Genomic and Proteomic Databases: Foundations, Current Status and Future Applications

Shamkant B. Navathe
College of Computing, Georgia Institute Of Technology, USA
sham@cc.gatech.edu

Upen Patil
School of Biology/ Bioinformatics Program, Georgia Institute Of Technology, USA
udpatil@yahoo.com

Wei Guan
College of Computing, Georgia Institute Of Technology, USA
wguan@cc.gatech.edu

In this paper we have provided an extensive survey of the databases and other resources related to the current research in bioinformatics and the issues that confront the database researcher in helping the biologists. Initially we give an overview of the concepts and principles that are fundamental in understanding the basis of the data that has been captured in these databases. We briefly trace the evolution of biological advances and point out the importance of capturing data about genes, the fundamental building blocks that encode the characteristics of life and proteins that are the essential ingredients for sustaining life. The study of genes and proteins is becoming extremely important and is being known as genomics and proteomics, respectively. Whereas there are numerous databases related to various subfields of biology, we have maintained a focus on genomic and proteomic databases which are the crucial stepping stones for other fields and are expected to play an important role in the future applications of biology and medicine. A detailed listing of these databases with information about their sizes, formats and current status is presented. Related databases like molecular pathways and interconnection network databases are mentioned, but their full coverage would be beyond the scope of a single paper. We comment on the peculiar nature of the data in biology that presents special problems in organizing and accessing these databases. We also discuss the capabilities needed for database development and information management in the bioinformatics arena with particular attention to ontology development. Two research case studies based on our own research are summarized dealing with the development of a new genome database called Mitomap and the creation of a framework for discovery of relationships among genes from the biomedical literature. The paper concludes with an overview of the applications that will be driven from these databases in medicine and healthcare. A glossary of important terms is provided at the end of the paper.

Journal of Computing Science and Engineering, Vol. 1, No. 1, September 2007, Pages 1–30.

Categories and Subject Descriptors: H.2.1 [**Database Management**]: Logical Design; J.3 [**Life and Medical Science**]: Medical information systems

General Terms: Database, Bioinfomatics

Additional Key Words and Phrases: Database management, data management, genomic database, proteomic database, bioinformatics, biological database, genome, proteome, health-care, medical application

1. BIOINFORMATICS: AN AMALGAMATION OF MULTIPLE DISCIPLINES

The biological science studies the phenomenon of life and encompasses an enormous variety of information. This wealth of information that has been generated, classified, and stored for centuries has only recently become a major application of database technology. The first genome of RNA bacteriophage MS2, was sequenced in 1976, in a truly heroic feat of direct determination of an RNA sequence. In those early years when scientists were just beginning to understand the complexity of viral genetic material, sequencing the 3,569 nucleotides long sequence of viral bacteriophage MS2 helped biologist in establishing the complete primary chemical structure of this viral genome [Fiers et al. 1976] as the first successful effort of this kind. However, it was not until August of 1995 that the complete genome sequence of the parasitic bacterium haemophilius influenza, ushered the era of real genomics, the study of complete genomes of cellular organisms [Fleischmann et al. 1995].

The U.S. government launched the Human Genome Project [Pearson and Soll 1991] in 1988 with the hope of sequencing the entire genome by 2005. However, with the major advances in the sequencing machinery, scientists in the private sector in conjunction with the HGP achieved the long desired goal already in 2002 by obtaining the sequence of the 3 billion base pairs making up the human genome. This sequenced human genome did not actually belong to a single human being but was in fact obtained from many volunteers because all humans share the same basic set of genes and other DNA regions, so this "reference" sequence represents every person. This landmark was the start of the post genomic era. Currently it is widely accepted that only powerful computational tools can achieve identification of protein coding genes in the genome sequence and determination of protein functions encoded by such genes with a variety of experimental approaches from the arsenals of biochemistry, molecular biology, genetics and cell biology [Koonin and Galperin 2003]. Thus deciphering the evolutionary history of life and maybe in future to manipulate and induce favorable evolution is the fundamental task of biology.

Bioinformatics is a field, which studies the information content of life. Bioinformatics has been defined as a combination of mathematics, computer science and molecular biology to analyze large scale genomic data. Computers have become an essential tool in biology to gather, store and analyze data, which ranges from research articles to complex metabolic pathways. According to NIH definition[1]: Bioinformatics is defined as "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data". They define another related term, Computational Biology as

[1]http://www.bisti.nih.gov/CompuBioDef.pdf

Table I. Classification of Organisms.

Aristotle (4th century B.C.)	Classified organisms as Plants and Animals
Ernst Haeckel	Promoted genealogy of life, as analogous to a tree. Classified organisms as Plants, Animals and Microbes.
Current	Concept of three domains of life based on rRNA studies making the tree of life. Organisms are classified into Eucarya, Archaea and Bacteria.

"the development and application of data-analytical and theoretical methods, and mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems". While the terms bioinformatics and computational biology get used interchangeably, the former is geared more toward the development of algorithms for analysis of biological data while the latter is concerned with the discovery of new biological knowledge by applying computing resources to large scale modeling and simulation coupled with experimental data. System Biology is viewed from multiple perspectives: one of them [Kitano 2001] considers it as a discipline that is trying to study the interactions between components of biological systems that gives rise to the structure and function of the systems such as tens of thousands of genes and proteins working together in interconnected networks to orchestrate the chemistry of life. The large volume and range of molecular, biochemical, genetic, anthropological and medical information has given rise to a very large variety of databases due to a phenomenal number of organisms produced in nature - we summarize briefly the entire evolution history of organisms (see Table I). Bioinformatics knowledge has become a necessity for any laboratory in developmental genetics. The future research advances in computational biology, molecular biology, genomic medicine as well as pharmacogenomics have now become critically dependent on what information such databases will contain including their data organization, accessibility and connectivity to other related databases [François and Peer 2001].

Currently, the people working in this field in most cases have training either in biology or computer science, but not both. After complementing their skill sets with the missing knowledge, they are gradually getting ready to deal with the problems in computational biology. This paper is intended to give a broad survey of the basic biological conceptual foundation necessary for understanding this field, the state of the art of databases and related tools available to the biologists today, and the problems that confront the database researcher in helping the biologists. We end the paper with several important upcoming application areas to which the ever-growing data in these databases may be applied.

2. BIOLOGY CONCEPTS AND PRINCIPLES

The field of biology is vast and it is impossible to summarize the basic concepts in a couple of pages. What is presented below are few terms essential for understanding the complexity of biological information, its representation and management. Some terms appear in the glossary at the end of the paper. Scientists have tried to classify organisms since the time of Aristotle. Darwin in the 17th century proposed

the fundamental ideas of natural selection and evolution. Table 1 summarizes the changing view of the tree of life.

2.1 Genomes

The term genome refers to the totality of the genetic code present in the cells of an organism. The genetic code is in the form of the nucleotides A,T,G,C which respectively stand for Adenine, Thymine, Guanine, Cytosine, in the form of a double helix; the protein information is made up of sequences containing 20 different amino acids, eventually forming a 3-D complex protein molecule.

Genomics

In a cell, information flows from DNA to RNA to Protein. There exists a need to develop a mechanistic understanding of protein speciation and understanding the role of genome in transcription and to develop precise models of where and when the initiation and termination takes place for transcription. Developing precise models of alternative RNA splicing (In RNA splicing introns are removed and exons are joined together and when a splicing signal in an intron is hidden by a regulatory protein, then the process is called alternative splicing.) and signal transduction pathways (Series of signals passed through receptors in the cell membrane to activate transcription are called signal transduction pathways) is essential along with determining mechanistic understanding of protein evolution by protein:DNA, protein:RNA, protein:protein recognition codes, for an accurate *ab initio* prediction of protein structure. To understand such complex models of cellular functions, of transcription and translation, there is a need to have models generated from evaluating numerous gene and protein sequences to compare and reconstruct the earliest stages of evolution. More details on these and subsequent biological concepts are outside the scope of this paper and the reader is advised to consult [Snyder and Champness 2007; Lewin 2007].

Genomics consists of two component areas:

(1) Structural genomics.
(2) Functional genomics.

Scientists are currently using a combination of functional genomics approaches and microarray analysis techniques to identify and clone human genes. Functional genomics is currently considered a major driving force behind such impending revolution in the field of genomics. Structural genomics mainly deals with approaches related to traits which are controlled by one or only a few genes, and mostly ends up providing information related to the location of a gene or genes in the genome.

Such gene position information is an essential preliminary step; functional genomics helps us further analyze the interrelationships and interactions between many genes to help us understand expression of certain traits and the role of certain genes in expressing them.[2] Together, with this functional genomics and structural genomic information, scientists will be well equipped for efficiently creating species with exact combinations of traits. Such new genes introduced into a species through genetic engineering are called as transgenes. For example, functional and structural genomic knowledge can potentially play an important role in producing

[2] http://www.cimmyt.org/

special variety of plants that can optimize the yield potential under any given set of conditions; this could potentially be an answer to the growing demand for food.

Due to the fact that functionality in genome and proteins is conserved, the idea of comparative genomics arose with a potential to solve the mystery of the Origin of Life. Comparative genomics started of with help of viral genomes and proteins, which in turn was a blessing in disguise [Koonin and Galperin 2003]. Due to the difficulty in understanding the sequence conservation in viral proteins, certain crucial approaches of sequence comparison had to be laid down.

Always compare protein sequences, rather than nucleotide sequences

Rely on multiple, rather than pairwise comparisons.

Search for conserved patterns or motifs in multiple sequences

Try to visualize potential relationships in sequences or structures

Comparative genomics is informative in principle, especially in homologs (Homologs have common origins but may or may not have common activity) but lacks to shed light on evolutionary distances and similarity seen between vertebrates and bacteria. Sequence similarity not only exists between organisms with common ancestry but also can exist in convergence from unrelated sequences in which only a limited similarity is observed and is poorly supported by current search techniques. Thus, whenever statistically significant sequence or structural similarity between proteins or protein domains is observed, it indicates their divergent evolution from a common ancestor or evidence of homolog [Koonin and Galperin 2003].

