Medical Genomics

Primer on Medical Genomics Part V: Bioinformatics

PETER L. ELKIN, MD

Bioinformatics is the discipline that develops and applies informatics to the field of molecular biology. Although a comprehensive review of the entire field of bioinformatics is beyond the scope of this article, I review the basic tenets of the field and provide a topical sampling of the popular technologies available to clinicians and researchers. These technologies include tools and methods for sequence analysis (nucleotide and protein sequences), rendering of secondary and tertiary structures for these molecules, and protein fold prediction that can lead to rational drug design. I then discuss signaling pathways, new standards for data representation of genes and proteins, and finally the promise of merging these molecular data with the clinical world (the new science of phenomics).

***Mayo Clin Proc*. 2003;78:57-64**

BLAST = Basic Local Alignment Search Tool; cDNA = complementary DNA; CML = chronic myelogenous leukemia; EST = expressed sequence tag; NCBI = National Center for Biotechnology Information; NLM = National Library of Medicine; PSI = Positive Specific Iterative; SOM = self-organizing map

Biologic systems are inextricably entwined with the management of information. Cellular information transfer is crucial at every functional level in biologic systems. Some of the principal areas of information transfer include receptor-mediated signaling, secretion of cellular effectors, and translation of stored genomic information, which generates new proteins and leads to the expression of new phenotypes. The need for computational biology is emphasized in this quote from M. J. Pallen: "Genome sequencing risks becoming expensive molecular stamp collecting without the tools to mine the data and fuel hypothesis driven laboratory-based research."[1]

Bioinformatics includes the establishment of and the information retrieval from databases of molecular sequence and structure information. It also includes the rendering and analyzing of molecular structure information. Bioinformatics extends to cover the generation and analysis of gene expression data. Bioinformatics can be used to facilitate the modeling of molecular interactions and is linking clinical data to gene expression, proteomic, and genomic data, which is the new field of phenomics. Phenomics has the potential to bring molecular biology from the bench to the bedside.

From the Division of Area General Internal Medicine, Mayo Clinic, Rochester, Minn. Dr Elkin is a member of the Mayo Clinic Genomics Education Steering Committee.

This work was supported in part by a grant (LM06918-A101) from the National Institutes of Health's National Library of Medicine.

DEFINITIONS

Bioinformatics is the application of the science of informatics to molecular biology. Informatics is the application of information science, computer science, mathematics, and their associated technologies to a discipline or domain. Finally, biomedical informatics is the application of the science of informatics to the domains of biology and health.

The implications of bioinformatics can be seen no more strongly than in its contribution to the quickly completed first mapping of the human genome (GenBank) (http://www.ncbi.nlm.nih.gov/Genbank/index.html). The scheduled completion of the human genome mapping was advanced by approximately 3 years. (It has been available and in the public domain through the National Center for Biotechnology Information [NCBI] Web site since June 26, 2000.) Rapid identification of homologous regions among single nucleotide polymorphisms contributed to our understanding of the sequence of these genetic fragments. Because molecular biology focuses more on the understanding of information exchange (messaging) within the cell, the field of bioinformatics will continue to be central to the effort. Models of cellular behavior and patterns of information exchange between molecules and within functional biologic systems will be essential for predicting function and dysfunction. Molecular modeling of disease must include a detailed understanding of the systems theory and information (knowledge) management in the subcellular environment.

Sequence Analysis

The application of bioinformatics tools to sequence analysis is perhaps the most well-recognized component of

