellette (2001), Lacroix & Critchlow (2003) and Krane & Raymer (2003), Krawetz & Womble (2003), Pevsner (2003), Lesk (2002), Misener & Krawetz (2003), Mount (2001), Orengo, Jones & Thornton (2003), Westhead, Parish & Twyman (2002) and Dwyer (2003). Barnes & Gray (2003), Gibson & Muse (2002), Mewes, Seidel & Weiss (2003), Primrose (2003), Sensen (2002), Wang, Wu, & Wang (2003), and Winter, Hickey & Fletcher (2002) speak to geneticists; Tözeren (2004) to engineers and computer scientists. There are also many article-length discussions of bioinformatics (e.g., Kossida, Tahri, & Daizadeh, 2002).

Finally, Luscombe, Greenbaum & Gerstein (2001, 347) propose a reasoned definition which they submitted to the Oxford English Dictionary, and is adopted by this review, with the small addition of "information science": "bioinformatics is conceptualising biology in terms of molecules (in the sense of Physical chemistry) and applying **'informatics techniques'** (derived from disciplines such as applied maths, computer [and information] science and statistics) to **understand** and **organise** the **information** associated with these molecules, on a **large scale**. In short, bioinformatics is a management information system for molecular biology and has many **practical applications**" [emphasis in original].

As these definitions demonstrate, the emphasis in bioinformatics is on the data sources, the innovative and efficient manipulation thereof, and improving understanding of the data and aid in its discovery and extraction. The challenge for computer and information science is to help biologists "organize data in a way that allows researchers to access existing information and to submit new entries as they are produced …, to develop tools and resources that aid in the analysis of data …, and to analyse the data and interpret the results in a biologically meaningful manner" (Luscombe, Greenbaum & Gerstein, 2001, 358).

Professional Communication

Journals

Professional communication plays a vital role in bioinformatics and is surprisingly dependent upon the Internet as a communications medium. Another important avenue is serials. Both McCain (2003) and Garfield (2002) offer ranked lists of journals (Table 2) using different selection criteria and which suggest that the field is evolving:

McCain	Garfield
Journal of Computational Biology	*Bioinformatics*
Pacific Symposium on Biocomputing	*Genetic Engineering News*
Proceedings of the International Conference on Intelligent Systems for Molecular Biology	*Abstracts of Papers of the American Chemical Society*
Bioinformatics	*Nature*
In Silico Biology	*Scientist*
Briefings in Bioinformatics	*Nature Biotechnology*
Journal of Molecular Graphics and Modeling	*Science*

Table 2: Ranked list of bioinformatics journals

A complete list of journals publishing, 25 or more items annually on bioinformatics, is in Appendix, 2.

Very large databases

The other and most critical form of professional communication is through very large Internet-based databases and software packages to access them. Table 3 outlines some of the databases by theme or function.

Some of the packages support particular functions of gene sequencing. For example, there are at least 150 free software applications (Gilbert, 2000, pp. 157-184) addressing all aspects of sequencing. One popular package is Accelryus' "GCG Wisconsin Package" (http://www.accelrys.com/about/gcg.html). An integrated suite of over 130 program tools for manipulating, analysis, and comparing nucleotide and protein sequences, this package also includes a graphic user interface, SeqLab, to interact with color-coded graphic sequences. The package includes sequence comparison statistics (alignment of two sequences to indicate gaps, best fit and *x*/*y* plotting of sequence similarity), database searching tools (LookUp, StringSearch for biological literature, BLAST (basic local alignment search tool), NetBLAST, FASTA and others for sequence strings, PAUPSearch, GrowTree, Diverge for phylogenetic relations, Fragment Assembly, gene finding and pattern recognition tools, protein analysis (e.g., PeptideMap), ChopUp, Reformat), and others for manipulating text files. Similarly, DNASTAR's Lasergene Sequence Analysis Software is a suite of eight applications to trim and assemble sequence data; discover and annotate genes patterns; predict protein secondary structure; create Boolean queries from sequence similarity, consensus sequence and text terms; sequencing, hybridization, and transcription; create maps; and import data from other sources (Burland, 2001, p. 71). It isn't possible to review all applications and their capabilities, but the reader can see these desktop software applications emphasize the graphic display of data, pattern detection and sequence prediction, and integration of the literature.

Examples of Data Models and Sources and their Uses

Bioinformatics encompasses all aspects of molecular biology research and has made amazing advancements in understanding and sharing information about molecular processes. Not the least of which is the structure of protein molecules themselves and the biological sequences from which they are derived. Applying statistical,

mathematical and computer techniques, however, has pushed bioinformatics into fuller explanations of the invisible and unanticipated, such as energy equations to model the dynamic behavior of molecules, linear and 3D protein functions, and probabilistic models of sequences (Durbin, Eddy, Krogh, & Mitchison, 1998), especially Hidden Markov Models (Koski, 2001). The computer flat-files that contain these data are now used to visualize known biological structures (e.g., using Cn3D available from NCBI) and to help predict macromolecule structure in 3D. Furthermore, and of particular interest to pharmaceutical companies, bioinformatics' application of genomic data applied to the study of variation in host and pathogen DNA and disease helps design drug treatments (Lengauer, 2002). Of significance to information science are the databases and software used to store, retrieve, and display data.

The vast amount of data generated in sequencing and in the support literature has resulted in a tremendous number of databases and the attendant difficulties of querying across a heterogeneous environment. As a result portals like PubMed and topic- or function-specific databases have been created. Zdobnov, Lopez, Apweiler, & Etzold (2002a) suggest categorizing the databases by these themes:

- bibliographic;
- taxonomic;
- nucleic acid;
- genomic;
- protein and specialized protein databases;
- protein families, domains and functional sites;
- proteomics initiatives; and
- enzyme/metabolic pathways.

In addition, data models specific to molecular biology and genomics have appeared. One such model is the Sequence Retrieval System (SRS) (Zdobnov, Lopez, Apweiler & Etzold, 2002b; Etzold, Ulyanov & Argos 1996); another is the "NCBI Data Model." These are detailed below, after a brief exposition of the main information resources employed in bioinformatics.

This section outlines some of the major databases and tools, arranged by general databases, protein sequence databases and tools, nucleotide sequences, classification schema, error reduction techniques, NCBI data model, and XML. To ease reading the text, websites are listed in Appendix 1 and Table 3 will help the reader classify the many resources.

PUBLICLY ACCESSIBLE DATABASES	Examples	
Nucleotide sequencing	GenBank, DDBJ, EMBL, MGDB, GSX, NDB	13 billion bases from > 100,000 species DNA Database of Japan European Molecular Bio Lab Mouse Genome Database Mouse Gene Expression DB Nucleic Acid DB
Protein sequences	SwissPrott, TrEMBL, TrEMBLnew PIR	 Annotated Supplement to SwissProt Weekly, pre-processed update to TrEMBL Protein Information Resource
3D Structure Data	PDB	
	MMDB	
	Cambridge	Small molecule structural database
Enzymes & Com-	LIGAND	Chemical compounds and reactions

pounds		
Sequence motifs (alignments)	PROSITE BLOCKS PRINTS Pfam ProDom	Sequence motif Derived from Prosite Protein families database of alignments and hidden Markov Models Protein domains
Pathway & Complexes	Pathway	Metabolic and regulatory pathway maps
Molecular disease	OMIM	
Biomedical Literature	PubMed MedLine	
Vectors	UniVec	Identification of vector contaminates
Protein mutation	PMD	Protein Mutant DB
Gene Expression	GEO	Gene Expression Omnibus
Amino Acid Indices	Aaindex	Amino Acid Index
Protein/Peptide Literature	LITDB	
Gene Catalog	GENES	KEGG Gene DB

TABLE 3. EXAMPLES OF PUBLICLY ACCESSIBLE BIOINFORMATIC DATABASES (Bergeron, 2003, 45-46)

Bibliographic and taxonomic databases

There are far too many resources to be reviewed here. A comprehensive list is available at http://www.expasy.ch/alinks.html and in the first January issue each year of *Nucleic Acid Research*.

The most commonly used, publicly available resource is Medline (or PubMed). Commercial databases include Embase (biomedical and pharmacological abstracts), Agricola, and Biosis (the former Biological Abstracts). Interestingly taxonomic databases reflect an old issue in librarianship, that of controlled vocabularies (ontologies) reflecting the knowledge and modes of expression of a given field. NCBI maintains the most important taxonomic databases, whose hierarchical taxonomy is used by Nucleotide Sequence Databases, SWISS-PROT, and TrEMBL (along with derivatives such as NiceProt). Another important source is the Chemical Abstracts database, which includes the bibliographic file ("CAPlus") and a file of compounds, the Registry File. According to the STN Database Summary (http://info.cas.org/ONLINE/DBSS/registryss.html), there are about 52 million records, of which 30,818,220 were sequences for either proteins or nucleic acids. Another commercial database on the STN system is DGENE, Derwent's Geneseq database covering sequences form patents published by 40 patent offices worldwide. The December, 2003 brochure for the database states "More than half of the sequence data that appear in DGENE is not available in any other public sequence database" (http://www.derwent.com/geneseqweb/).

Genome databases

The popularity of the Human Genome Project has introduced to the public three significant databases for human genes. The primary human genome database is the Genome Database (GDB). Related to this is the Online Mendelian Inheritance in Man (OMIM) that catalogues all human genes and genetic disorders. The Sequence Variation Database, like OMIM and GDB, maps genetic variation, but has links to many sequence variance databases (EBI-Mutations) and via the Sequence Retrieval System (SRS) interface to other human mutation databases. Increasingly portals, such as GeneCard (GeneCard), are instituted to harmonize searches.