Homologs can be of two types, namely, Orthologs and Paralogs, and crucial to the understanding of evolutionary relationships between genomes and gene functions. This information is also essential in understanding the concepts behind some of the databases e.g. COG [Tatusov et al. 2003]. Another evolutionary concept, which needs to be understood while dealing with genomes are standard (vertical) vs. horizontal (lateral) gene transfer and single nucleotide polymorphism (SNP).

Conservation of function

Functions which are important for a cell's existence are always preserved; or in other words, if there is a mutation in the genes of important functions then those cells do not always survive. But on the other hand, because of the mutation, if the same important function can be achieved in a much more efficient way, then those cells tend to thrive more than their ancestors in the same environment. Disease causing mutations are those that have been shown to be linked to a disease with high probability. Research on Drosophila blue cheese (bchs) gene has shown that mutation can lead to a neuronal apoptosis with the age-dependent formation of protein aggregates throughout the neurons of the central nervous system and thus leading to a neuron-degenerative disease [Finley et al. 2003].

[3]http://www.ddbj.nig.ac.jp/
[4]http://www.ncbi.nlm.nih.gov/Taxonomy/
[5]http://www.ncbi.nlm.nih.gov/PubMed/
[6]http://www.ncbi.nlm.nih.gov/omim/

Table II. Current Genomic and Proteomic Databases.

Database	Format	Size
Nucleotide Databases		
Genbank [Benson et al. 2007]	Abstract Syntax Notation 1 (ASN.1) syntax	GenBank (Aug. 2006) contains 61,132,599 entries totaling 65,369,091,950 base pairs
European Molecular Biology Laboratory (EMBL) [Kulikova et al. 2007]	-Flat file format -A more syntax-oriented structure adopted	EMBL release 91 (May 2007) contains 197,361,640 entries and 170,766,876,848 nucleotides
DNA Data Bank of Japan (DDBJ[3])	-Flat file format	DDBJ release 70 (Jul. 2007) contains 72,801,67 enteries and 76,788,510,646 nucleotides
Nucleotide Structure Database		
Nucleic acid database (NDB) [Berman et al. 1992]	-PDB and mmCIF format	NDB (Jun. 2007) contains 3,557 Structures
Protein Databases		
SWISSPROT/TrEMBL (ExPASy) [Boeckmann et al. 2003]	-Currently flat file -Future Relational Database system with XML format	UniProtKB/Swiss-Prot release 54.1 (Aug. 2007) contains 277,883 sequence entries, comprising 101,975,253 amino acids abstracted from 158,679 references UniProtKB/TrEMBL release 37.1 (Aug. 2007) contains 4,754,787 sequence entries, comprising 1,543,116,088 amino acids
Protein International Resource (PIR) [Wu et al. 2002]	-Flat files with XML format	PIR-PSD final release 80.00 (Dec. 2004) contains 283,416 entries
Protein Structure and Interrelationship Databases		
Protein Data Bank (PDB) [Weissig et al. 2000]	A collection of mmCIF data files	It has a total of 45,506 molecular type files, 27,984 structure factor files, 3,601 NMR restraint files (Aug. 2007)
Structural Classification of Proteins (SCOP) [Andreeva et al. 2004]	Tightly linked hypertext documents	SCOP release 1.71 (Oct. 2006) contains 27,599 PDB entries, 75,930 domains, 971 folds, 1,589 superfamilies and 3,004 families
CATH [Pearl et al. 2003]	-Flat file format	CATH release 3.1.0 (Jan. 2007) contains 30,028 PDB entries, 93,885 domains, and 63,453 chains
Families of Structurally Similar Proteins (FSSP) [Holm et al. 1992]	-Flat file format -A more syntax-oriented structure adopted	FSSP current release (Nov. 2004) has 2,860 entries
Restriction Enzyme Database (REBASE) [Roberts et al. 2007]	-Flat file format	It contains 3,805 biochemically-characterized restriction enzymes (Sep. 2006)

Protein Domain, Motif & Family Databases		
PROSITE (ExPASy) [Hulo et al. 2006]	ASCII text file -Prosite.dat is a computer readable file -Prosite.doc contains textual information	PROSITE release 20.17 (Jul. 2007) contains 1,489 documentation entries that describe 1,319 patterns, 739 rules and 763 profiles/matrices
BLOCKS Database [Henikoff et al. 2000]	Data files of the Blocks database are disseminated as ASCII text files.	BLOCKS version 14.2 (Mar. 2006) consists of 29,767 blocks representing 6,149 groups documented in InterPro 12.0 keyed to SWISS-PROT 48.3 and TrEMBL 31.3 obtained from the InterPro server
Protein Families Database (PFAM) [Finn et al. 2006]	Traditional- Structure of one directory of text files for each family. Also has a PfamRDB, a MySQL relational database	PFAM version 22.0 (Jul. 2007) contains 9,318 families, 73.23% of all proteins in Pfamseq contain a match to at least one Pfam domain. 50.79% of all residues in the sequence database fall within Pfam domains
Protein Fingerprints Database (PRINTS) [Attwood 2002]	Former-ASCII text Current-Relational database with a new display	PRINTS release 38.1 (May 2007) contains 1,904 entries, encoding 11,451 individual motifs
Simple Modular Architecture Research Tool (SMART) [Letunic et al. 2006]	-Relational database	SMART current release shows a total of 726 domain families
Clusters of Orthologous Groups (COG) [Tatusov et al. 2003]	Tightly linked hypertext documents	COGs are delineated by comparing protein sequences encoded in 68 complete genomes and representing 14 major phylogenetic lineages
Conserved Domain Database(CDD) [Marchler-Bauer et al. 2005]	ASN.1 syntax	CDD currently contains domains from Smart and Pfam and COG
Species specific databases		
Saccharomyces Genome Database (SGD) [Wood et al. 2002]	- A collection of Locus Pages [Wood et al. 2002] for Yeast	It contains 8,001 genomic entries, including 6,609 ORFs, 382 long terminal repeats, 299 tRNA, etc.; 14,650 GO annotations on 4,214 gene products of S. cerevisiae (Sep. 2007)
FlyBase [Crosby et al. 2007]	- Flat file format	Flybase current release (Aug. 2007) contains 184,745 references, 78,122 research papers, 35,866 abstracts, 85,022 fly stocks, 870 fly images, etc.
Mouse Genome Database (MGD) [Blake et al. 2006]	-Relational database implemented in Sybase	Currently it contains 30,881 genes, of which 27,617 genes with sequence data, 19,580 genes with protein sequence information, 16,572 genes with GO annotations, 15,849 mouse/human orthologies, 15,532 mouse/rat orthologies, 94,891 reference, etc.

Human Genome Database (GDB) [Pearson and Soll 1991]	An object-oriented data model was used to implement the system, data arranged in a hierarchy of object classes	Currently it contains 6,458,156 objects in HGD, including: 209,070 alleles, 103 gene families, 507 gene products, 4,915,369 genomic segments, 4,537 maps, etc.
European mutant mouse pathology database (Pathbase) [Schofield et al. 2004]	- Images	It currently holds over 1,000 images of lesions from mutant mice and their inbred backgrounds
Specialized Genomic (Gene and Protein) Databases		
MitoMap [Kogelnik et al. 1998]	ASN.1, Current: Relational database.	Mitochondrial Genome: Total number of nucleotides per mitochondrion : 16,569
Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa et al. 2006]	Integration is at the level of data entries in different databases, retrieved uniformly with links.	KEEG release 43.0 (Jul. 2007) contains 56,987 pathways generated from 333 reference pathways, 8,049 hierarchies generated from 44 reference hierarchies, 10,222 KO groups, 2,603,477 genes, 14,877 compounds, 6,572 drugs, 10,972 glycans, 7,217 reactions, and 7,251 reactant pairs
The Encyclopedia of Escherichia coli Genes and Metabolism (EcoCyc) [Keseler et al. 2005]	An object-oriented data model was first used to implement the system, with data stored on Ocelot, a frame knowledge representation system. Data arranged in a hierarchy of object classes.	EcoCyc release 11.5 (Aug. 2007) contains 224 pathways, 1,615 reactions, 1,342 enzymes, 224 transporters, 4,471 genes, 4,413 gene product summaries, 3,187 transcription units, and 16,154 citations
Transcription Factor Databases		
Gene Transcription Factor Database (TRANSFAC) [Matys et al. 2003]	- Flat file format	TRANSFAC release 11.1 (Apr. 2007) contains 9,621 factors, 19,114 sites, 24,340 factor-site links, 19,338 genes, 16,884 chip-chip fragments, 821 matrices, and 14,783 references
Transcription Regulatory Regions Database (TRRD) [Kolchanov et al. 2002]	- Relational database	TRRD current release (Sep. 2005) contains 2,334 genes, 14,407 expression patterns, 14 locus control regions, 3,490 regulatory units, 10,135 transcription factor binding sites, 7,609 publications
Taxonomy Databases		
NCBI (Taxonomy[4])	Includes a UNIX compressed tar file called "taxdump.tar.Z". A note for *.dmp files - they are not human-friendly files, but can be uploaded into SyBase with the BCP facility.	As of 2007, it contains 260,876 entries, of which, there are 739 entries on Archaea, 18,639 entries on Bacteria, 210,423 entries on Eukaryota, 21,915 entries on Fungi, 101,365 entries on Metazoa, 77,968 entries on Viridiplantae, and 27,134 entries on Viruses

Text and Ontology Databases		
NCBI- PUBMED[5]	Medline database is available in XML with updates every two weeks	PubMed includes over 17 million citations for biomedical articles back to the 1950's. Growing at roughly 20,000 entries per week
Online Mendelian Inheritance In Man (OMIM[6])	The full-text entries were converted to an ASN.1 structured format when OMIM was transferred to the NCBI	As of September, 2007, OMIM contains more than 18,049 entries.
Gene Ontology (GO) [Ashburner et al. 2000]	Implemented using MySQL, with a monthly database release which is available in SQL and XML formats	As of September, 2007, it contains 23,803 GO terms, and 664,463 associations between 219,335 gene products and the GO terms.