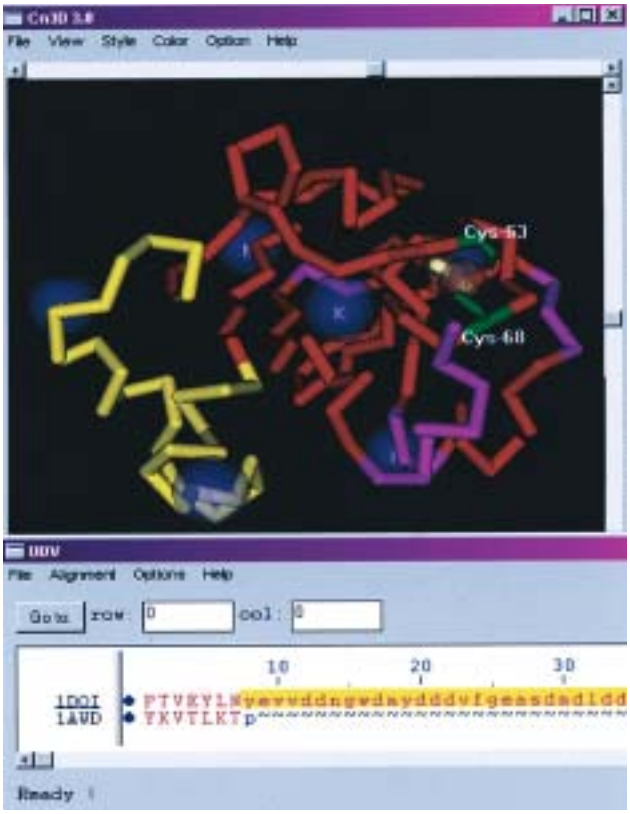

Figure 1. Two aligned ferredoxins and 1D0I, which has a substitution containing a potassium (K) ion–binding site. Cys = cysteine. From www.ncbi.nlm.nih.gov/Structure/CN3D/Cn3d.shtml.

the bioinformatics armamentarium. In this section, I review the Basic Local Alignment Search Tool (BLAST) and its descendant Positive Specific Iterative (PSI)-BLAST. Motif searching and sequence prediction are also described.

These alignment algorithms compare experimentally derived sequences to one or more databases of either nucleotides or amino acids. Homology in these sequences is usually determined by a greater than 30% match. Once a similarity in sequence is identified in one organism, other organisms can also be searched for the same homologue. This activity can generate structural protein families and can yield information regarding evolutionary relationships (ie, they are derived from a common ancestral sequence).

BLAST is designed to identify all similar nucleotide and protein sequences. The algorithm has been optimized to provide reasonable performance with minimal sacrifices in its sensitivity. BLAST 2.0 is the current version of the software and is often referred to as gapped BLAST because it is able to accommodate for intervening introns. The software can be accessed from the Web site of the National Library of Medicine (NLM) of the NCBI (http://www.ncbi.nlm.nih.gov/BLAST). The application allows the user to specify the type of search: genetic sequence to nucleotide databases, amino acid sequences to protein databases, gene sequences to protein databases (translated and mapped), and amino acid sequences to nucleotide databases (remember that DNA ≥ RNA ≥ proteins). Users then can specify whether they want to search all databases (which takes longer) or more specific databases (eg, human genome). The user can specify an E value, which is the number of false-positive matches that are expected to be returned by chance alone. This allows the user to vary the sensitivity and specificity of the search. Weighted results based on statistical algorithms are returned to the user.

PSI-BLAST is a method to extend the sensitivity of a BLAST search. With PSI-BLAST the results of an initial BLAST search (the profile) are iteratively repeated. Highly conserved regions from the initial BLAST search are used to generate new BLAST searches. Poorly conserved regions receive a score of zero. The results of each iteration are used to refine the profile. This continues until a low-level E value is met. In PSI-BLAST, an upper level E value is used to specify the number of false-positive matches expected in the initial BLAST search, and a lower level E value is used to specify the number of false-positive matches expected in the subsequent PSI-BLAST searches. This strategy can identify even weak (subtle) homologies to annotated entries in the database. The sequences must be from globular proteins (which tend to use all 20 amino acids) as opposed to transmembrane regions and signal peptides (which use only a limited set of amino acids).

Another powerful method is motif searching. Protein motifs are sequences associated by known function or structural feature (eg, fold). These databases provide analytic tools that recognize specific amino acid patterns with functional significance. The results of these searches allow the assignment of proteins to functional and structural families. This assists researchers in developing hypotheses regarding the function of novel proteins. Examples of annotated databases that contain protein motifs are Pfam (http://www.sanger.ac.uk/Software/Pfam/) and PROSITE (http://www.expasy.ch/prosite/).