Perhaps the most well-known are the protein sequence databases. Like the nucleotide databases, the protein sequence databases fall into two groups: all species' data or specific organisms'. Of interest to information science is the further division of these databases into "sequence data" or "annotated sequence data." SWISS-PROT (Bairoch & Apweiler, 2000) is an annotated universal protein sequence database (Swiss-Prot) and strives to quality in the annotations and integration with other biomolecular databases. Each entry is analyzed by biologists: as of

May, 2000, there were more than 85,000 annotated sequence entries from more than 6,000 different species. A sister product was created from Swiss-Prot, called TrEMBL (Translation of EMBL nucleotide sequence database), to speed new sequence information to the public. SP-TrEMBL focuses on entries to be incorporated later into Swiss-Prot. REM-TrEMBL contains other data that will not be integrated because it may be redundant or are truncated or are not proteins or fragments legitimately translated *in vivo*. SPTR (SWALL) is another protein sequence database that provides non-redundant sequence data by focusing on data currency in SWISS-PROT, ignoring REM-TrEMBL, and by performing sequence comparisons against a database of all known isoforms. Table 4 outlines some genomic applications by usage.

User Function	Example resource (database or software application)
Sequence Search	BLAST, BLASTN, CLUSTALW, Fasta, Motif, PBLast, TBLASTIN
Submission	AceDB, Audet, BankIT, Sakura, Sequin, WebIN
Information Retrieval	Entrez, DBGET, IDEAS
Linkage	LocusLink
Portal	KEGG
Structure Match	DC, DALI, SCOP, Searchlite, Structure Explorer, VAST
Visualization	CAD, Cn3D, Mage, RasMol/MolMol, Swiss-PDB Viewer, VRML, WebMol
Protein-Protein Interaction	BRITE
Microarray Gene Expression Profiles	Expression
Open Reading Frame Locator	ORF Index

TABLE 4. GENOMIC APPLICATIONS (after Bergeron, 2003, p. 62)

Nucleotides

GenBank (NCBI), the European Bioinformatics Institute (EBI) (Apweiler et al., 2003), and the DNA Data Bank of Japan have joined to create the International Nucleotide Sequence Database Collaboration. The quality and currency of the data vary between databases. The quality of the data in the nucleotide sequence databases is the responsibility of the authors or submitters (the scientists themselves, no professional enforcement of standards). With more than 10 billion nucleotides in more than 10 million individual entries, one can imagine the potential error rate (EBI-Stats). See Rodriguez-Tomé (2001) for a description of EMBL and examples of interfaces for submitting and searching the databases and the Genome Monitoring Table for updates on the progress of genome sequencing projects (GMT).

Protein Sequence Databases

The specialized protein sequence databases perform different functions with the data – such as pre-clustering of SwissProt records (CluSTr) (Apweiler et al., 2001) catalogues and structure-based classification of peptidases (MEROPS, and "PepCards", or classification, nomenclature and hyperlinks for each peptidase, and Fam[ily]Cards and ClanCards). The Yeast Protein Database (YDP for Saccharomyces cerevisiae) details about 6,000 yeast proteins. The protein classification schemas define the cellular role, function, and pathway, and other information about the functional data in the "YPD Protein Reports."

The finding of relationships when an unknown protein cannot be matched to other known structures calls for examining the "sequence signatures." PROSITE, PRINTS, PFAM, ProDom (ProDom), and especially InterPro attempt in one form or another to derive patterns from sequence databases, using various sequencing and clustering algorithms (Guerra & Istrail, 2000; Guigó & Gusfield, 2002; Benson & Page, 2003, Gascuel & Moret, 2001). InterPro (Integrated Resource of Protein Families, Domains and Functional Sites) in an integrated documentation resource for PROSITE, PRINTS, and Pfam, which helps address the question of ambiguous biological relevance when

a pattern is detected (e.g., by ignoring family discriminators), by linking to known protein sequences in SWISS-PROT and TrEMBL. InterPro entries are available as XML-formatted files (EBI-InterPro).

Some work focuses on learning more about organisms are various levels. For instance, the Kyoto Encyclopedia of Genes and Genomes (or "KEGG") and Proteome Analysis Initiative (PAI) provide information about the gene, transcript, protein, and function level. In additional all completely sequenced organisms in Swiss-Prot and TrEMBL have the proteome set information available through InterPro and CluSTr. This includes the amino acid composition and links to the homology (HSSP, Homology-derived Secondary Structure of Proteins (HSSP)). As evidence of the need to understand better the function of proteins (Ashburner et al., 2000), genes and how to associate the literature more successfully, there is growing interest in ontologies (e.g., Bard, 2003). As Paris (2002, p. XX) notes "Recognizing the existing problems in classifying and organizing information about cell and molecular biology, especially in this era of exponentially exploding data from genomics and proteomics experiments, a consortium was proposed by Michael Ashburner in 1998 (ISMB), and eventually established in 1999 to create and promote a consistent, scientifically-sound, useful "gene ontology". The result (GO) is a troika of tree-schemes based on molecular function, biological processes, and cellular component; genes and gene products can map to multiple locations in multiple trees, reflecting biological diversity and (to some extent) ambiguity of knowledge. "If used to support the annotation process, this is one approach that will help eliminate many problems …" (Paris, 2002).

Classification Schema – Enzyme Commission

Given that so many bioinformatics computer applications rely on post-coordinate retrieval techniques, it is interesting to note that some researchers have turned to a pre-coordinate technique, classification schema. One example of a protein schema is Enzyme Commission Classification which assigned an "E.C. no.": "The most useful definition of enzyme function is through biochemical reaction, i.e., the chemical reaction catalyzed by the enzyme. The biochemical reaction catalyzed by an enzyme is assigned as an EC (Enzyme Commission Classification) number. If a new enzyme activity is discovered then a new EC number has to be created. The EC scheme is hierarchical with four levels hence the four elements to an EC number, e.g., alcohol dehydrogenase has an EC number of 1.1.1.1. The first number is the top level of the classification (in this case an oxidoreductase), the last number refers to the specific enzyme (the specific oxidoreductase alcohol dehydrogenase). The middle numbers give details about the reaction catalysed" (http://www.bru.edu.ac.uk/~rhamilto/rsh_phd.html).

In addition, there are also standards being developed for model organism databases, such as Lincoln Stein's MOD project (http://www.genome.gov/10006365; Harris & Parkinson, 2003).

Error Correction

Other projects address error potentials by clustering and specialization. Reminiscent of latent semantic indexing (Gordon & Dumais, 1998), clustering of data to remove redundant record is performed by UniGene (UniGene) and STACK (Sequence Tag Alignment and Consensus Knowledgebase). The Ribosomal Database Project (RDP), HIV Sequence Database (HIV), IMGT database (IMGT), Transfac (transcription factors and transcription factor binding sites, EPD (Eukaryotic Promoter Database), REBASE, and GoBase are all examples of specialty resources.

NCBI Data Model

Of particular interest to information scientists is the NCBI Data Model. "This new and more powerful model made possible the rapid development of software and the integration of databases that underlie the popular

Entrez retrieval system and on which the GenBank data is now built (Schuler, Epstein, Ohkawa, & Kans, 1996). The advances of the model (e.g., the ability to move effortlessly from the published literature to DNA sequences to the proteins they encode, to chromosome maps of the genes, and to the three-dimensional structures of the proteins) have been apparent for years to biologists using Entrez …" (Baxevanis & Ouellette, 2001, p., 20). The Entrez database (Tatusova, Karsch-Mizrachi & Ostell 1999) has the complete genome of about 300 archaeal, bacterial and eukaryotic organisms, including exon and intron data.

The NCBI data model is an implementation of Abstract Syntax Notation 1, an ISO standard (ISO/IEC 8824-1:2002 (ASN.1)) for reliable encoding of data that is data-centered (here the DNA), human-interpretable, in computer-readable flat files. Based on this, NCBI's website (NCBI) is a portal, offering PubMed, Entrez, BLAST, OMIM and other services. One can search by author and journal (PubMed), protein, nucleotide, structure, genome, PMC, LocusLink, PopSet, OMIM, Taxonomy, book, ProbeSet, 3D Domain, UniSTS, Domain, SNP, Journal and UniGene. The definitions of each are available at this URL: http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/helpdoc.html.

Although readers can explore these sites for themselves, it seems useful to demonstrate here an example of the results of a search, to suggest the file structure and how one might want to parse the file. Below is the result of a search using Entrez (for a gene in humans; http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=6849043&form=6&db=n&Dopt=g). The reader will identify immediately with some of the "access points" in the NCBI Sequence Viewer (Figure 1) as well as notice the sequence of bases. Hyperlinks (in this case, organism, MedLine, PubMed, exon) in the record link the biological data to support literature. Notice the traditional access points (authors, title, journal), links to other literature entries (the Medline unique identifier (MUID) and PubMed identifier (PMID)), and among others, the base. This record also includes the sequence identifiers (Seq-id) because NCBI integrates sequence data from multiple sources.

```
LOCUS       HSDDT1                  166 bp    DNA     linear   PRI 01-FEB-2000
DEFINITION  Homo sapiens D-dopachrome tautomerase (DDT) gene, exon 1.
ACCESSION   AF012432
VERSION     AF012432.1  GI:2352911
KEYWORDS    .
SEGMENT     1 of 3
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 166)
  AUTHORS   Esumi,N., Budarf,M., Ciccarelli,L., Sellinger,B., Kozak,C.A. and
            Wistow,G.
  TITLE     Conserved gene structure and genomic linkage for D-dopachrome
            tautomerase (DDT) and MIF
  JOURNAL   Mamm. Genome 9 (9), 753-757 (1998)
  MEDLINE   98384542
   PUBMED   9716662
REFERENCE  , 2  (bases 1 to 166)
  AUTHORS   Esumi,N. and Wistow,G.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-JUL-1997) Molecular Structure and Function, NEI,
            Building 6, Rm. 331, NIH, Bethesda, MD, 20892, USA
FEATURES             Location/Qualifiers
     source          1..166
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="22"
     exon            1..146
                     /gene="DDT"
```

```
                     /number=1
BASE COUNT      , 24 a     61 c     50 g     31 t
ORIGIN
        1 cttcttccgc cagagctgtt tccgttcctc tgcccgccat gccgttcctg gagctggaca
       61 cgaatttgcc cgccaaccga gtgcccgcgg ggctggagaa acgactctgc gccgccgctg
      121 cctccatcct gggcaaacct gcggacgtaa gcgtgggccg ggcagc
```

Figure 1: NCBI example

The heterogeneity of data outside the various sequence databases and resources described above calls for greater cross-discipline mapping (similar to the UMLS) or semantically independent modeling schema, such as XML.