2.2 Proteins

Structure: Primary, Secondary, Tertiary, and Quaternary

Proteins are functional products of genes, which have evolved over years under selective pressure, to perform very specific and essential functions. These functions depend on their structures, which arise due to particular amino acid sequence folding to generate linear chains, and compact domains with specific three-dimensional structures.

Proteomics

The objective of proteomics is the quantitative measurement of protein expression, particularly under the influence of drug or disease perturbations [Anderson and Anderson 1998]. To understand proteomics it is important to know the basic subunits of proteins, mainly domain and motif, which help in defining the structure of the protein. Domains are considered to be the natural independent units of protein structure and evolution, to the extent that they can be excised from the chain, and still be shown to fold correctly, and often still exhibit biological activity. These folding units of protein vary in length from 80-120 amino acids and may include two or more motifs. Motifs are associated with a particular function. Motifs are patterns of amino acid residues that can highlight the characteristic regions of proteins with similar function and common ancestral background [Lones and Tyrrell 2005; Seehuus et al. 2005]. Motifs are mostly confined to short stretches of protein of varying length of about 10-30 amino acids.

3. FUNDAMENTAL CHARACTERISTICS AND CHALLENGES OF MOLECULAR BIOLOGY DATABASES

In this section we briefly review the peculiar characteristics of the data arising from experiments and natural observations, which originate from the above concepts in biology. We experienced this first hand during the process of creating MITOMAP[7], a mitochondrial genome database. Then we present in Table II, a summary of the available databases.

[7]Details of MITOMAP and its information complexity can be seen in [Kogelnik et al. 1996; 1998] and at http://www. mitomap.org. The database is currently actively maintained.

3.1 Characteristics of biological data and related data management problems

Characteristic 1: *Biological data is highly complex when compared with most other domains or applications.* Definitions of such data must thus be able to represent a complex substructure of data as well as relationships and to ensure that no information is lost during biological data modeling. The structure of biological data often provides an additional context for interpretation of the information. Biological information systems must be able to represent any level of complexity in any data schema, relationship, or schema substructure-not just hierarchical, binary, or table data. As an example, MITOMAP is a database documenting the human mitochondrial genome. This single genome is a small, circular piece of DNA encompassing information about 16,569 nucleotide bases; 52 gene loci encoding messenger RNA, ribosomal RNA, and transfer RNA; 1000 known population variants; over 60 known disease associations; and a limited set of knowledge on the complex molecular interactions of the biochemical energy producing pathway of oxidative phosphorylation. We initially proposed to design a database using the traditional RDBMS or ODBMS approaches to capture all aspects of the data, but were not fully satisfied. Later we ended up designing our own structure which would be "natural" and "optimal" for understanding and navigation by a biological scientist [Navathe and Kogelnik 1999]. Later, for the sake of long-term maintainability and ease of curation, we ended up going back to a standard relational DBMS. We believe that a number of databases that are publicly available today may have undergone a similar evolution.

Characteristic 2: *The amount and range of variability in data is high. Hence, biological systems must be flexible in handling data types and values.* With such a wide range of possible data values, placing constraints on data types must be limited since this may exclude unexpected values- e.g., outlier values-that are particularly common in the biological domain. Exclusion of such values results in a loss of information. In addition, frequent exceptions to biological data structures may require a choice of data types to be available for a given piece of data.

Characteristic 3: *Schemas in biological databases change at a rapid pace.* Hence, for improved information flow between generations or releases of databases, schema evolution and data object migration must be supported. The ability to extend the schema, a frequent occurrence in the biological setting, is unsupported in most relational and object database systems. Presently systems such as GenBank re-release the entire database with new schemas once or twice a year. Such an evolutionary database would provide a timely and orderly mechanism for following changes to individual data entities in biological databases over time. This sort of tracking is important for biological researchers to be able to access and reproduce previous results.

Characteristic 4: *Representations of the same data by different biologists will likely be different (even when using the same system).* Hence, mechanisms for "aligning" different biological schemas or different versions of schemas should be supported. Given the complexity of biological data, there are multitudes of ways of modeling any given entity, with the results often reflecting the particular focus of the scientist. While two individuals may produce different data models if asked to interpret the same entity, these models will likely have numerous points in common.

In such situations, it would be useful to biological investigators to be able to run queries across these common points. By linking data elements in a network of schemas, this could be accomplished.

Characteristic 5: *Most users of biological data do not require write access to the database; read-only access is adequate.* Write access is limited to privileged users called curators. For example, the database created as part of the MITOMAP project has on average more than 15,000 users per month on the Internet. There are fewer than twenty non-curator-generated submissions to MITOMAP every month. Thus, the number of users requiring write access is small. Users generate a wide variety of read- access patterns into the database, but these patterns are not the same as those seen in traditional relational databases. User requested ad hoc searches demand indexing of often un-expected combinations of data instance classes.

Characteristic 6: *Most biologists are not likely to have any knowledge of the internal structure of the database or about schema design.* Biological database interfaces should display information to users in a manner that is applicable to the problem they are trying to address and that reflects the underlying data structure. Biological users usually know which data they require, but have little technical knowledge of the data structure or how a DBMS represents the data. They rely on technical users to provide them with views into the database. Relational schemas fail to provide cues or any intuitive information to the user regarding the meaning of their schema. Web interfaces in particular often provide preset menus supporting search which in turn do a limited type of querying of the database. However, if these interfaces are generated directly from database structures, they are likely to produce a wider possible range of access, although they may not guarantee usability.

Characteristic 7: *The context of data gives added meaning for its use in biological applications.* Hence, context must be maintained and conveyed to the user when appropriate. In addition, it should be possible to integrate as many contexts as possible to maximize the interpretation of a biological data value. Isolated values are of less use in biological systems. For example, the sequence of a DNA strand is not particularly useful without additional information describing its organization, function, the organism, etc. A single nucleotide on a DNA strand, for example, seen in context with non-disease causing DNA strands, could be seen as a causative element for Lebers Hereditary Optical Neuropathy (LHON) in the case of MITOMAP.

Characteristic 8: *Defining and representing complex queries is extremely important to the biologist.* Yet, the biologists are not likely to use any detailed syntax-based queries. Hence, biological systems must support complex queries and yet provide an easy interface to do so. Without any knowledge of the data structure (see Characteristic 6), average users cannot construct a complex query across data sets on their own. Thus, in order to be truly useful, systems must provide some tools and menu based or iconic or forms based interfaces for building these queries. As mentioned previously, many systems provide predefined query templates.

Characteristic 9: *Users of biological information often require access to "old" values of the data-particularly when verifying previously reported results.* Hence, changes to the values of data in the database must be supported through a system of archives. Access to both- the most recent version of a data value and its previous

version are important in the biological domain. Investigators consistently want to query the most up-to- date data, but they must also be able to re-construct previous work and reevaluate prior and current information. Consequently, values that are about to be updated in a biological database cannot simply be thrown away.

All of these characteristics clearly point to the fact that today's DBMSs do not fully cater to the requirements of complex biological data. A new set of features in database management systems is necessary (Chap.30, [Navathe and Elmasri 2007]).

3.2 State of the art of database creation and management for applications in genomics and proteomics

Genome research projects generate enormous quantities of data. GenBank is the National Institutes of Health (NIH) molecular database, which is composed of an annotated collection of all publicly available DNA sequences [Benson et al. 2000; Benson et al. 2003; 2007]. There exist many standalone databases, which harbor important scientific data and are goldmines for a biologist. These databases have expanded exponentially and typically double in size every 12-18 months due to development of advanced DNA sequencing technologies. GenBank statistics show that in 1995 GenBank had less than 0.3 million sequences and today it has over 10 million of them.[8]

In biological information management, two levels of heterogeneous database problem exist: one is across diverse systems housing the same types of information (for example, genetic maps in RiceGenes and MaizeDB), and a second is across different types of data that needs to be related and made accessible for analysis through a single interface (for example, genetic maps and DNA sequences differ from each other, which are different from temporal profiles of gene expression, but they all are relevant to the inquisitive scientist engineering a new species or variety) [Sobral 1999].

As described in the survey of Pearson and Soll [Pearson and Soll 1991], genome databases are used for the storage and analysis of genetic and physical maps. Just as the information inside a cell flows from DNA to RNA to Protein, there are three main categories of major types of public databases, which include Genomic DNA databases, RNA databases (complementary DNA - cDNA, Expressed Sequence Tags - ESTs, Unigene), and Protein Databases. There are numerous bioinformatic databases located around the world, which are growing very rapidly, consistent with the growth of GenBank. They harbor complete sequence and annotation data and hold the potential to be a vital resource for researchers for years to come. Most of these databases are stand-alone text-only repositories containing highly specialized medical, mutational, sequence or coordinate data. The number, size, redundancy and limited query capabilities of the current databases sometimes prevents many researchers from making full use of the information contained within them. The limitations of present-day bioinformatic databases could largely be overcome if many of them could be combined, reorganized and integrated. In Table II, we summarize some of the main databases available to the biological researcher today.