Because the function of a protein depends on its 3-dimensional structure, searching by the structural characteristics of a protein can yield identification of analogous proteins (proteins with differing amino acid sequences but with similar protein folds). These folds can reflect the lock-and-key structure of an enzyme substrate complex. To effect this type of search, one must first predict the protein's structure. Prediction of structure begins with identifying the amino acid sequence associated with the protein. Methods for predicting the arrangement of α-helices and β-

sheets (secondary structure) have been available for decades. The NCBI provides a public domain program for display of the 3-dimensional structure of proteins (Cn3D). The Cn3D can display structure, sequence, and alignment information. It is located on the NCBI Web site at http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml. The sophistication of the publicly available software provided by the NLM to assist molecular biology researchers is impressive. The model shown in Figure 1 contains 2 aligned ferredoxins and 1D0I, which has a substitution containing a potassium ion–binding site. Graphically, it is clear that this potassium ion–binding site is integral in the association of these 2 molecules. Another example of the power of 3-dimensional rendering of the tertiary protein structure is the dopamine D_2 receptor (Figure 2) and its associated genetic sequence (Figure 3). This receptor is an important molecular target in psychiatry. A structural understanding of the dopamine D_2 receptor's 3-dimensional structure can accelerate the design of drugs that affect dopaminergic function by binding to this receptor.

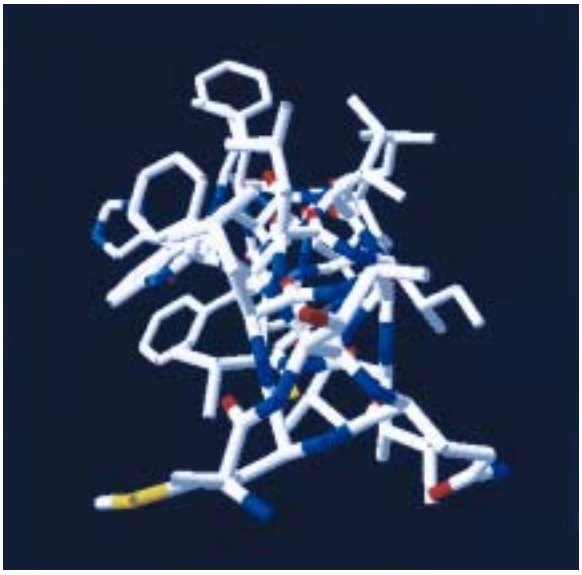

Figure 2. Tertiary protein structure of the dopamine D_2 receptor.

GENE EXPRESSION

The differential expression of genes in healthy and diseased tissue (eg, cancer) can suggest specific genomic markers of disease. This understanding can lead to identification of protein products associated with these genes that may be candidate drug targets. Because microarray technology has become more prolific, we will see an increase in gene expression data associated with patients' health records. These data, when aggregated across populations, will be powerful for identifying both genetic markers of disease and molecular targets for intervention. Microarrays allow researchers to gather thousands of data points regarding individual genes in one experiment.[3] These experiments are usually performed multiple times in both healthy and diseased tissue, and then the gene expression profiles are compared. This generates a considerable influx of data, which require annotation, analysis, and storage. The analysis is most commonly accomplished by using either hierarchical clustering or self-organizing maps (SOMs).[4]

Hierarchical Clustering

This technique uses iterative pairwise gene distance matrices, in which paired values are tested by using a modified Pearson correlation. As multiple experiments are performed, similarities in expression patterns are identi-

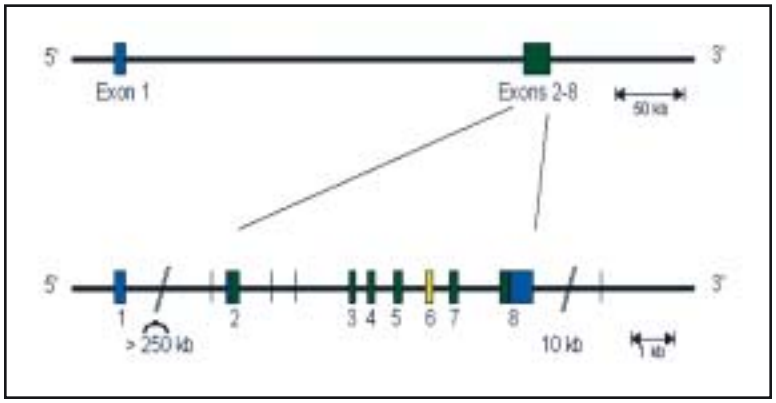

Figure 3. Dopamine D_2 receptor's genetic sequence. This sequence has a number of exons (coding regions) and intervening introns (noncoding segments). The exons are indicated by the colored bars within a relatively small coding region (exons 2-8). Modified from Gandelman et al.[2]