XML

There exist many xml-schema for biology and a glance at their numbers suggests the variety of specific needs for semantically useful descriptors. Guerrinia & Jackson (2000) offer an introduction to xml and document type declarations (DTDs) in biology; Maher (2001, §3) demonstrates applying BioML2SVG to gene sequences. The variety of XML-DTDs available today suggests, too, opportunities for mapping across schema. A list of over twenty distinctive schema is available at http://www.xml.com/pub/rg/Bioinformatics . Some are domain-specific, such as Neuron Markup Language (NeuroML), Genome Annotation Markup Elements (GAME), Ribonucleic Acid Markup Language (RiboML); others emphasize integration (Architecture for Genomic Annotation, Visualization and Exchange [AGAVE], Genbank to xml conversion [GB2XML], Integrated Taxonomic Information System [ITIS]), and even an XML-based Ontology Exchange Language (XOL). Visualization of data is critical in bioinformatics and several products are available that combine xml records and display techniques (EBI-Vis) and some work emphasize fine-grain parsing of XML records with interactive information retrieval.

Data mining and visualization

Some activities in molecular biology focus on predicting sequences where there are missing values or on establishing patterns that otherwise would be impossible to be detected. Data mining, defined as an "exploration and analysis by automatic and semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules" (Benoît, 2000) , when focused on biological processes turns bioinformatics into a data mining activity. See Benoît (2000) for a detailed overview of data mining or Kantardzic (2003) for an excellent course in statistical foundations. EBI (European Bioinformatics Institute, 1999) hosted a conference exploring the intersection of bioinformatics as the field matures and data mining: "During the last few years bioinformatics has been overwhelmed with increasing floods of data, both in terms of volume and in terms of new databases and new types of data. We are now entering the post-genomic age, where, in addition to complete genome sequences, we are learning about gene expression patterns and protein interactions on genomic scales. This poses new challenges. Old ways of dealing with data item by item are no longer sustainable and it is necessary to create new opportunities for discovering biological knowledge 'in silico' by data mining."

There have been other recent conferences discussing the union of data mining and bioinformatics. One example is "BioKDD, Workshop on Data Mining in Bioinformatics," sponsored by the Association for Computing Machinery's Knowledge Discovery and Data Mining special interest group, SIGKDD. The results of this conference, to be published in book form by Springer, demonstrates some of the product of biology + information technology: gene expression ("Determination of RNA folding pathway functional intermediates using a massively parallel genetic algorithm", "Extracing knowledge form gene expression data: a case study of Batten Disease"), microarrays

("Mining microarray expression data for classifier gene-cores", "Analysis of an associative memory neural network for pattern identification in gene expression data"; "Probabilistic approach to sequence assembly validation"); sequence modeling and clustering ("Maximum entropy methods for biological sequence modeling"; "Scalable algorithm for clustering protein sequences") (www.cs.rpi.edu/~zaki/BIOKIDD01). Another is the "Data Mining for Bioinformatics – towards in silico biology", supported by EBI-Welcome Trust (http://industry.ebi.ac.uk/datamining99). Jenssen, Öberg, Andersson, & Komorowski (2001) explore methods for mining networks of human genes.

Several monographs have appeared focusing on the union of data mining and biology: Schlichting & Egner (2001), Perner (2002), Calvanese, Lenzerini & Motwani (2003), and Ye (2003).

NCBI offers a "Tools for data mining" website (http://www.ncbi.nih.gov/Tools) describing BLAST with variants, such as BLAST, 2 and "specialized BLASTs for human, microbial, malaria and other genomes," and other tools; Clusters of Orthologous Groups (COGs), Map Viewer showing integrated views of chromosome maps, LogusLinks that "combines descriptive and sequence information on genetic loci through a single query interface", UniGene, ORF finder, to identify all possible ORFs in a DNA sequence, Electronic PCR ( ) to search DNA sequences for sequenced tagged sites (STSs), VAST Search (for structure-to-structure similarity searches), Cancer Chromosome Aberration Project, Human-Mouse Homology Maps, VecScreen (for identifying segments of a nucleic acid sequence), dbMHC (for human Major Histocompatibility Complex), Spidey (to align "one or more mRNA sequence to a single genomic sequence" and to try to "determine exon/intron structure") and the Cancer Genome Anatomy Project.

Computer software has been written to visualize atomic and molecular structures and information visualization, or what is sometimes called "alternative metaphor" (Bergeron, 2003). The various visualization tools differentiate between structures, say where the protein backbones are the same it is possible to articulate between differences by hiding or showing atoms, etc.

Most imaging uses data from PDB or the Molecular Modeling Databases (MMDB). After searching for a structure by protein name or identification label, e.g., "Glutamine synthetase" or "1FPY"), the data can be used by different applications for different presentations. For instance, wireframe diagrams can be generated by PyMol to emphasize the atomic bonds; Chimera to create ribbon diagrams to highlight the protein' secondary structure. Figures 2-5 are examples of visualization. Figure 2 is an example of the secondary structure of Colicin Ia, from *E. coli*, PDB ID number, 1CII:

Figure 2: Secondary structure of Colicin Ia

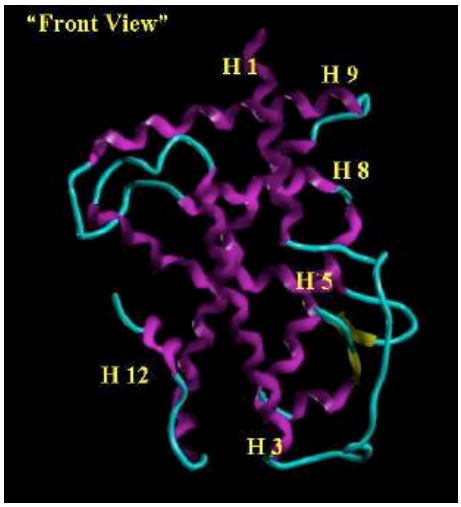

Figure 3: Orthographic view of retinoic acid receptor, created using SYBYL 6.6 (Taylor, 2000)

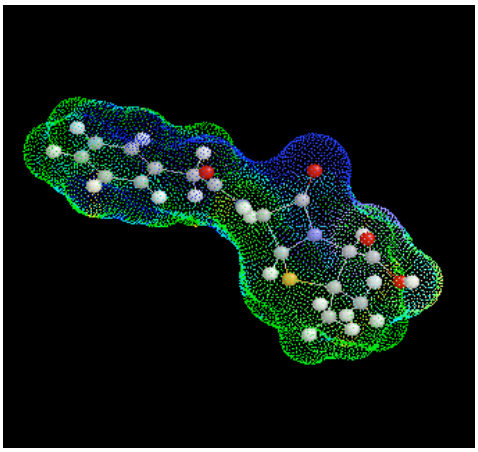

Figure 3: Surface potential of G-penicillin (RasTop, 2003)

Figure 4: Tap1 conformation, composite file (RasTop, 2003).

There is a range of options for visualization. The data can be modeled using general purpose software, such as Excel, Strata Vision 3D, Max3D, 3D-Studio, Ray Dream Studio, StatView, SAS/Insight, MatLab (Bergeron, 2003, p. 121) or creating one's own user interface using Java, VRML (Virtual Reality Modeling Language), C++, Python or computer aided design (CAD) applications. There are also specific tools already created that visualize sequences (NCBI-MapViewer), or structures. To visualize nucleotide locations use MapViewer; for protein structures there are many products (Swiss-PDB Viewer, WebMol, RasMol, Protein Explorer, Cn3D, VMD,MolMol, MidalPlus, PyMol, Chime, Chimera).

In addition to this type of visualization, researchers are developing applications with greater interactive and output opportunities. For instance, Böhringer et al. (2002) have developed a software package for visualizing genome wide information and generating "arbitrary karyograms, banding patterns and chromosome groupings" (p. 51).

From an information science and computer science perspective, many bioinformatics projects use familiar visualization techniques, applied to a specific user function. One example is the TimeSearcher that parses microarray data, sampling every few hours, and graphically representing the specific cellular functions (http://www.cs.umd.edu/hcil/bioinfovis/timesearcher.shtml). Similarly, the concept map idea has been applied to the gene ontology to generate directed acyclic graphs, according to biological functions (http://www.cs.umd.edu/hcil/bioinfovis/treemap.shtml). Indeed, there is opportunity on most university campuses for molecular biologists to work with budding programmers. Monographs by Chen (1999), Ware (2000), Card, MacKinlay and Shneiderman (2002), online lectures by Keim (1997), and a series of articles may guide in selecting the types of interfaces that may be useful and identify some pitfalls (e.g., Schroeder, Gilbert, van Helden, & Noy, 2001; Lanzenberger et al., 2003; Rhodes, Bergeron, & Sparr, 2002).