[8] http://www.ncbi.nlm.nih.gov/Taxonomy/

4. CRITICAL DATABASE RESEARCH NEEDS FOR CURRENT ISSUES OF GENOMIC AND PROTEOMIC DATABASES

4.1 Current capabilities vs. needed features

As we can see from the description of the databases in Table II, many of them still use the simple flat file organization; in terms of the data formats, ASN.1 (Abstract Syntax Notation 1) which was originally proposed for defining the syntax of the telecommunication protocols is used by several databases; we also employed it in MITOMAP due to its popularity. Relational DBMSs are getting more popular for storing genomic and proteomic information in public databases - SYBASE and MySQL seem to be two of the more frequently used RDBMSs by this community. They are used basically as storage managers without really utilizing the complex querying and transaction processing as well as concurrency control and recovery functions because they are simply not required, as biological databases are not for heavy-duty transaction processing. RDBMSs fail to meet the semantic demands of irregular, incomplete, overlapping and ill defined data which is rampant in biology. There is no support for subtyping and inheritance which is needed - e.g., a mutation in a gene could be subtyped into replacement or insertion or deletion subtypes. Object oriented DBMSs would meet the needs of several of the above databases in terms of complex structuring of objects, dealing with data types that need type constructors like sets, bags and lists. It would be easy to capture the behavior of the biological objects and to identify properties like homologs, orthologs, paralogs etc. using inheritance and encapsulation. However, the query processing techniques of the OODBMSs are not adequate and leading commercial OODBMSs like Objectstore, Versant, Objectivity and Gemstone have not been able to establish themselves to an extent where the general users felt sufficiently confident about their long term existence in the market. Unfortunately, the marketshare of OODB technology is at most only about 2 to 3% of the total database market today. We experimented with a relational DBMS and an OODBMS ourselves before settling on our own approach for modeling and implementing the mitochondrial genomic database in MITOMAP. However, most recently, we re-implemented the system using Oracle 9i for the sake of long term maintainability. Generally speaking, the relational model has been adopted for the genomic and proteomic databases at large, not because it is ideally suited for the data, but because of the support provided by the vendors and the expected longevity of the relational DBMSs in the marketplace. Although all the major relational vendors offer Object-relational features in their systems (IBM in DB2, Oracle in Oracle 10g and Microsoft in MS SQL Server), the use of these two data models concurrently has not been widely adopted by the application developers. See (Chap.22, [Navathe and Elmasri 2007]) for further details of the object-relational approach that still has good potential of applicability to the genomic and proteomic databases.

4.2 Main thrust areas of research for data management

It is obvious from the above discussion that there is a need to develop solutions for managing data in biology that go beyond the scope of the current relational and object oriented databases that will have the scalability as well as the semantic capability to handle the characteristics of biological information that we pointed

out in the section above. We highlight the issues below. Some of these coincide with those mentioned by [Jagadish and Olken 2003] in their report of the Workshop on Data Management for Molecular and Cell Biology held in February 2003 at NIH.

4.2.1 *Non-standard and unstructured data.* As we show above, this data includes sequence data for DNA, RNA, mRNA sequences and protein sequences. It is unclear whether every position in the sequence should be treated as a data object and its related information stored around it. (This is the strategy we have in MIT-OMAP). Typically sequence data itself is stored independently; other non-sequence data which describes various aspects about the sequence such as function, products, SNP (single nucleotide polymorphisms), mutations etc. would constitute the main content. Scientists are currently using tools like BLAST or PSBLAST to do pattern searches; this capability needs to be integrated into biological DBMSs. There is structural data about proteins, carbohydrates etc., that needs a 3-D representation. Techniques from GIS (geographic information systems) and CAD (computer aided design) as well as from geometric modeling need to be applied to such databases for efficient indexing and querying. Chemical pathway databases such as [Kanehisa et al. 2006] essentially store a graph where the links represent some chemical reactions/phenomena. These are hard to represent in most conventional databases and support for graph queries including recursive queries is almost completely non-existent. Data in the form of matrices occur in microarray experiments (where each cell has some intensity value) and needs further analysis in terms of clustering, classification and matrix operations.

4.2.2 *Complex query processing.* As we suggested in the characteristics of biological data, the queries tend to be complex involving paths and links along connected objects of data. Similarity of sequences, graphs and 3-D shapes is typically beyond the basic querying capabilities of RDBMSs and OODBMSs. Similarity is a vague notion in bioinformatics and becomes context-dependent as well as problem specific - e.g., consider terms like homologs, paralogs and orthologs - they are all based on notions of similarity and derived functions due to ancestry. Pattern matching using Hidden Markov Models (HMM), complex language grammars and regular expressions are common in sequence analysis and protein motifs identification. Computational biology involves indexing and processing of in-memory data; current DBMSs do not support much in-memory processing or operations between main memory and disk. Recursive query processing support is almost non-existent in today's database products; it is needed wherever graphical data is used in terms of pathways. TIGR's gene indices clustering tool [Pertea et al. 2003] uses transitive closure to form clusters on the graph with sequences as nodes. Transitive reduction queries are opposite of transitive closure and find minimal subgraphs whose transitive closure includes the original graph. Matching queries for graphs are also very complex and use properties such as homomorphism and isomorphism. Graph matching typically is analogous to global sequence alignment and subgraph matching tries to find embedded subgraphs within a graph. Matrix multiplication is another need in dealing with results of microarray or other types of data.

All of these queries have currently little or no support in the existing DBMS products. Oracle 10g includes a network data model (NDM) as part of its spatial feature

that allows users to model and analyze data as a graph. IBM has also reported work on graph data management [Eckman and Brown 2006]. DBMS developers must particularly address the requirements of these complex, sometimes NP-hard problems to make them suitable for dealing with path-oriented queries. Specialized libraries have been created (e.g., [Pitt et al. 2001]) for aiding the bioinformatics researcher with some special functions.

4.2.3 *Data interpretation and metadata management.* A very important problem with biological data is to provide enough metadata so as to allow the scientist to interpret it. Toward this end, various techniques are employed:

A) *Annotation*: The annotation process involves adding reliable and up-to-date information as possible to describe a sequence, or in other words adding biological meaning to raw sequence data using known publications, articles and databases. Proteins are better sources for annotation than DNA because each amino acid has different properties like size, charge, etc. and are closer to biological function in terms of evolution.

Two main types of annotation are present: Structural and Functional annotation.

Structural annotation: Finding genes and other biologically relevant sites thus building up a model of genome as objects with specific locations

Functional annotation: This type of annotation attaches biologically relevant information to whole sequence and individual objects

Automatic annotation is carried out with the help of already known data in current databases after confirming sequence homology rules, and then transferring the information to raw biological data eventually classifying them to specific families with similar functions. Gene clusters are used in functional prediction. The COG database [Tatusov et al. 2003] also plays an important role in this. This automatic classification is done through pattern matching, sequence clustering, comparing protein structure and function information. Automatic functional characterization is done with the help of functional databases. Context information is added with the help of comparative genome analysis, and metabolic pathway databases. Good examples of a well-annotated reference database are SWISS-PROT [O'Donovan et al. 2002; Boeckmann et al. 2003] or PIR [Wu et al. 2002].

B) *Ontology (Controlled Vocabulary)* and other language systems Biologists waste a lot of time and effort in searching for relevant and specific information related to their research. For successful data mining of related literature on diseases or gene/protein data we need a dictionary of standardized keywords, which should interface with current genetic repositories and medical terminologies. Some examples are:

MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus (MESH [Lipscomb 2000]). It mainly consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure, using to index MEDLINE article. Currently there are 22,997 descriptors in MeSH.

UMLS (Unified Medical Language System) consists of the metathesaurus, the semantic network and the specialist lexicon and is provided with a variety of tools

for installation and searching of the ontology. Metathesaurus contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases and expert systems (UMLS [Humphreys et al. 1998]). Names of concepts are present in 17 languages. The 2007AB edition of the Metathesaurus includes 1,436,586 concepts and 7,243,751 concept names in its source vocabularies.

GO (Gene ontology) consortium is also another project, which is a collaborative effort to address the need for consistent descriptions of gene products in different databases (GO [Ashburner et al. 2000]). The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The GO datasets are freely available and have three different formats: flat files (updated daily), XML (updated monthly) and MySQL (updated monthly). Currently there are 23,803 GO terms (13916 biological process, 2007 cellular component and 7880 molecular function) in its vocabulary, and over 664,463 associations between 219,335 gene products and the GO terms.

Open Biomedical Ontologies (OBO[9]) is a US National Center for Biomedical Ontology (NCBO)'s effort to create well-structured controlled vocabularies for shared use adhering to a set of principles specifying best practices in ontology development. OBD (Open Biomedical Database) is the database for OBO expressed using relational semantics. OBO will form a central element of the NCBO's BioPortal. The OBO foundry[10] contains over 65 ontologies.

Another resource is the Ontology Lookup Service (OLS[11]) part of the PRIDE project which is a central query interface for many ontologies and controlled vocabularies lookup. Multiple ontologies can be queried from a single location with a unified output format using the OLS web service interface. OLS contains 50 ontologies and has a total of 517989 terms loaded in it. Some important ontologies that can be accessed through OLS include UniProt Taxonomy Database (NEWT) - 381,608 terms, Chemical Entities of Biological Interest (CHEBI) - 16054 terms, Human Disease (DOID) - 16054 terms, Mammalian Phenotype (MP) - 5923 terms.

The PRoteomics IDEntifications database (PRIDE[12]) is a centralized, standards compliant, public data repository for proteomics data. It has been developed to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications. The PRIDE database currently contains: 3,182 experiments, 340,493 identified proteins, 2,145,587 identified peptides, 309,971 unique peptides, and 2,582,616 spectra.