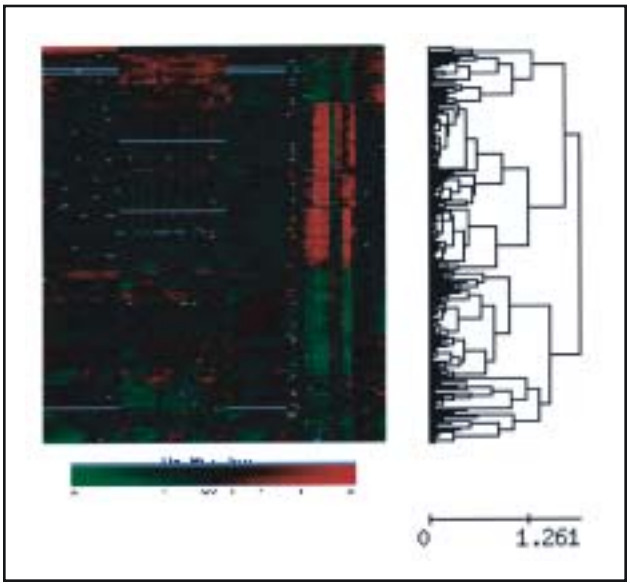

Figure 4. Hierarchical clustering from the output of gene expression data. From http://ep.ebi.ac.uk/.

fied. These groups are then joined iteratively with other groups, and the process continues until a hierarchy of aggregation is constructed (Figure 4). These arrays are built from either complementary DNA (cDNA) libraries or oligonucleotide arrays. The cDNA strands are manufactured by messenger RNA sequences and are often identified by using expressed sequence tag (EST) libraries that are the 3′ or 5′ ends of the cDNA (Figure 5). This information is available in databases such as DbEST and Unigene, both of which can be accessed from the NCBI Web site. The analysis is a 2-step process. First, cDNA target sites are identified by segmenting the digital image into regions and locating each cDNA site within the region. Second, the gene expression is measured by averaging the region's background to mitigate the effects of noise; then one takes the differential intensity averages between the cDNA site and its background as the gene expression level (Figure 6). Oligonucleotide arrays can be designed in silico (by computer), which allows one to exclude sequences showing homology to other genes. Additionally, oligonucleotide arrays can be smaller components of genes, allowing researchers to examine separately individual exons from a particular gene of interest. These results are usually visualized by using a dendogram or tree view (Figure 7). Hierarchical clustering can identify genes that are functionally related. These hierarchies also serve to partition the massive amount of data coming out of the microarray into manageable aliquots. These groups can be sorted according to clinical features, such as certain diseases or conditions that may demonstrate differential expression in some groups as opposed to healthy subjects. However, hierarchical clustering can force unrelated clusters to converge.

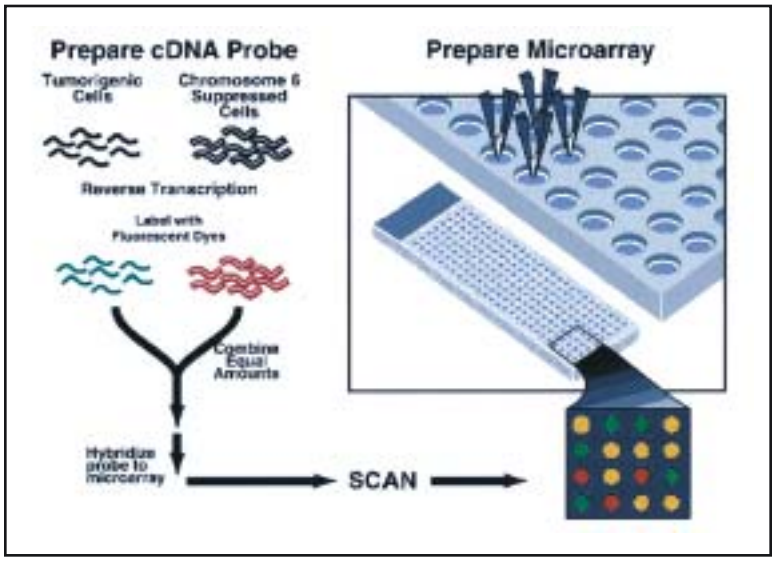

Figure 5. Hybridization of a complementary DNA (cDNA) microarray. The DNA of interest is labeled with a fluorescent dye and fixed (hybridized) to a well in the microarray. This is available to bind the matching RNA should it be expressed in a subject's tissue sample, causing activation of the fluorescent dye. From http://ee.tamu.edu/~camdi/subpages/cdna.html.