There are also many workshops and conferences, encouraging interdisciplinary visualization projects, such as the Bioinformatics Visualization Workshop sponsored by the Human-Computer Interaction Lab, University of Maryland, College Park, the IEEE Symposium on Information Visualization, and the various lectures of EBI.

Visualization in bioinformatics is exceptionally valuable and as the volume of data increases and becomes even more multidimensional, visualization techniques will integrate more data mining techniques. The following chart (Table 5) suggests how data can be subjected to visualization techniques and then combined with data mining approaches to be applied to biologically relevant tasks.

Data (examples)	Microarray, Genomic sequences, Protein structures, Gene ontologies, Biological pathway data
Visualization techniques (examples)	2D, 3D scattergrams, Heatmaps, Hierarchical treemaps, Temporal data searching, Topographic displays, Hierarchical presentations
Combined with data mining techniques	Supervised and unsupervised classification/categorization, Principle component analysis, Multi-dimensional scaling
Applied to biologically relevant tasks	Comparing samples, Identifying similar and different genes, Identifying targets, Defining pathways, Hypothesis generation

Table 5: Visualization Applications (based on Bergeron, 2003)

Collaborative opportunities

The aim of bioinformatics is to foster the organization and understanding of biological data so the future of collaboration between biology and various information- and computer sciences is bright indeed. Denn and MacMullen identify information science research areas ("domain analysis, information use, communication, and theories of information (basic); systems analysis and design, data modeling, classification, storage and retrieval, and human-computer interaction (applied)") and map them "onto a generalized model of a molecular biology experimental cycle" (2002, p. 556). They classify the "insertion points" of information science research into this experimental cycle as:

- "(a) the development of new tools and methods for managing, integrating, and visualizing data;
- (b) the application of tools and methods for integration, inference and discovery; and
- (c) theoretical approaches to biological information."

[The following authors were extracted from Denn and MacMullen and made available for interested readers: (a) Brazma, 2001; Gene Ontology Consortium, 2001; Nucleic Acid Research, 2002; NCBI, 2002; Paris, 1997; Searls, 2000; (b) Chen, 1997; Corruble & Ganascia, 1997; Juvan, 2001; Karp et al., 2001; Raychaudhuri, 2002; Wise, 2000; (c) de Jong, 2002; Poinçot, Lesteven, & Murtagh, 2000; Smalheiser & Swanson, 1996.]

This interpretation suggests closer personal collaboration and creates a reasoned foundation for future interactions.

Information Retrieval (IR) strives to improve the locating, retrieval, and ranking of full-text documents according to a number of ranking algorithms, which requires some knowledge of the collection (Baeza-Yates & Ribiero-Neto, 2000). So, from an IR point of view, the stored data in bioinformatics describing and representing biological elements and functions and published literature may also be viewed as decontextualized tokens. The manipulation of tokens from full-text literature and metadata can be lexical, based on the parsed texts and regions in flat-files, in which case there is greater opportunity in a purely information retrieval sense. Using these same tokens from an entirely biological perspective suggests the files can be mined to reflect manipulations of biological processes to expose unanticipated, interesting phenomena. For instance, in addition to mining databases of biological processes, research is also turning to mining the literature of molecular biology to expose and visualize unanticipated relationship among the records. In "biobibliometrics", Stapley and Benoît (2000) describe a visualization technique and retrieval system from co-occurrences of gene names in Medline abstracts.

Swanson has long been recognized as engaged in knowledge discovery across disciplines based on poor citations or missing data (e.g., 1999, 1997, 1991). Swanson's valuable literature investigations and Swanson & Smalheiser's (1997) work in finding complementary literatures are credited with stimulating the type of biological data/literature integration called for by Altman and others. Such work includes Cory (1997), Davies (1988), Lu, Janssen, Milios & Japkowicz (2001), ~~Jenssen & Lisa (2001)~~, Valdés-Pérez (1999) and many others. Swanson and collaborators continue to further his literature-based discovery of scientific knowledge: Gordon & Lindsay (1996), Weeber (2001), Weeber, Vos, Klein, & de Jong-van den Berg (2001), and Swanson, Smalheiser & Bookstein (2001) and expressly in biology (Gardy & Brinkman, 2003).

Molecular biology's information resources have progressed from just flat-files, through relational databases to web pages and portals hints at the need to construct robust digital library architectures to automate some of the bibliographic searches and automatically integrate knowledge bases from the literature. Such automation necessarily incorporates improved string comparison methods and one-dimensional alignment algorithms, but also linking specific genes to different disease traids as represented in the structured data. This suggests data modeling projects such as the Gene Ontology but also xml-schema for mapping between biological databases.

Naturally more computationally complicated proteome research requires greater data mining interaction, to aid in prediction of nucleic acid structures (Tsigelny, 2002), motif and pattern identification, structural comparison and 3d matching. Ultimately, to aid understanding of the scientific processes, computer and information science can step up efforts in several endeavors:

- machine learning (Baldi & Brunak, 2001),
- clustering algorithms (e.g., Jajugo, Sokolowski & Bock, 2002; Laender & Oliveira, 2002; Toronen 1999; Eisen 1998),
- connectionist systems in bioinformatics (Kasabov, 2003),
- artificial intelligence and heuristic methods in bioinformatics (Frasconi, 2003),
- visualization of literature,
- metadata and
- graphic representations of surfaces and volumes in general and for disease-specific analysis.

Some researchers suggest that "systems biology" is the next wave in bioinformatics. For instance, the Institute for Systems Biology (ISB, ¶1, http://www.systemsbiology.org/Default.aspx?pagename=predictiveandpreventive) claims that the HGP "has catalyzed two paradigm changes in contemporary biology and medicine – systems biology and predictive, preventative and personalized medicine." Similarly, there are recent conferences (e.g., International Conference on Systems Biology). Systems biology emphasizes causal relationships in cellular dynamics (e.g., Funahashi et al., 2003; Kitano, 2003) and even its own Systems Biology Markup Language (Hucka et al., 2003).

Another collaborative opportunity is in teaching computer programming skills. It is telling that biology feels under-served or needs to help itself more in manipulating the data better to suit the researcher's needs. Several monographs have been recently published that focus on the string manipulating strength of the programming language, Perl. For instance, *Genomic Perl: from Bioinformatics Basics to Working Code* (Dwyer, 2003), *Perl Programming for Biologists* (Jamison, 2003), *Beginning Perl for Bioinformatics* (Tisdall, 2001) and *Developing Bioinformatics Computer Skills* (Gibas & Jambeck, 2001) were recently published as manuals to instruct biologists how

to parse full-text records. Dwyer (2003) cogently and effectively demonstrates the merging of computer technology to biological processes.

Using Perl emphasizes the string-processing power needed for manipulating full-text files, although Java and C++ certainly offer equivalent file manipulation power along with greater computational efficiency and, at least for Java, a large graphic library. Montgomery (2003) offers Java code demonstrating how to access web-based databases GenBank, EMBL, and DDBJ. He describes several products to perform bioinformatic analysis: TIGR MultiExperimental Viewer, J-Express, BioJava, Apollo, and Sockeye and WebMol for 3D displays. Base4 hosts a large collection of Java applets, code, and "biowidgets" (Base4), as well as references to Tk/Tcl and Corba projects.

There are several free JavaBeans tools available for manipulating BLAST and GSDB (e.g., BlastView and AnnotView, http://www.cbil.upenn.edu/bioWidgets/). Anyone familiar with information storage and retrieval (IR) will see immediately the similarities between methods employed in full-text retrieval and the potential application of parsing, matching, clustering, similarity measurements, and display from IR to bioinformatics records. And that the biologists are learning the computing skills themselves suggests opportunities for information- and computer-science to assist.

Training

Futhermore, there are at least 45 undergraduate and 63 graduate programs (NIBIB) in bioinformatics (UNC-CH, 2003), suggesting the need for integrated training has been so acute that universities and the Federal government are willing to invest in it.

Another stimulus a rise in grants aimed at cross-training. For example, the National Science Foundation and the National Institutes of Health (both in the United States) support training in bioengineering and bioinformatics (http://www.nibib1.nih.gov/training/trainingopps.html). These grants support several trainee programs through the country, along with post-doctoral grants, e.g., the Institute for Pure and Applied Mathematics at UCLA. (http://www.ipam.ucla.edu/programs/fg2000/fellowship_100301.html).

Finally, there are opportunities for molecular biology that are both outside the ken of the computer- and information-scientist and scope of this review. The two samples introduced below are certainly not a complete presentation of what is available for molecular biology but may suggest two authors' perspectives.

Integration with clinical informatics

Altman (2000) describes the Stanford Medical Informatics program as the next step in a "post-genome age, [where] the interplay between basic biological data (sequences, structures, pathways, and genetic networks) and clinical information systems is, clearly, critical" (p. 442). The primary concerns for the future (http://bits.stanford.edu/) emphasize robust computing to issues of "information acquisition, storage, retrieval, and management" (p. 442). By outlining six "affinity groups", Altman suggests greater integration (and hence opportunity for computer- and information-science) in

- "image acquisition and analysis (physical systems)
- structural biology and genetics bioinformatics (physical systems)
- biomechanical modeling for macroscopic systems (physical systems)
- computer-assisted interventions and robotics (physical systems)
- data modeling, statistics, and informatics (informatics)
- networked and computer-enabled education (informatics)"

Furthermore Altman (1998, p. 53) suggests that "DNA sequence information and sequence annotations will appear in the medical chart with increasing frequency" which suggests both the ethical issues of making such data publicly accessible but also the computerization issue of what data are stored and how to integrate an expanded data model. Certainly the advancements in controlled vocabularies in clinical informatics can be applied to representing bioinformatic data.