4.2.4 *Data integration across related databases.* We pointed out in Table II the various databases related to information on genes and proteins from organisms including the human (homo sapiens). There has been an intensive study and reporting in the literature of certain species like E-coli, yeast, C-elegans, drosophila (fruit fly), or mouse. In the botanical domain, a lot of data pertaining to the Rice or Maize

[9]http://obo.sourceforge.net
[10]http://obofoundry.org/
[11]http://www.ebi.ac.uk/ontology-lookup/
[12]http://www.ebi.ac.uk/pride/

genomes has been collected. It is not possible with the available technology of web links to get a complete and uniform picture of where science stands today on any one of these organisms unless a scientist spends many hours using search engines with some associated frustration. While annotations and links take the biologist from one database to the next, currently no uniform interfaces or consolidation of data has been done so that information can be accessed in an integrated fashion in any given context or by any particular classification. Scientists at SRI have embarked on a bold effort to consolidate all metabolic pathways and provide a complete view of one organism- the bacterium E-coli based on their existing EcoCyc database [Karp et al. 2002; Keseler et al. 2005]. In general, there is a tremendous need to bring together related heterogeneous information under one uniform interface to support a variety of ambitious applications which we have stated later in Section 6. This problem is saddled with the typical problems in databases of heterogeneous data integration - multiple models, multiple formats, different underlying files and database systems, and a large amount of context-sensitive semantic content. The general advances in heterogeneous database integration and multi-database query processing ought to come to the rescue of the biologist in this domain. Many biological integration efforts have been reported recently, like BioMOBY [Wilkinson and Links 2002], and DiscoveryLink [Haas et al. 2001]. A survey of approaches to biological source integration appears in [Hernandez and Kambhampati 2004]).

4.3 Need for a uniform set of data management solutions

On the laboratory technology front, the challenge will be to devise more efficient and cost effective technologies for identifying and scoring all types of genetic variants (at the structural level) in a given genome, with the human genome taking the lead [Chakravarti 1999]. Of special interest, the development of high-throughput methods to monitor and analyze responses at the level of regulatory and biochemical networks, will allow enhanced understanding of genetic control.

The shift in emphasis from data accumulation to data interpretation has already begun and will continue to expand. Integration of data types, provision of unified interfaces to complex biological data sets and provision of distributed data acquisition, storage and analysis is a current focus of many public and private efforts in the broadly defined field of bioinformatics [Sobral et al. 2001].

To prevent further continued anarchy in terms of handling of raw data production and its analysis, a collaborative effort is needed to reduce redundancy and improve the quality by curation of data.

5. RESEARCH CASE STUDIES

We have been involved with two pertinent projects related to the database creation of a mitochondrial genome database and the mining of a medical text (PubMed[13]) database for the benefit of biologists conducting microarray studies. We briefly summarize these as examples of research case studies where in the first, a new database has been designed and populated and in the second, an existing very rich database of literature is being utilized to support interpretation of experimental data by the biologists.

[13] http://www.ncbi.nlm.nih.gov/PubMed/

Mitomap: (www.mitomap.org) - this is a database of mitochondrial genome information developed as the result of the Ph.D. work of Andy Kogelnik (with advisor Navathe at Georgia Tech and with Prof. Doug Wallace as advisor in molecular genetics at Emory). Mitochondria are monomorphic little sausage like structures present in each cell and only cellular organelles known to have their own deoxyribonuclease acid DNA (mtDNA). They are normally considered the powerhouses of a cell since they generate adenosine triphosphate (ATP) but are also involved in many other cellular functions [Naviaux 1997]. MITOMAP is the most complete database of published data relating to the human mitochondrial genome. This single genome is a small, circular piece of DNA encompassing information about 16,569 nucleotide bases; 52 gene loci encoding messenger RNA, ribosomal RNA, and transfer RNA; 1000 known population variants; over 60 known disease associations; and a limited set of knowledge on the complex molecular interactions of the biochemical energy producing pathway of oxidative phosphorylation [Kogelnik et al. 1996; 1998; Navathe and Kogelnik 1999]. Many important diseases like Parkinsons, Optical Neuropathy, Cardiomyopathy, and Deafness have been linked to mitochondrial disorders and hence there is a vast potential use for this data. Our goal is to make this a comprehensive repository of mitochondrial genome information that includes sequence data as its core and relates it to biochemical functional data, anthropological data, gene-gene interactions (including those with nuclear genes) and records disease causing as well as natural mutations. Having experimented with the RDBMS and OODBMS approaches, we stored it in our own data model using the ASN.1 notation. Recently we have moved it to the Oracle RDBMS for long term maintainability. The database is maintained and curated at the Center for Molecular and Mitochondrial Medicine and Genetics (MAMMAG) at the university of California, Irvine. MITOMAP receives a few new entries every week and is regarded as a standard source of information on mitochondrial genome.

Genetrek: under this project, our goal is to create a system which extracts relevant keywords from medical text for a given set of genes that may have been identified as having a differential or temporal pattern of interest from microarray experiments. We have been mainly dealing with neurological disorder related genes. The data is from Medline, a database of over 17 million abstracts dating back to 1950's. We do extraction of important keywords based on statistical weighting analysis [Liu et al. 2004a; Liu et al. 2006]). We employ a variety of techniques like stemming, stop lists, background sets of documents to extract the words. We manipulated gene vs. keyword matrices with a variety of clustering algorithms and eventually developed our own algorithm called BEA-PARTITION [Liu et al. 2005] based on the Bond Energy Algorithm which was used in [Navathe et al. 1984] for attribute partitioning in databases. We showed in [Liu et al. 2005] that our clustering yields better results compared with k-means, self-organizing maps, and hierarchical clustering, which have been popular clustering algorithms among bio-scientists. We also showed [Liu et al. 2004b] that the traditional information retrieval measure of TFIDF (term frequency*inverse document frequency) yields better results than the Z-score technique commonly used in the prior information extraction work from biomedical text [Andrade and Valencia 1998]. So far our results are extremely encouraging and we have come up with new clusters of genes

on known gene sets that lend new biological insights as well as we have found that for certain genes, we get keywords indicative of their functions that are otherwise difficult to determine from publicly available databases. The keywords have been used as features to perform an SVM (support-vector-machines)-based classification of literature for public health with very high precision and recall in a project with CDC [Polavarapu et al. 2005].

6. LINKING GENOMIC AND PROTEOMIC DATABASES TO FUTURE HEALTHCARE APPLICATIONS

In this section we highlight several important areas and issues where genomic and proteomic databases as well as the various technologies of data generation that feed into these databases will play a significant role in the next few decades. Database technologists need to realize this vast potential application area.

Genomic Medicine

In the new century the healthcare industry faces a variety of new challenges as consumers are becoming more active in the search for healthcare information and treatment alternatives. Healthcare community faces several technological and scientific difficulties due to lack of universal gateways, partial interoperability and no standardized template for researchers and clinicians to record their findings [Kumar 2007b]. The growing use of the Internet is also pushing the envelope on e-health and the demand for higher levels of service. Hence there exists a need to invest in resources, research, and partnerships with the main objective of making a positive impact on the quality and efficiency of services offered by healthcare organizations.[14] In future, a patient's medical record will also include his/her complete genome as well as a catalog of single base-pair variations depending on his/her family history, race and geographical location. It will be used to accurately predict a patient's predisposition to not only the rare Mendelian diseases but also susceptibility to a whole range of common but complex medical diseases like cancers, coronary heart disease, stroke, diabetes, hypertension, neurodegenerative disorders and psychiatric illness like bipolar disorder and schizophrenia [Kumar 2007b; Cardon and Bell 2001]. A patient's response to the therapeutic dosage of drugs could be accurately measured based on their genetic makeup to metabolize a particular drug. The genetic variation between individuals together with environmental factors probably determines disease susceptibility and protection, and is important for drug efficacy and side effects [Holden 2000; Chakravarti 2000]. This will permit a patient to be treated as a biochemical and genetic individual, thus making medical interventions more specific, precise, and successful. The PharmGKB project at Stanford (pharmgkb.org) is aiming to develop a pharmacogenetics and phramacogenomics database as an integrated resource to study how variation in human genetics leads to variation in response to drugs. In addition, the increased power of medicine to predict susceptibility to specific diseases will allow a patient to alter one's lifestyle in order to reduce the likelihood of developing such diseases or to be treated with preventive or disease-delaying medicine. Such an approach has

[14]IBM Life Sciences, Pharmaceutical Clinical Development: The future of clinical trials- How genomics, proteomics and technology are changing the clinical development process, 2002
IBM Life Science Solutions: Integrated infrastructure to accelerate and enhance research, 2002

a tremendous potential to reduce morbidity and mortality [Nebert and Bingham 2001]. To reach such capability we need to understand all human genetic variations which include all mutations and gene polymorphisms. Scientists have already taken steps to tackle this enormous challenge by initiating projects like the Human Genome Variation Database, Indian Genome Variation Project and human gene mutation database [Kumar 2007a; 2007b].

Disease Studies for Large Populations

With the completion of Human Genome Project, researchers predict the use of personalized genomic information will transform the way clinicians & epidemiologists combine medicine and public health [Guttmacher and Collins 2002; Collins and Guttmacher 2003; Khoury et al. 2000]). Analyzing genetic data, interpreting genetic risks, and then formulating and testing new concepts will require significant computational and storage power. Initiatives like the HuGE Net for appraisal and integration of epidemiological data on human genome will provide some of the refined ever increasing databases which collect, analyze and correlate medicine and public health together in the 21st century [Khoury 2002]. The implementation of a high-performance information infrastructure to facilitate and support the work of life scientists in genomics, proteomics, and drug discovery and development is needed. In future, genomic and proteomic databases will have to handle genomic data from almost every individual in the world. Examples of such large population databases include UK Biobank [Wright et al. 2002], CartaGene[15], Decode Genetics [Hakonarson et al. 2003], Estonia[16], GenomeEUtwin[17]. Manipulation of such vast data needs durable, robust databases with trouble-free implementation/ data cleaning and data curation capability. Databases should be able to accommodate data changes, allow flexible searches, and be scalable for different data sources.

Physician Aids

Physicians will require interactions of different databases, which could potentially include information from electronic medical records database, genome databases, laboratory data, detailed disease information (the OMIM database[18]), and pharmacological data before they make a diagnosis and prescribe a personalized drug. The Marshfield Clinic Personalized Medicine Research Project (PMRP [McCarty et al. 2005]) is a population-based DNA biobank database created to serve research focused on pharmacogenetics, genetic epidemiology and population genetics. Interactions of multiple databases will not only require enormous computational capability to manage large datasets and analyze them but also requires robust and scalable data integration. New paradigms such as systems biology are predicated on the availability of such large integrated data sets of many different types [Sternberg 2000]. Data integration is a critical aspect of disease diagnosis and drug discovery in the future.

Role of Gene Therapy

The potential for using genes themselves to treat disease–gene therapy–is the most exciting application of DNA science. Although still in its infancy and plagued

[15]http://www.cartagene.qc.ca/en/index.htm
[16]http://www.geenivaramu.ee/index.php?show=main&lang=eng
[17]http://www.genomeutwin.org/
[18]http://www.ncbi.nlm.nih.gov/omim/

by technical difficulties, gene therapy for single-gene diseases will almost certainly be routine and successful in the next 20 years.[19] Certain aberrant disease-associated genes will be replaced with normally functioning versions, and several hundred diseases will be curable. Gene therapy treatments utilize retro virus to deliver such damaged or missing genes to the targeted cells, which in turn gets assimilated into the cellular genome over time. But such therapies require in depth understanding of not only the genes but also their products (i.e., proteins) which results due to their expression and their delivery systems. Many researchers have created their own ad hoc databases with detailed information on a particular disease or condition under study for gene therapy. Such ventures, though rich in information, still pose enormous challenges for the database community at the time of integration with other public databases. Recently gene therapy appears to have cured Myeloid Blood Disease which includes a variety of bone marrow failure syndromes.[20] Similar studies have shown that expressing an engineered wild-type copy of the gene in the nucleus can rectify the effects of a pathogenic mutation in a human mitochondrial gene. But the uniqueness of the mitochondrial genome presents a number of obstacles to the successful use of gene therapy for the treatment of mitochondrial DNA disease [Douglass and Robert 2002]. Our long term goals with MitoMap are to make use of it as a resource for the physician and eventually in gene therapy.

Pharmacogenomics: Designer drugs

Until the late 20th century, drug discovery was mainly a slow tedious process based on the screening and testing of thousands of chemical and natural substances for potential therapeutic activity. Identification of new drug targets was always the main bottleneck of the drug discovery process. For efficiency and cost-cutting, nearly the entire pharmaceutical industry has been developing systems for automating the steps in drug discovery to streamline the entire process. With the advent of genome research during the past decade, pharmaceutical companies were exposed to a new frontier for drug development. In-silico genetic models could identify polymorphisms in drug metabolizing enzymes that contribute to differential drug performances. Such genetic models could also formulate novel and accurate hypotheses about traits of biomedical importance [Wang et al. 2005; Kaminsky and Zhang 1997]. Scientists predict that instead of trying to replace a gene, it will be more effective and simpler to manipulate cellular functions by replacing proteins made by that gene. But a single protein can potentially also have multiple functions which could be interpreted as side effects in a patient. Targeting such a protein to a relevant cell is indeed a challenge and needs a detailed study across different human races or even species. Instead of having to rely on a chance and screening thousands of molecules to find an effective drug, which is how most drugs used today were found, scientists will begin the process of drug discovery with a clearer notion of what they're looking for and where it will act by searching protein structure databases There have been estimates that the number of these targets would range from 3000 to 10,000 [Minoru 2001].

Structural biology & drug discovery

The elucidation of the 3D structure of potential drug targets, in particular in

[19] Genetics: The future of Medicine, National Human Genome Research Institute

[20] http://www.cincinnatichildrens.org/about/news/release/2006/3-gene-therapy.htm

the case where complexes between drug and target can be obtained, is another very powerful approach to target validation. Specific structural and physiochemical properties are required for such macro drug molecules for efficacy, bioavalibility and safety [Lipinski et al. 2001]. These requirements often limit the number of drug-gable macro drug molecules which can be found but at the same time some drug-gable macro drug molecules are also shared among different diseases like Beta-adrenoreceptor for circulatory system diseases, nervous system disorders and respiratory system diseases [Hopkins and Groom 2002; Hardy and Peet 2004]). Structural data is being analyzed computationally and stored in huge protein-structure databases like Strucural Folds of Protein database (SCOP [Andreeva et al. 2004]) and Protein Data Bank database (PDB [Weissig et al. 2000]). And because rationally designed drugs are more likely to act very specifically, they will be less likely to have damaging side effects such as drug interactions and drug allergies.

Systems Biology

Microbes can thrive mostly in any environment and have mostly helped this planet to sustain life. It is essential to understanding the intricate details of their functions which can eventually enable us to harness their sophisticated biochemical abilities towards energy production, global climate change mitigation, and toxic waste cleanup. This study of tens of thousands of genes and proteins working together in interconnected networks to orchestrate the chemistry of life is known as whole-systems biology. The changes in the metabolism are often expressed through transcriptional changes introduced by complex regulatory mechanisms coordinating the activity of different metabolic pathways [Patil and Nielsen 2005]. Systems biology is creating a context for interpreting the vast amounts of genomic and proteomic data being produced by pharmaceutical companies in support of drug development.

Harnessing Microarray Data

Gene discovery and expression: Traditional methods in a molecular biology lab take a long time to discover all the genes and their expression since scientists work on a one gene on one experiment basis. With the discovery of a recent technology, called DNA microarrays, biologist can see a better picture of the interactions among thousands of genes simultaneously. Microarray technology provides a tool to potentially identify and quantify levels of gene expression for all genes in an organism. With 10,000 to 80,000 genes per experiment, the problem of how to further study the identified genes remains. The large set of important databases we described in Table 2 can be employed for a better interpretation of the results of microarray experiments. This is where the project Genetrek we have described above comes in.

Disease diagnosis

Microarrays have helped us understand the functions of individual genes. They are also proving to be an essential tool for understanding disease processes and identifying new diseases [Meltzer 2001]. Until recently disease diagnosis was done by identifying defects at the tissue or cellular level, but with microarrays the shift towards using positional cloning in identifying the underlying patho-mechanisms in human disease, has proved to be a successful step. There are many emerging variants on microarray technology, such as expression arrays, exon arrays, array-based

comparative genomic hybridization and sequencing arrays [Dobrin and Stephan 2003]. The power of mRNA and its association with the disease process has only recently been acknowledged by both scientists and physicians. Certain transcriptional changes can be tracked to predict a disease state or exacerbation of a disease state. Laboratory tests based on this concept will be able to predict cardiovascular diseases in patients months before a patient can even manifest its symptoms. Such preemptive disease diagnosis has a huge potential in reducing long term healthcare costs. Such studies require complex algorithm development which can analyze both patient phenotype data against transcriptional changes at cellular level via microarray/PCR laboratory processes [Patil and Nielsen 2005].

Overall, these futuristic medical applications can only be successful if they are supported by a standardized but robust technological growth.

Glossary

Basic terms

DNA: (Deoxyribonucleic acid) DNA molecules carry the genetic information necessary for the organization and functioning of most living cells and control the inheritance of characteristics.

RNA: (Ribonucleic acid) RNA is involved in the transcription of genetic information; the information encoded in DNA is translated into messenger RNA (mRNA), which controls the synthesis of new proteins with help of tRNA and rRNA.

Base-pair: Nucleotides in a pair with each other: A pairs with T, and G pairs with C.

Gene: Basic unit of heredity made up of a string of DNA base pairs.

Genetics: Study of genes and their inherited properties.

Allele: Any of the alternative forms of a gene that may occur at a given gene locus.

Genotype: Allelic composition of an individual.

Phenotype: Any morphological, biochemical, behavioral properties of an individual.

Terms related to Genomics and Proteomics

(1) Prokaryotes Genomes: They have inhabited the earth for billions of years and are small cells with relatively simple internal structures e.g. bacteria.

(2) Eukaryotic Genomes: The eukaryotic cell probably originated as a complex community of prokaryotes. Humans, called as Homo Sapiens (Homo: genus, sapiens: species) are eukaryote organism.

(3) Archael Genomes: Archaea and bacteria are the two main branches of prokaryotic evolution.

(4) Genomics is defined as the scientific discipline, which focuses on the systematic investigation of genomes, i.e., the complete set of chromosomes and genes of an organism [Houle et al. 2000].

(5) Structural genomics refers to the large-scale determination of DNA sequences and gene mapping.

(6) Functional genomics is a systematic scientific study that seeks to recognize and describe the function of genes, and expose when and how genes work together to produce traits.

(7) Homologs can be of two types and crucial to the understanding of evolutionary relationships between genomes and gene functions. Homologs have common origins but may or may not have common activity.

(8) Orthologs: Evolutionary counterparts derived from a single ancestral gene in the last common ancestor of the given two species. They retain the same ancestral function.

(9) Paralogs: Evolutionary counterparts evolved from duplication within the same ancestral genome. They tend to evolve new function.

(10) Horizontal gene transfer occurs due to acquisition of genes from other species, genera or even higher taxa. A good example of such horizontal gene transfer is Archaeal genomes where some genes are close homologs between eukaryotes and bacteria.

(11) Vertical or standard gene transfer occurs due to acquisition of genes from same species.

(12) Protein: Primary, Secondary, Tertiary, Quaternary structures: The amino acids sequence forming a polypeptide chain is called its primary structure. Certain regions of this polypeptide chain form local regular secondary structures in the form of alpha helix and beta strands. Further packing leads to formation of tertiary structure, which are compact globular units, also called as domains. The quaternary structure may be formed by a bunch of tertiary structures formed from polypeptide chain [Branden and Tooze 1999].