Burge (2002) outlines several challenges:

- "Precise, predictive model of transcription initiation and termination…
- Precise, predictive model of RNA splicing/alternative splicing…
- Precise, quantitative models of signal transduction pathways: ability to predict cellular responses to external stimuli
- Determining effective protein, DNA:protein, protein:RNA and protein:protein recognition codes
- Accurate *ab initio* protein structure prediction
- Rational design of small molecule inhibitors of proteins
- Mechanistic understanding of protein evolution…
- Mechanistic understanding of speciation: molecular details of how speciation occurs
- Continued development of effective gene ontologies – systematic ways to describe the functions of any gene or protein
- Education: development of appropriate curricula for secondary, undergraduate and graduate education"

The porous borders of clinical information systems, medical informatics and bioinformatics, especially in light of Burge's and Altman's predictions of tighter integration of these fields, implies great opportunities for data mining. Medical research and practice have generated tremendous amounts of data, beyond that created by pharmaceutical and biomedical research. For instance, electronic patient records and integrated medical-information systems provides a great warehouse of clinical data online. By mining these data, bio- and other informaticians can detect trends and surprising events from the data, to support informed decision making by clinicians (e.g., evidence-based medicine) and even create "intelligent" system that respond to the data (evidence-based adaptive medicine), to improve health care.

Conclusions

The trend in bioinformatics education is to train computer and information scientists in basic biology and genomics and to emphasize computerized database creation, searching, and parsing of files for particular research projects. This is supported by the growing number of undergraduate and graduate programs in "bioinformatics" throughout the world. In addition, the contents of conferences and monographs demonstrate the results of biologists working with computer and information science. Moreover, there are many training grants available. Therefore, the future is bright for training and hiring new bioinformaticians.

The technology employed in the lab to create new genomic data and to process the data form a regular part of the molecular biologist's work. There are so many advancements to this equipment for general and specialized research that the Internet is the primary means for keeping informed, along with conferences and annual reviews. Merely keeping abreast of the developments is a challenge, which could be supported by selective dissemination of information and other "push" technologies, such as could be provided by centralized biotechnical digital libraries.

Scientists submit and search for records in a heterogeneous computing environment. Therefore, there is considerable opportunity for both computer science and information science to focus on their shared areas of expertise. The literature presented above demonstrates seemingly limitless opportunity for computer- and information science to provide biology with its expertise in machine-centric work. Under this rubric, CS/IS can develop high performance computing, compression algorithms, user interface design, refined similarity measures, metadata applications, interactive graphic user interfaces and 3D modeling in ways that ultimately will integrate sequence and expression data with associated research literature. The results of shared efforts naturally would create a more robust computational environment and further efforts at more flexible data manipulation, which may help answer increasingly sophisticated research questions.

Acknowledgements

The author wishes to thank Connie Chow, Ph. D., Harvard University School of Public Health, and the ARIST reviewers for their helpful comments.

Appendix 1 – Resource Websites

Base4	http://www.cis.udel.edu/~vagrawal/bioinformatics/code/java/base4javabioresources
BIOML	http://bioperl.org/Projects/XML/
bioXML	http://stateslab.bioinformatics.med.umich.edu/
CluSTr	http://www.ebi.ac.uk/clustr
EBI	http://www.ebi.ac.uk/
EBI-InterPro	ftp://ftp.ebi.ac.uk/pub/databases/interpro
EBI-Mutations	http://www.ebi.ac.uk/mutations/index.html
EBI-Stats	http://www3.ebi.ac.uk/services/DBStats/
EBI-Vis	http://industry.ebi.ac.uk/~alan/VisSupp/VisAware/index.html
EPD	ftp://ftp.ebi.ac.uk/pub/databases/epd
GDB	http://www.gdb.org
GeneCard	http://bioinfo.weizmann.ac.il/cards/
GMT	http://www.ebi.ac.uk/~sterk/genome-MOT/MOTgraph.html
GO	http://www.geneontology.org/
GoBASE	http://megasun.bch.umontreal.ca/gobase/gobase.html
HIV	http://hiv-web.lanl.gov/
HGPI	http://www.ornl.gov/sci/techresources/Human_Genome/glossary
HSSP	http://www.sander.ebi.ac.uk/hssp
IMGT	http://imgt.cnusc.fr:8104/texts/info.html
KEGG	http://www.genome.ad.jp/kegg/
MEROPS	http://www.merops.co.uk
NCBI	http://www.ncbi.nlm.nih.gov/
NIBIB	http://www.nibib1.nih.gov/training/coursetable.html
OMIM	http://www3.ncbi.nlm.nih.gov/Omim/
PAI	http://www.ebi.ac.uk/swissprot/hbi/hpi.html
ProDom	http://www.toulouse.inra.fr/prodom.html
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
RasTop	http://www.geneinfinity.org/rastop/gallery.htm
RDP	http://rdp.life.uiuc.edu/index.html
REBASE	http://rebase.neb.com/rebase
STACK	http://www.sanbi.ac.za/Dbases.html
Swiss-Prot	http://www.expasy.ch
Transfac	http://transfac.gbf.de/TRANSFAC/index.html
UniGene	http://www.ncbi.nlm.nih.gov/UniGene
YDP	http://www.proteome.com/databases/

Appendix 2: Bioinformatics Journals

Supplemental to those listed by McCain and Garfield

Biochemie
Biochemistry
Bio-IT World
Biophysical Journal
Biotechniques
British Medical Journal
Cell
Cell Biology Education
Chemical and Engineering News
Drug Discovery Today
Engineering in Medicine and Biology
Genetic Epidemiology
Genetics
Genome Research
IEEE Transactions on Bio-Medical Engineering
Journal of Biological Chemistry
Journal of Molecular Biology
Journal of the American Medical Informatics Association
Journal of Theoretical Biology
Methods in Biochemical Analysis
Methods of Molecular Biology
Molecular Biology and Evolution
Molecular Medicine Today
Nature Biotechnology
Nature Genetics
Nature Structural Biology
New England Journal of Medicine
Nucleic Acid Research
Online Journal of Bioinformatics
Pharmacogenomics
Proceedings of the National Academy of Science
Proteins
Science
Science Next Wave
The Scientist
Trends in Biotechnology

References

Abstract Syntax Notation One (ASN.1). ASN.1 Information. (n.d.). Retrieved December, 21, 2003, from http://asn1.elibel.tm.fr/en/index.htm

Adler, D. A., & Conklin, D. (2000, December). Bioinformatics. *Encyclopedia of Life Sciences*. New York: Nature Publ. Group.

Altman, R. B. (1998). Bioinformatics in support of molecular medicine. In C. G. Chute, (Ed.), *AMIA Annual Symposium*, pp. 53-61.

Altman, R. B. (2000, Sept/Oct.). The Interactions between clinical informatics and bioinformatics: a case study. *Journal of the American Medical Informatics Association, 7* (5), 439-443.

Altman, R. B. (2003, May). The expanding scope of bioinformatics: sequence analysis and beyond. *Heredity 90*(5), 345. Retrieved December, 21, 2003, from http://www.nature.com/cgi-taf/DynaPage.taf?file=/hdy/journal/v90/n5/full/6800225a.html

Altman, R. B., & Dugan, J.M. (2003). Defining bioinformatics and structural bioinformatics. *Methods in Biochemical Analysis, 44*, 3-14.

Altman, R. B., & Koza, J. (1996). A programming course in bioinformatics for computer and information science students. In L. Klein & T. E. Hunter (Eds.), *Pacific Symposium in Biology*. Singapore: World Scientific, pp. 73-84.

Altshul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., & Zhang, Z. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Research, 25*(17), 3389-402.

Andersson, S., Larsson, K., Larsson, M., & Jacob, M. (1999). *Biomathematics: mathematics of biostructures and biodynamics*. Amsterdam: Elsevier.

Apweiler, R., Birney, E., Brazma, A., Brooksbank, C., Cameron, G., Camon, E., Harris, M. et al. (2003). The European Bioinformatics Institute's data resources. *Nucleic Acids Research, 31*, 43-50.

Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y, et al. (2001, Jan 1). Proteome analysis database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Research, 29*(1), 44-48.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25*, 25-29.

Attwood, T. (2000). The Babel of Bioinformatics. *Science, 290*, 471-472.

Attwood, T. K., & Miller, C. J. (2001). What craft is best in bioinformatics. *Computers in Chemistry, 25*, 329-339.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Reading, MA: ACM Press and Addison-Wesley.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Research, 28*, 45-48.

Baldi, P., & Brunak, S. (2001). *Bioinformatics: the machine learning approach*. Cambridge: MIT Press.

Bard, J. (2003, May). Ontologies: Formalising biological knowledge for bioinformatics. *Bioessays, 25*(5), 501-6.

Barnes, M. R., & Gray, I. C. (Eds.). (2003). *Bioinformatics for geneticists*. Hoboken: Wiley.

Basi, G., Clum, R., & Modi, C. (2003, April 1). An Array of Opportunities. *Science Next Wave*. Retrieved December, 21, 2003, from http://nextwave.sciencemag.org/cgi/content/full/2003/04/09/2

Baxevanis, A. D., & Ouellette, B. F. F. (2001). *Bioinformatics: a practical guide to the analysis of genes and proteins*. (2nd ed.). Methods of biochemical analysis, vol. 43. New York: Wiley-Interscience.

Bayat, A. (2002, April). Science, medicine, and the future: Bioinformatics. *British Medical Journal 324*: 1018-1022. Retrieved June, 24, 2003 from http://bmj.com/cgi/reprint/324/7344/1018

Benaim Jalfon, C. (2001). *Analysis of the bioinformatics industry*. Unpublished M. S. Thesis. MIT Sloan School of Management . Cambridge, MA: MIT.

Benoît, G. (2000). Data mining. In Blaise Cronin (Ed.), *Annual Review of Information Science and Technology*, vol. 39. Medford: Information Today.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., & Wheeler, D. L. (2000). GenBank. *Nucleic Acids Research, 28*(1), 15-8.

Benson, G., & Page, R. (Eds.). (2003). *Algorithms in bioinformatics: Third International Workshop, WABI, 2003, Budapest, Hungary, September 15-20, 2003. Proceedings*. Berlin: Springer.

Bergeron, B. P. (2003). *Bioinformatics Computing*. Upper Saddle River, NJ: Prentice Hall.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235-42.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodger, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *European Journal of Biochemistry 80*(2), 319-24.

Bernstein, R. L. (2001). What is bioinformatics? Retrieved December 22, 2003, from http://www.swbic.org/education/bioinfo/

Boguski, M. S. (1999). Biosequence exegesis. *Science, 286*(5439), 453-5.

Böhringer, S., Gödde, R., Böhringer, D., Schulte, T., & Epplen, J. T. (2002). A software package for drawing ideograms automatically. *Online Journal of Bioinformatics 1*, 51-61.

Bowtell, D., & Sambrook, J. (2003). *DNA Microarrays: a molecular cloning manual*. Cold Spring Harbor, NJ: Cold Spring Harbor Laboratory Press.

Brass, A. (2000). Bioinformatics Education – A UK Perspective. *Bioinformatics, 16*(2), 77-78.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellmann, P., Stoeckert, C., Aach, J., Ansorge, W. et al. (2001). Minimum information about a microarray experiment (MIAME) – towards standards for microarray data. *Nature Genetics, 29*(4), 365-371.

Brzeski H. (2002). An introduction to bioinformatics. *Methods of Molecular Biology 187*: 193-208.

Burge, C. (2002). Top ten future challenges for bioinformatics. *Genome Technology, 17*. Retrieved June, 24, 2003 from http://genes.mit.edu/burgelab/topten.htm

Burland, Timothy G. (2001). DNASTAR's Lasergene Sequence Analysis Software. In Stephen Misener & Stephen A. Krawetz, (Eds.). *Bioinformatics: methods and protocols*. Totowa, NJ: Humana, pp. 71-91.

Butte, A. J. (2001). Challenges in bioinformatics: infrastructure, models and analytics. *Trends in Biotechnology, 19*, 159-160.

Calandra, B. (2002, September, 2). Bioinformatics Knowledge Vital to Careers. *The Scientist, 16*(17), 53-54.
Calvanese, D., Lenzerini, M., & Motwani, R. (Eds.). (2003). *Database theory, ICDT, 2003: 9th international conference, Siena, Italy, January 8-10, 2003. Proceedings*. Lecture notes in computer science, 2572. New York: Springer.
Card, S. K., MacKinlay, J. D., & Shneiderman, B. (2002). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann.
Chawla, H. S. (2002). *Introduction to plant biotechnology*. Enfield, NJ: Science Publishers.
Chen, H. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval. An experiment on the Worm Community System. *Journal of the American Society for Information Science, 48*(1),17-31.
Chen, Z. (1993). Let documents talk to each other: A computer model for connection of short documents. *Journal of Documentation 49*(1), 44-54.
Cheung, V. A., Dalrymple, H. L., Narasiman, S., Watts, J., Schuler, G., & Raap, A. K. (1999). Making and reading microarrays. *Nature Genetics, 21* (1 Supplement), 15-19.
Chicurel, M. (2002). "Bioinformatics: Bringing It All Together. *The Scientist 419*, 751-757.
Conn, P. M. (Ed.). (2003). *Handbook of proteomic methods*. Totowa, NJ: Humana Press.
Corruble, V., & Ganascia, J.-G. (1997). Induction and the discovery of the causes of scurvy: a computational reconstruction. *Artificial Intelligence, 91*, 205-223.
Cory, K. A. (1997). Discovering hidden analogies in an online humanities database. *Computers and the Humanities 31*(1), 1-12
Cottle, H. (2001, June 29). Bioinformatics for beginners. *Science Next Wave*. Retrieved December 23, 2003, from http://nextwave.sciencemag.org/cgi/content/full/2001/06/27/
Davies, R. (1988). The creation of new knowledge by information retrieval and classification. *Journal of Documentation 45*(4), 273-301.
de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology 9*(1), 67-103.
Denn, S. O., & MacMullen, W. J. (2002). The ambiguous bioinformatics domain: a conceptual map of information science applications to molecular biology. In *Proceedings of the 65th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*, pp. 556-558.
Dougherty, T. J., & Projan, S. J. (2003). *Microbial genomics and drug discovery*. New York: Marcel Dekker.
Duggan, D. J. (1999). Expression profiling using cDNA microarrays. *Nature Genetics, 21* (1 Supplement), 10-14.
Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.
Dwyer, R. A. (2003). *Genomic Perl: from bioinformatics basics to working code*. Cambridge: Cambridge University Press.
Eisen, M. B. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academic of Science 95*(25), 14863-8.
Eisen, M. B., & Brown, P. O. (1999). DNA arrays for analysis of gene expression. *Methods of Enzymology, 303*, 179-205.
Elkin, P. L. (2003, January). Primer on medical genomics part V: bioinformatics. *Mayo Clinic Proceedings, 78*(1), 57-64.
Ellis, L. (2003a) What is bioinformatics?, 2000-2001. Retrieved June, 24, 2003 from http://www.binf.umn.edu/whatsbinf2000.html
Ellis, L. (2003b). What is bioinformatics?, 2003. Retrieved June, 24, 2003 from http://www.binf.umn.edu/whatsbinf.html
Etzold, T., Ulyanov, A., & Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods of Enzymology, 266*,114-28.
European Bioinformatics Institute. (1999). Data Mining for Bioinformatics – towards in silico biology. Retrieved June, 24, 2003 from http://industry.ebi.ac.uk/datamining99/
Ewens, W. J., & Grant, G. R. (2001). *Statistical methods in bioinformatics: an introduction*. New York: Springer.
Fagan, R., & Swindells, M. (2000). Bioinformatics, target discovery and the pharmceutical/biotechnical industry. *Current Opinion in Molecular Therapeutics, 2*, 655-661.
Fogel, G. B., & Corne, D. W. (2003). *Evolutionary computation in bioinformatics*. San Francisco, CA: Morgan Kaufman.
Frasconi, P. (Ed.). *Artificial intelligence and heuristic methods in bioinformatics*. Amsterdam: IOS Press.
Fuchs, R. (2001 April). From sequence to biology: The impact on bioinformatics. *Bioinformatics, 18*, 505-506.
Funahashi, A., Tanimura, N., Morohashi, M., & Kitano, H. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico 1*, 159-162.

Gardiner, K. (2001). Bioinformatics for biologists. *Trends in Genetics, 17*, 736-737.
Gardy, J. & Brinkman, F. (2003, January 17). The Benefits of Interdisciplinary Research: Our Experiences with Pathogen Bioinformatics. *Science Next Wave*. Retrieved June, 24, 2003 from http://nextwave.sciencemag.org/cgi/content/full/2003/01/15/1
Garfield, E. (1994). Linking literatures: An intriguing use of the citation index. *Current Contents, 21*, 3-5.
Garfield, E. (2002). *Bioinformatics*. Retrieved December, 21, 2003, from http://www.garfield.library.upenn.edu/papers/bio/bioinformatics112002.html
Gascuel, O., & Moret, B. M. E., (Eds.). (2001). *Algorithms in bioinformatics: first international workshop, WABI, 2001, Århus, Denmakr, August, 28-31, 2001. Proceedings.* Lecture notes in computer science, 2149. New York: Springer
Gascuel, O., & Sagot, M-F. (2000). *Computation biology: First International Conference on Biology, Informatics, and Mathematics, JOBIN, 2000, Montpellier, France, May 3-5, 2000*. Lecture notes in computer science, 2066. New York: Springer.
Gatto, J. G. (2003). The Changing Face of Bioinformatics. *Drug Discovery Today 8*, 375-376.
Gene Ontology Consortium. (2001). Creating the Gene Ontology Resource: design and implementation. *Genome Research, 11*, 1425-1433.
Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nature Structural Biology 7*, Supplement 960-3.
Gerstein, M., & Jansen, R. (2000). The current excitement in bioinformatics, analysis of whole-genome expression data: how does it relate to protein structure and function. *Current Opinions in Structural Biology, 10*, 574-84.
Gibas, C., & Jambeck, P. (2001). *Developing Bioinformatics Computer Skills*. Sebastopol: O'Reilly.
Gibson, G., & Muse, S. V. (2002). *A primer of genomic science*. Sunderland, MA: Sinauer.
Gilbert, D. (2001). Free Software in Molecular Biology for Macintosh and MS Windows Computers. In Stephen Misener & Stephen A. Krawetz, (Eds.). *Bioinformatics: methods and protocols*. Totowa, NJ: Humana, pp. 149-184.
Goodfellow, J. M. (Ed.). (1995). *Computer Modeling in Molecular Biology*. Weinheim: VCH.
Goodman, L. (2003, July 15). Making a genesweep: it's official! *Bio-IT World News*, p. 12.
Gordon, M. D., & Dumais, S. (1998). Using latent semantic indexing for literature-based discovery. *Journal of the American Society for Information Science, 49*(8), 674-685.
Gordon, M. D., & Lindsay, R.K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's disease and fish-oil. *Journal of the American Society for Information Science, 47*(2), 116-128.
Grant, A. M., Moshyk, A. M., Kushniruk, A., & Moehr, J.R. (2003). Reflections on an Arranged Marriage between Bioinformatics and Health Informatics. *Methods of Information in Medicine, 42*(2):116-20.
Guerra, C., & Istrail, S. (2000). *Mathematical models for protein structure analysis and design: C. I. M. E summer School, Martina Franca, Italy, July 9-15, 2000*. Lecture notes in computer science, 2666. New York: Springer.
Guerrinia, V. H., & Jackson, D. (2000). Bioinformatics and extended [sic] markup language (XML). *Online Journal of Bioinformatics, 1*, 1-13. Retrieved December 22, 2003, from
Guigó, R., & Gusfield, D. (Eds.). (2002). *Algorithms in bioinformatics: Second International Workshop, WABI, 2002, Rome, Italy, September 17-21, 2002. Proceedings*. Lecture notes in computer science, 2452. New York: Springer.
Gywne, P. (2002, June 14). Informatics: Integrating Information. *Science*. Retrieved June, 24, 2003 from http://recruit.sciencemag.org/feature/advice/informatics.shtml
Harris, M. A., & Parkinson, H. (2003). Conference Report: standards and ontologies for functional genomics: towards unified ontologies for biology and biomedicine. *Comparative and Functional Genomics, 4*, 116-120.
Henry, C. M. (2002, January 7). Careers in Bioinformatics: Field is not significantly affected by economic downturn; qualified people are still hard to find. *Chemical and Engineering News, 79*(1), 47-55.
Hightower, C. (2002, Winter). Guide to selected bioinformatics Internet Resources. *Issues in Science and Technology Librarianship, 33*. Retrieved December, 21, 2003 from http://www.istl.org/istl/02-winter/internet.html
Hillisch, A., & Hilgenfeld, R. (2003). *Modern methods of drug discovery*. Boston: Birkäuser.
Howard, M. (2000). The bioinformatics gold rush. *Scientific American, 283*, 58-63.
Hucka, M., Finney, A., Sauro, H. M., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics, 19*, 524-531.
Human Genome Project Information. (n.d.). Retrieved June, 24, 2003 from http://www.ornl.gov/hgmis/faq/seqfacts.html