(13) Proteomics: Defined as the use of quantitative protein-level measurements of gene expression in order to characterize biological processes and elucidate the mechanisms of gene translation.

(14) Domain: In structural biology they are defined as structurally compact independently folding parts of protein molecules. In comparative genomics the central atomic objects are parts of proteins that have distinct evolutionary trajectories and can have single domain or multi-domain architecture.

(15) Motif: They are groups of highly conserved amino acid residues in multiple alignments of domain that tend to be separated by regions of less pronounced sequence.

7. CONCLUSION

This paper is an attempt to take stock of the current situation related to a large number of public databases containing genomic and proteomic information repositories. To help the average (non-biologist) reader, we provided the important basic biological concepts and phenomena that provide the experimental and factual data to these databases. A glossary of important terms is provided at the end of the paper. Typically, most of these databases undergo "curation", a process of quality and reasonableness checking that precedes data entry.

We have tried to highlight the unique characteristics of biological data and the challenges faced by computers scientists and database professionals in the creation,

management and querying of these databases. Finally, we have provided a brief glimpse of the myriad applications in medicine and health sciences that relate to diagnosis of disease, treatments based on knowledge of genetic screens, designer drug development, gene therapies, and all types of assistance that will be made available to physicians. With the systemic view of biology as an integrative discipline and with the help of bioinformatics tools at the disposal of biologists, the next several decades will see a major revolution in improving the health of the human race.

ACKNOWLEDGMENTS

The authors are grateful for the contributions of Ying Liu, Saurav Sahay and Neha Narkhede during the development of this paper. The anonymous referees also provided useful comments.

REFERENCES

Anderson, N. L. and Anderson, N. G. 1998. Proteome and proteomics: new technologies, new concepts and new words. *Electrophoresis 19,* 11, 1853–1861.

Andrade, M. and Valencia, A. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics 14,* 2, 600–607.

Andreeva, A., Howorth, D., Brenner, S., Hubbard, T., Chothia, C., and Murzin, A. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acid Res 32,* Database Issue, 226–229.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., and Cherry, J. 2000. Gene ontology: tool for the unification of biology. *Nature Genet. 25*, 25–29. The Gene Ontology Consortium http://www.geneontology.org.

Attwood, T. 2002. The PRINTS database: a resource for identification of protein families. *Briefings in Bioinformatics 3,* 3, 252–263.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B., and Wheeler, D. 2000. Genbank. *Nucleic Acids Research 28,* 1, 15–18.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. 2003. Genbank. *Nucleic Acids Research 31,* 1, 23–27.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. 2007. Genbank. *Nucleic Acids Research 35,* Database issue, D21–25.

Berman, H., Olson, W., Beveridge, D., Westbrook, J., Gelbin, A., and Demeny, T. 1992. The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J. 63*, 751–759.

Blake, J. A., Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., and Group, M. G. D. 2006. The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res 34,* Database issue, 562–567.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., and Gasteiger, E. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research 31,* 1, 365–370. http://us.expasy.org/sprot/.

Branden and Tooze. 1999. *Introduction to Protein Structure*, Second Edition ed. Garland Publishing.

Cardon, L. and Bell, J. 2001. Association study designs for complex diseases. *Nature Reviews Genetics 2*, 91–99.

Chakravarti, A. 1999. Population genetics-making sense out of sequence. *Nature Genet. 21,* sup. 1, 56–60.

Chakravarti, A. 2000. To a future of genetic medicine. *Nature 409*, 822–823.

Collins, F. and Guttmacher, A. 2003. Welcome to the genomics era. *New England Journal of Medicine 349*, 996–998.

Crosby, M., Goodman, J., Strelets, V., Zhang, P., Gelbart, W., and the FlyBase Consortium. 2007. FlyBase: genomes by the dozen. *Nucleic Acids Research 35,* Database issue, 562–567.

Dobrin, S. and Stephan, D. 2003. Integrating microarrays into disease-gene identification strategies. *Expert Rev. Mol. Diagn. 3,* 3, 375–385.

Douglass, M. and Robert, N. 2002. A roundabout route to gene therapy. *Nat Genet. 30,* 4, 345–346. http://genetics.nature.com.

Eckman, B. and Brown, P. 2006. Graph data management for molecular and cell biology. *IBM Journal of Res and Dev Nov.*, 545–560.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., and Merregaert, J. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature 260,* 5551, 500–507.

Finley, K., Edeen, P., Cumming, R., Mardahl-Dumesnil, M., Taylor, B., and Rodriguez, M. 2003. blue cheese mutations define a novel, conserved gene involved in progressive neural degeneration. *Journal of Neuroscience 23*, 1254–1264.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., and Lassmann, T. 2006. Pfam: clans, web tools and services. *Nature 34,* Database Issue, 247–251.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., and al. et. 1995. Whole genome random sequencing and assembly of haemophilus influenzae rd. *Science 269,* 5223, 496–512.

François, O. B. and Peer, K. 2001. *A Molecular Biology Database Digest*. Institute for Computer Science, University of Munich, Germany. http://www.pms.informatik.uni-muenchen.de.

Guttmacher, A. and Collins, F. 2002. Genomic medicinea primer. *New England Journal of Medicine 347*, 1512–1520.

Haas, L. M., Schwarz, P. M., Kodali, P., Kotlar, E., Rice, J. E., and Swope., W. C. 2001. Discoverylink: A system for integrated access to life sciences data sources. *IBM Systems Journal 40*, 489–509.

Hakonarson, H., Gulcher, J., and Stefansson, K. 2003. decode genetics inc. *Pharmacogenomics J. 4*, 209–215.

Hardy, L. and Peet, N. 2004. The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. *Drug Discov Today 9*, 117–126.

Henikoff, J., Greene, E., Pietrokovski, S., and Henikoff, S. 2000. Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res. 28,* 1, 228–230.

Hernandez, T. and Kambhampati, S. 2004. Integration of biological sources: Current systems and challenges ahead. *ACM SIgmod Record 33*, 51–60.

Holden, A. 2000. The SNP consortium: a case study in large pharmaceutical company research and development collaboration. *Journal of Commercial Biotechnology 6*, 320–324.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. 1992. A database of protein structure families with common folding motifs. *Protein Science 1*, 1691–1698.

Hopkins, A. and Groom, C. 2002. The druggable genome. *Nat Rev Drug Discov 1*, 727–730.

Houle, J., Cadigan, W., Henry, S., Pinnamaneni, A., and Lundahl, S. 2000. Database mining in the human genome initiative. Tech. rep., AMITA.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Castro, E. D., Langendijk-Genevaux, P., Pagni, M., and Sigrist, C. 2006. The PROSITE database. *Nucleic Acids Res. 34,* Database Issue, 227–230. http://us.expasy.org/prosite/prosuser.html#convent, http://us.expasy.org/prosite/.

Humphreys, B., Lindberg, D., Schoolman, H., and Barnett, G. 1998. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc 5*, 1–11.

Jagadish, H. and Olken, F. 2003. Database management for life science research: Summary report of the workshop on data management for molecular and cell biology at the national library of medicine. *OMICS 7,* 1, 131–137.

Kaminsky, L. and Zhang, Z. 1997. Human P450 metabolism of warfarin. *Pharmacology & Therapeutics 73*, 67–74.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., and Kawashima, S. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res. 34,* Database Issue, 354–357.

Karp, P., Riley, M., Saier, M., Paulsen, I., Collado-Vides, J., and Paley, S. 2002. The EcoCyc database. *Nucleic Acids Research 30,* 1, 56.

Keseler, I., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I., Peralta-Gil, M., and Karp, P. 2005. EcoCyc: A comprehensive database resource for Escherichia coli. *Nucleic Acids Research 33,* Database Issue, 334–337.

Khoury, M. 2002. Epidemiology and the continuum from genetic research to genetic testing. *Am J Epidemiol 156*, 297–299.

Khoury, M., Burke, W., and Thomson, E. 2000. Genetics and public health in the 21st century: Using genetic information to improve health and prevent disease. *New York: Oxford University Press*, 3–23.

Kitano, H. 2001. *Foundations of Systems Biology.* MIT Press.

Kogelnik, A., Lott, M., Brown, M., Navathe, S., and Wallace, D. 1996. MITOMAP: a human mitochondrial genome database. *Nucleic Acids Research 24,* 1, 177–179.

Kogelnik, A., Lott, M., Brown, M., Navathe, S., and Wallace, D. 1998. MITOMAP: a human mitochondrial genome database-1998 update. *Nucleic Acids Research 26,* 1, 112–115.

Kolchanov, N., Ignatieva, E., Ananko, E., Podkolodnaya, O., Stepanenko, I., and Merkulova, T. 2002. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res. 30,* 1, 312–317.

Koonin, E. and Galperin, M. 2003. *Sequence-Evolution- Function. Comparative Approaches in comparative genomics.* Kluwer Academic Publishers.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., and Andersson, M. 2007. EMBL nucleotide sequence database in 2006. *Nucleic Acids Research 35,* Database issue, 16–20. http://www.ebi.ac.uk/embl/.

Kumar, D. 2007a. Genome mirror - 2006. *Genomic Med. 1,* 1-2, 87–90.

Kumar, D. 2007b. Genomic medicine: a new frontier of medicine in the twenty first century. *Genomic Med. 1,* 1-2, 3–7.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. 2006. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Research 34,* Database Issue, 257–260.

Lewin, B. 2007. *Gene IX.* Jones & Bartlett.

Lipinski, C., Lombardo, F., Dominy, B., and Feeney, P. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev 46*, 3–26.

Lipscomb, C. 2000. Medical subject headings (MeSH). *Bull. Med. Libr. Assoc. 88*, 265–266. http://www.nlm.nih.gov/pubs/factsheets/mesh.html.