Ibba M. (2002). Biochemistry and bioinformatics: when worlds collide. *Trends in Biotechnology, 20*, 53-4.
Jain, E. (2003). Practical bioinformatics. *Pharmacogenomics 4*(2), 119-21.
Jajuga, K., Sokolowski, A., & Bock, H-H. (Eds.). (2002). *Classification, clustering and data analysis: recent advances and applications*. New York: Springer.
Jamison, D. C. (2003). *Perl programming for biologists*. Hoboken: Wiley-Liss.
Jenders, R., Sideli, R., & Hripcsak, G. Introduction to Medical Informatics. Retrieved June, 24, 2003 from http://www.cpmc.columbia.edu/edu/textbook
Jenson, D. (2002, October 18). Job Market Hype: Media Misinformation Puts a Spin on Biotech Job Market. *Science Next Wave*. Retrieved June, 24, 2003 from http://nextwave.sciencemag.org/cgi/content/full/2002/10/17/2
Jenssen, T-K, Öberg, L. M. J., Andersson, M. L., & Komorowski, J. (2001). Methods for large-scale mining of networks of human genes. In R. Grossman, V. Kumar, and J. Han (Eds.). *Proceedings SIAM Conference on Data Mining (SDM, 2001)*. Retrieved December, 21, 2003 from http://www.siam.org/meetings/sdm01/pdf/sdm01_10.pdf
Jiang, T., Xu, Y, Zhang, M. Q. (2002). *Current topics in computational molecular biology*. Cambridge: MIT Press.
Juvan, P. (2001). Web-enabled knowledge-based analysis of genetic data. Lecture Notes in Computer Science, 2119. Berlin: Springer, pp. 113-119.
Kalow, W., Meyer, U. A., & Tyndale, R. F. (2001). *Pharmacogenomics*. New York: Marcel Dekker.
Kanehisa, M. & Bork, P. (2003, March). Bioinformatics in the post-sequence era. *Nature Genetics 33* Supplement, 305-10.
Kantardzic, M. (2003). *Data Mining: concepts, models, methods, and algorithms*. Piscataway, NJ: IEEE Press/Wiley-Interscience.
Karp, D. D., Paley, S., & Zhu, J. (2001). Database verification studies of Swiss-Prot and GenBank. *Bioinformatics 17*(6), 526-532.
Kasabov, N. K. (2003). *Evolving connectionist systems: methods and applications in bioinformatics, brain study and intelligent machines*. New York: Springer.
Keim, D. A. (1997). Visualization Techniques in Exploring Databases. Invited tutorial. International Conference on Knowledge Discovery in Databases (KDD'97). Newport Beach, CA.
Kitano, H. (2003). A graphical notation for biochemical networks. *Biosilico, 1*, 169-176.
Kohane, I. S., Kho, A. T., & Butte, A. J. (2003). *Microarrays for an integrative genomics*. Cambridge, MA: MIT Press.
Kohlatkar, P. (2002). Biocomputing at Singapore's Top R&D Institutes. *Science Next Wave*. Retrieved from http://nextwave.sciencemag.org/car.dtl, December, 22, 2003.
Korf, I., Yandell, M., & Bedell, J. (2003). *BLAST*. Farnham: O'Reilly.
Koski, T. (2001). *Hidden Markov Models for Bioinformatics*. Computational biology series, vol., 2. Dordrecht: Kluwer Academic.
Kossida, S., Tahri, N., & Daizadeh, I. (2002, December). Bioinformatics by Example: From Sequence to Target. *Journal of Chemical Education 79*,1480-1485.
Krane, D. E., & Raymer, M. L. (2003). *Fundamentals of bioinformatics*. San Francisco: Benjamin Cummins.
Krawetz, S. A., & Womble, D. D. (2003). *Introduction to bioinformatics: a theoretical and practical approach*. Totowa, NJ: Humana Press.
Lacroix, Z., & Critchlow, T. (2003). *Bioinformatics: managing scientific data*. San Francisco: Morgan Kaufmann.
Laender, A. H. F., & Oliveira, A. L. (2002). (Eds). *String processing and information retrieval: 9th International Symposium, SPIRE, 2002, Lisbon, Portugal, September 11-13, 2002. Proceedings*. Lecture notes in computer science, 2476. New York: Springer.
Lander, E. S. (et al.) (2001). Initial sequencing and analysis of the human genome. *Nature 409*, 860-921.
Lanzenberger, M., Miksch, S., Ohmann, S., & Popow, C. (2003). Applying information visualization techniques to capture and explore the course of cognitive behavioral therapy. *Proceedings of the Symposium on Applied Computing*.
Lengauer, T. (2002). *Bioinformatics – from genomes to drugs*. Methods and principles in medicinal chemistry, vol. 14. Weinheim: Wiley-VCH.
Lesk, A. M. (2002). *Introduction to bioinformatics*. New York: Oxford.
Lesk. A. M. (Ed.) (1988). *Computational Molecular Biology: sources and methods for sequence analysis*. Oxford: Oxford Univ. Press.
Leszczynski, J. (Ed.). (1999). *Computational Molecular Biology*. Theoretical and computational chemistry, vol. 8. Amsterdam: Elsevier.
Lim, H. A. (2002). *Genetically yours: bioinforming, biopharming, biofarming*. River Edge, NJ: World Scientific.

Lindsay RK, Gordon MD. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science, 50*(7): 574-587.
Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., & Lockhard, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics, 21*(1), 20-24.
Lu, W., Janssen, J., Milios, E., & Japkowicz, N. (2001). Node similarity in networked information spaces. Technical Report CS-2001-03, Dalhousie University. Retrieved December 23, 2003, from http://www.cs.dal.ca/research/techreports/2001/CS-2001-03.html.
Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001, May). What is bioinformatics? A proposed definition and overview of the field. *Method Inform Med 40*, 346-58.
Lynn, D. J., Lloyd, A.T., & O'Farrelly, C. (2003, April). Bioinformatics: implications for medical research and clinical practice. *Clinical Investigations in Medicine, 26*(2), pp. 70-74.
Maher, R. (2001). BioML2SVG. In *Stirring XML: Visualisations in SVG*. Retrieved December 22, 2003, from http://www.svgopen.org/2003/papers/StirringXml-VisualisationsInSVG/#S3.
McDonald, C. J. (2001). Hickham 2000: the maturation of, and linkages between, medical informatics and bioinformatics. *Journal of Laboratory and Clinical Medicine, 138*, 359-366.
Mewes, H.-W., Seidel, H., & Weiss, B. (Eds.). (2003). *Bioinformatics and genome analysis*. Berlin: Springer.
Miller, C., Gurd, J., & Bass, A. (1999). A RAPID algorithm for sequence database comparison: application to the identification of vector contamination in the EMBL databases. *Bioinformatics 15*(2), 111-21.
Miller, C.J., Attwood, T.K. (2003, Feb.). Bioinformatics goes back to the future. *Nature Review of Molecular Cell Biology 4*(2), pp. 157-62.
Misener, S. & Krawetz, S. A., (Eds.). (2000). *Bioinformatics: methods and protocols*. Methods in molecular biology, vol. 132. Totowa, NJ: Humana.
Montgomery, S. (2003). Java for Bioinformatics. Retrieved December, 23, 2003, from http://www.onjava.com/pub/a/onjava/2003/09/24/java_bioinformatics.html
Mount, D. W. (2001). *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
National Center for Biotechnical Information (NCBI). (2003, May 1). What is Bioinformatics? Retrieved June, 24, 2003, from http://www.ncbi.nlm.nih.gov/Education/
Nishikawa, K. (2002, May). Information Concept in Biology. *Bioinformatics, 18*, 649-651.
Nucleic Acids Research (NAR). (2002). Annual database issue. Retrieved June, 24, 2003 from http://nar.oupjournals.org/content/vol30/issue1/
Orengo, C. A. (1999). CORA – topological fingerprints for protein structure families. *Protein Science 8*(4), 699-715.
Orengo, C. A., & Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Methods of Enzymology, 266*, 617-35.
Orengo, C., Jones, D., & Thornton, J. (2003). *Bioinformatics: genes, proteins, and computers*. Oxford: BIOS Scientific/New York: Springer.
Ouzounis, C. (2002, March). Bioinformatics and the Theoretical Foundations of Molecular Biology. *Bioinformatics, 18*, 377-378.
Paris, C. G. (1997). Chemical structure handling by computer. In M. E. Williams, (Ed.), *Annual Review of Information Science and Technology 32*. Medford, NJ: ASIS, pp., 271-337.
Paris, G. (2002, November). Mining bioinformatics databases. In J. Hurd (Moderator), *The role of "unpublished" research in the scholarly communication of scientists: digital preprints and bioinformation databases*. Panel presentation at the 65th Annual Meeting of the American Society for Information Science and Technology, Philadephia.
Pearson, W. R. (2001). Training for bioinformatics and computational biology. *Bioinformatics 17*(6), 761-762.
Pearson, W. R., Lipman, D. L. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Adcademy of Science 85*(8), 2444-8.
Perner, P. (Ed.). (2002). *Advances in data mining: applications in E-commerce, medicine, and knowledge management*. Lecture notes in computer science, 2394. New York: Springer.
Pevsner, J. (2003). *Bioinformatics and functional genomics: a short course*. New York: Wiley-Liss.
Poinçot, P., Lesteven, S., & Murtagh, F. (2000). Maps of information spaces: assessments from astronomy. *Journal of the American Society for Information Science, 51*(12), 1081-1089.
Priami, C. (Ed.). (2003). *Computational methods in systems biology: first international workshop, CMSB, 2003, Rovereto, Italy, February, 24-26, 2003. Proceedings*. Lecture notes in computer science, 2602. New York: Springer.
Primrose, S. B. (2003). *Principles of genome analysis and genomics*. Maldon, MA: Blackwell.
Raychaudhur, S. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of bio-

medical literature. *Genome Research, 12*, 203-214.
Rhodes, P. J., Bergeron, R. D., & Sparr, T. M. (2002). Database support for multisource multiresolution scientific data. *Proceedings of SOFSEM, 2002*. New York: Springer.
Rikken, F. (1998). *Adverse drug reactions in a different context: A scientometric approach towards adverse drug reactions as a trigger for the development of new drugs.* (Ph. D. Dissertation). Groningen: Rijksuniversiteit, Groningen.
Rodriguez-Tomé, P. (2001). Resources at EBI. In Stephen Misener & Stephen A. Krawetz, (Eds.). *Bioinformatics: methods and protocols*. Totowa, NJ: Humana, pp. 313-335.
Roos, D. S. (2001). Bioinformatics – trying to swim in a sea of data. *Science, 291*, 1260-1261.
Rost, B., Honig, B., & Valencia, A. (2002, July). Bioinformatics in structural genomics. *Bioinformatics 18*, 897.
Sander, C. (2002). The Journal Bioinformatics, key medium for computational biology. *Bioinformatics 18*, 1-2.
Sansom, C. E., & Smith, C. A. (2000). Computer applications in biomolecular sciences. Part 2: bioinformatics and genome projects. *Biochemical Education, 28*, 127-131.
Schachter, B. (2002, June 12). Informatics Moves to the Head of the Class. *Bio-IT World*. Retrieved June, 24, 2003 from http://www.bio-itworld.com/archive/061202/class.html
Schlichting, I., & Egner, U. (Eds.). (2001). *Data mining in structural biology*. New York: Springer.
Schroeder, M., Gilbert, D., van Helden, J., & Noy, P. (2001). Approaches to visualization in bioinformatics: from dendrograms to Space Explorer. *Information Sciences, 139*(1-2), 19-57
Schuler, G. D., Epstein, J. A., Ohkawa, H., & Kans, J. A. (1996). Entrez: molecular biology database and retrieval system. *Methods of Enzymology, 266*, 141-62.
Searls, D. B. (2000). Bioinformatics tools for whole genomes. *Annual Review of Genomics and Human Genetics, 1*, 251-279.
Sensen, C. W. (Ed.). (2002). *Essentials of Genomics and Bioinformatics*. Weinheim: Wiley-VCH.
Smalheiser, N. R., & Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neuroscience Research Communications, 15*(1), 1-9.
Smalheiser, N. R., & Swanson, D. R. (1996). Linking estrogen to Alzheimer's Disease. An informatics approach. *Neurology, 47*(3), 809-810.
Smith, D. W. (Ed.). (1993). *Biocomputing: informatics and genome projects*. San Diego: Academic.
Stapley, B. J., & Benoît, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium on Biocomputing 5*, 538-549.
Swanson, D. R. & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91*(2), 183-203.
Swanson, D. R. & Smalheiser, N. R. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. *Library Trends 48*(1), 48-59.
Swanson, D. R. (1991). Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghavan (Eds.). *SIGIR '91*, (pp., 280-289). New York: Association for Computing Machinery.
Swanson, D. R. (1993). Intervening in the life cycles of scientific knowledge. *Library Trends 41*(4), 606-631.
Swanson, D. R., Smalheiser N. R., & Bookstein A. (2001, August). Information discovery from complementary literatures: categorizing viruses as potential weapons. *Journal of the American Society for Information Science, 52*(10): 797-812.
Tatusova, T. A., Karsch-Mizrachi, I., & Ostell, J. A. (1999). Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics 15*(7-8), 536-43.
Taylor, M.D., Mainprize, T.G., Rutka, J.T. (2003, April). Bioinformatics in neurosurgery. *Neurosurgery, 52*(4), 723-31.
Taylor, L. P. (2000). Ligand binding domain of nuclear receptors. Retrieved December 23, 2003, from http://www-personal.umich.edu/~lpt/mr.htm.
Tessier, D C., Benoît, F., Rigby, T., Hogues, H., van het Hoog, M., Thomas, D. T., & Brousseau, R. (2000). A DNA Microarray fabrication strategy for research laboratories. In C. Sensen (Ed.), *Essentials of Genomics and Bioinformatics*.
Tisdall, J. D. (2001). *Beginning Perl for bioinformatics*. Sebastopol, CA: O'Reilly.
Toronen, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., & Golub, T.R. (1999). Analysis of gene expression data using self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academic of Science 96*(6), 2907-12.
Tözeren. A. (2004). New biology for engineers and computer scientists. Upper Saddler River, NJ: Prentice Hall.
Tsigelny, I. F. (2002). *Protein structure prediction: bioinformatics approach*. La Jolla, CA: International University Line.

University of California, Davis, Genome Center. (2003). What is genomics? Retrieved June, 24, 2003 from http://genomics.ucdavis.edu/what.html
University of North Carolina, Chapel Hill. (2003). Bioinformatics Journal Club. http://ils.unc.edu/bmh/bioinfo/Bioinformatics_Programs_Brief_7-13-03.htm
Valdés-Peréz, R. E. (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence 107*(2), 335-346.
Valencia, A. (2002, December). Bioinformatics: Biology by Other Means. *Bioinformatics 18*, 1551-1552.
Van Haren, K. (2002 May, 24). Bioinformatics Funding Boost Means More Science Training. *Science Next Wave*. Retrieved June, 24, 2003 from http://nextwave.sciencemag.org/cgi/content/full/2002/05/23/1
Venter, J. C. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-51.
Wang, J. T. L., Wu, C. H., Wang, P. (Eds.). (2003). *Computational biology and genomic informatics*. River Edge, NJ: World Scientific.
Ware, C. (2000). *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufmann.
Watkins, K. J. (2001, Feb. 19). Bioinformatics: making sense of information mined from the human genome is a massive undertaking for the fledgling industry. *Chemical & Engineering News, 79*(8), 25-45.
Watson, J. D., & Crick, F, H. C. (1953). A structure for deoxyribose nucleic acid. *Nature, 171*, 737.
Weeber, M., Vos R, Klein H, & de Jong-van den Berg, L. T. W. (2001). Using concepts in literature-based discovery: simulating Swanson's Raynaud- fish-oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science, 52*(7): 548-557.
Weeber, M. (2001). Literature-based Discovery in Biomedicine. (Ph. D. Dissertation). Groningen: Rijksuniversiteit Groningen.
Westhead, D. R., Parish, J. H., & Twyman, R. M. (2002). *Bioinformatics*. Oxford: BIOS.
Wilson, C. A., Kreychman, J., & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology, 297*(1), 233-49.
Winter, P. S., Hickey, S. O. & Fletcher, H. L. (2002). *Genetics*. (2nd ed.). Oxford: Bios Scientific.
Wise, M. J. (2000). Protein annotator's assistant: a novel application of information retrieval techniques. *Journal of the American Society for Information Science, 51*(12), 1131-1136.
Ye, N. (2003). *Handbook of data mining*. Mahwah, NJ: Lawrence Erlbaum.
Zauhar, R. J. (2001, March). University bioinformatics programs on the rise. *Nature Biotechnology,19*, 285-286.
Zdobnov, E. M., Lopez, R., Apweiler, R., & Etzold, T. (2002). Using the molecular biology data. In C. Sensen (Ed.), *Biotechnology: Genomics and Bioinformatics* (vol. 5b). New York: Wiley-VCH, pp., 265-284.
Zdobnov, E. M., Lopez, R., Apweiler, R., & Etzold, T. (2002, Feb.). The EBI SRS server – recent developments. *Bioinformatics, 18*(2), 368-373.
Zimmerman, K-H. (2003). *An introduction to protein informatics*. Boston: Kluwer Academic.