Liu, Y., Brandon, M., Navathe, S., Dingledine, R., and Ciliax, B. 2004. Automatic keyword extraction from medline for functional gene clustering. In *11th MEDInfo 2004 (American Medical Informatics Association Official Annual Conference).* 292–296.

Liu, Y., Ciliax, B., Borges, K., Dasigi, V., Ram, A., Navathe, S. B., and Dingledine, R. 2004. Comparison of two schemes for automatic keyword extraction from medline for functional gene clustering. In *Proc. 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04).* 394–404.

Liu, Y., Navathe, S., Civera, J., Dasigi, V., Ram, A., Ciliax, B., and Dingledine, R. 2005. Text mining biomedical literature for discovering gene-to-gene relationships: A comparative study of algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 2,* 2, 62–76.

Liu, Y., Navathe, S., Pivoshenko, A., Dasigi, V., Ciliax, B., and Dingledine, R. 2006. Text analysis of medline for discovering functional relationships among genes: Evaluation of keyword extraction weighting schemes. *International Journal of Data Mining in Bioinformatics, (IJDMB) 1,* 1, 88–110.

Lones, M. A. and Tyrrell, A. M. 2005. The evolutionary computation approach to motif discovery in biological sequences. In *Proc. of the Genetic and evolutionary computation workshop*. 1–11.

Marchler-Bauer, A., Anderson, J., Cherukuri, P., DeWeese-Scott, C., Geer, L., and Gwadz, M. 2005. CDD: a conserved domain database for protein classification. *Nucleic Acids Res. 33,* Database Issue, 192–196.

Matys, V., Fricke, E., Geffers, R., Göling, E., Haubrock, M., and Hehl, R. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res 31,* 1, 374–378.

McCarty, C., Wilke, R., Giampietro, P., Wesbrook, S., and Caldwell, M. 2005. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Med. 2*, 49–70.

Meltzer, P. 2001. Spotting the target: microarrays for disease gene discovery. *Curr. Opin. Genet. Dev. 11,* 3, 258–263.

Minoru, S. K. 2001. Embryogenomics: developmental biology meets genomics. *TRENDS in Biotechnology 19,* 12, 511–518.

Navathe, S., Ceri, S., Wiederhold, G., and Dou, J. 1984. Vertical partitioning algorithms for database design. *ACM Transactions on Database Systems 9*, 680–710.

Navathe, S. and Elmasri, R. 2007. *Fundamentals of Database Systems*, 5th ed. Addison Wesley.

Navathe, S. and Kogelnik, A. 1999. *Lecture Notes in Computer Science*. Springer- Verlag, Chapter Mitomap: Addressing the Challenges of Modeling Biological Information in Conceptual Modeling: Current Issues and Future Directions.

Naviaux, R. 1997. The spectrum of mitochondrial disease, in mitochondrial and metabolic disorders-a primary care physician's guide. *Psy-Ed Corp., Oradell, NJ*, 3–10. http://biochemgen.ucsd.edu/mmdc/ep-3-10.pdf.

Nebert, D. W. and Bingham, E. 2001. Pharmacogenomics: out of the lab and into the community. *TRENDS in Biotechnology 19,* 12, 519.

O'Donovan, C., Martin, M., Gattiker, A., Gasteiger, E., Bairoch, A., and R., A. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL brief. *Briefings in Bioinformatics 3,* 3, 275–284.

Patil, K. and Nielsen, J. 2005. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci U S A 102*, 2685–2689.

Pearl, F., Bennett, C., Bray, J., Harrison, A., Martin, N., and Shepherd, A. 2003. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research 31,* 1, 452–455.

Pearson, M. and Soll, D. 1991. The human genome project: a paradigm for information management in the life sciences. *FASEB J. 5,* 1, 35–39. http://www.gdb.org.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., and Karamycheva, S. 2003. TIGR Gene Indices Clustering Tools (TGICL): a software system for fast clustering of large est datasets. *Bioinformatics 19,* 5, 651–652.

Pitt, W., Williams, M., M., S. S., Bleasby, A., and Moss, D. 2001. The bioinformatics template library-generic components for biocomputing. *Bioinformatics 17,* 8, 729–737.

Polavarapu, N., Navathe, S., Ramnarayanan, R., ul Haque, A., Sahay, S., and Liu, Y. 2005. Investigation into biomedical literature screening using support vector machine. In *Proc. IEEE Computational Systems and Bioinformatics Conference (CSB'05)*. 366–374.

Roberts, R., Vincze, T., Posfai, J., and Macelis, D. 2007. REBASE–enzymes and genes for DNA restriction and modification. *Nucl. Acids Res. 35,* Database Issue, 269–270.

Schofield, P. N., Bard, J. B., Booth, C., Boniver, J., Covelli, V., and Delvenne, P. 2004. Pathbase: a database of mutant mouse pathology. *Nucleic Acids Res 32,* Database issue, 512–515.

Seehuus, R., Tveit, A., and Edsberg, O. 2005. Discovering biological motifs with genetic programming. In *Proc. of the Genetic and evolutionary computation conference*. 401–408.

Snyder, L. and Champness, W. 2007. *Molecular Genetics of Bacteria*. ASM Press.

Sobral, B. 1999. Bioinformatics and the future role of computing in biology. http://www.agbiotechnet.com/proceedings/10_Sobral.pdf.

Sobral, B., Waugh, M., and Beavis, B. 2001. Information systems approaches to support discovery in agricultural genomics. *Advances in cellular and molecular biology of plants 6*, 139–166.

Sternberg, D. 2000. The diagnostic strategies in mitochondrial disorders: The example of the diagnostic group in LA SALPêTRIèRE Laboratoire de Biochimie B, CHU PitiéSalpêtrière, Juin. http://www.ifrns.chups.jussieu.fr/Sternberg-mito.pdf.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., and Koonin, E. V. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4,* 41, 871–880.

Wang, J., Liao, G., Usuka, J., and Peltz, G. 2005. Computational genetics: From mouse to man? *Trends in Genetics 21*, 526–532.

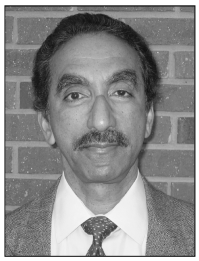
Weissig, H., Shindyalov, I., and Bourne, P. 2000. The Protein Data Bank. *Nucleic Acids Research 28,* 1, 235–242.

Wilkinson, M. and Links, M. 2002. BioMOBY: an open-source biological web services proposal. *Brief Bioinform 3*, 331–341.

Wood, V., Gwilliam, R., Rajandream, M., Lyne, M., Lyne, R., and Stewart, A. 2002. The genome sequence of Schizosaccharomyces pombe. *Nature 415,* 6874, 871–880. http://db.yeastgenome.org/cgi-bin/locus.pl?locus=act1.

Wright, A., Carothers, A., and Campbell, H. 2002. Gene-environment interactionscthe BioBank UK study. *Pharmacogenomics J. 2*, 75–82.

Wu, C., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., and Hu, Z. 2002. The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Ressearch 30,* 1, 35–37. http://pir.georgetown.edu/pirwww/dbinfo/pirpsd.html, http://pir.georgetown.edu/.



Shamkant B. Navathe received the PhD degree from the University of Michigan in 1976. He is a professor in the College of Computing, Georgia Institute of Technology. He has published more than 150 refereed papers in database research; his major contributions are in database modeling, database conversion, database design, database integration, distributed database allocation, and data mining and engineering and biological applications. Current projects include text mining of medical literature databases, creation of databases for biological applications, transaction models, security and privacy in P2P and Web applications, and data mining for better understanding of genomic/proteomic and medical data. His recent work has been focusing on issues of mobility, scalability, interoperability, and personalization of databases in scientific, engineering, and e-commerce applications. He is an author of the book, Fundamentals of Database Systems, with R. Elmasri (Addison Wesley, fifth edition, 2007) which is currently the leading database text-book worldwide. He also co-authored the book Conceptual Design:An Entity Relationship Approach (Addison Wesley, 1992) with Carlo Batini and Stefano Ceri. He was the general co-chairman of the 1996 International VLDB (Very Large Data Base) Conference in Bombay, India. He was also program co-chair of ACM SIGMOD 1985 at Austin, Texas. He is on the editorial boards of Data and Knowledge Engineering (North Holland), Information Systems (Pergamon Press),Distributed and Parallel Databases (Kluwer Academic Publishers), and World Wide Web Journal (Kluwer). He has been an associate editor of IEEE Transactions on Knowledge and Data Engineering and ACM Computing Surveys. He is a member of the IEEE and has lectured extensively in U.S.A, Canada,

Europe, India, Korea and Japan.

Upen Patil received his M.D training in India from the University Of Bombay. He practiced as a clinician at several hospitals in India before coming to United States. He holds two master's degrees, one in Sports Medicine from San Francisco State University and one in Bioinformatics from Georgia Institute of Technology. He also has a strong clinical research background and has published many papers in renowned journals. His experience spans the field of Biotechnology and Healthcare Information Technology that includes his past experience at Perot Systems & PersonalMD Inc. In his current role at XDx as a Clinical Data Manager he is involved with managing and analyzing clinical data from multiple clinical databases. Dr. Upen Patil has worked at various hospitals in California especially in the field of Cardiac Rehabilitation. Currently, he also consults for business research & marketing firms and contributes to market research & analysis for scientific products.

Wei Guan received the Bachelor of Engineering degree in computer science from University of Science & Technology of China in 2003. She is currently a PhD candidate in the College of Computing at Georgia Institute of Technology. Her research interests include bioinformatics, machine learning, text mining, information extraction and information retrieval. Her current research work focuses on the application of machine learning and text mining techniques to the bioinformatics field.