Self-organizing Maps

The SOMs are analogous to neural network–based clustering algorithms. They are particularly indicated when one needs to sort messy data, such as data that lack uniformity and contain a high percentage of irrelevant information. This property makes SOMs a useful modality to consider when performing an exploratory analysis. These maps start out by imposing a partial structure on the data. The expression data are iteratively mapped to the closest vector. As each datum is added to a vector, the vector adjusts itself to fit the mean of the data points categorized with the vector. Several passes of all the data are performed to make sure each data point is appropriately classified. The process stops when the changes to the mean vectors are less than a preset value. The vectors then represent the gene expression groupings, which can be sorted according to clinical features or diagnoses.

Gene Filtering

This technique is useful in the analysis of microarray data. Gene filtering differs from hierarchical clustering and SOMs, which are gene-grouping techniques. Instead, gene filtering concentrates on the overexpression or underexpression of genes in abnormal or diseased tissue (one or more biologic conditions) as opposed to their healthy counterparts. For example, one could identify genes that are overexpressed in cancerous tissue as opposed to normal tissue from the same individual. This technique is useful for identifying unusual patterns of expression in which the clustering and grouping techniques are more useful for identifying common patterns of expression.[5]

GENOMIC AND PROTEOMIC DATABASES

Databases of Nucleotide Sequences

The 3 major databases of publicly available information on nucleotide sequences are GenBank (www.ncbi.nlm.nih.gov/Genbank/index.html), EMBL (www.ebi.ac.uk/), and DNA Data Bank of Japan (www.ddbj.nig.ac.jp/fromddbj-e.html). Another useful site is the Human Genome Browser Gateway (http://genome.ucsc.edu/cgi-bin/hgGateway). Although their formats vary substantially, each of these data sets contains identical sequence information. All sites provide tools for the prediction of sequence and the structure information for a submitted nucleotide sequence. Because coding regions account for only 1% to 2% of the human genome, it is important to predict accurately where these coding regions are located. Therefore, accurate structural information is extremely important in the location of new genes.

These databases contain both DNA and cDNA sequences. The sequences that contain only coding regions can be replicated as ESTs, which are single-pass sequences obtained from the 3′ or 5′ ends of cDNAs. The volume of these ESTs makes them ideal for identifying new gene sequences. These ESTs serve to partition GenBank into nonredundant sets of gene-oriented clusters. Several databases of ESTs are publicly available, including dbEST (www.ncbi.nlm.nih.gov/dbEST/) and UniGene (www.ncbi.nlm.nih.gov/UniGene/index.html). UniGene, in particular, serves to provide GenBank clustering and can be useful in gene discovery and probe design for oligonucleotide and cDNA arrays.

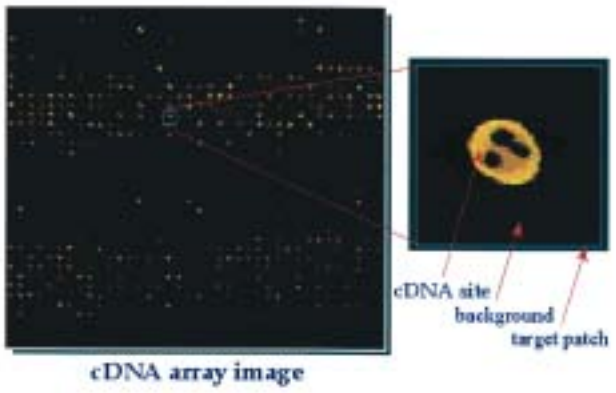

Figure 6. Expanded image of a single complementary DNA (cDNA) microarray well. The intensity of the fluorescence at the cDNA site is compared with the background to determine the relative binding of the subject's RNA and the cDNA probe. From http://ee.tamu.edu/~camdi/subpages/cdna.html.

Databases of Amino Acid Sequences

Many of the publicly available databases dedicated to proteins include both sequence information and tools for

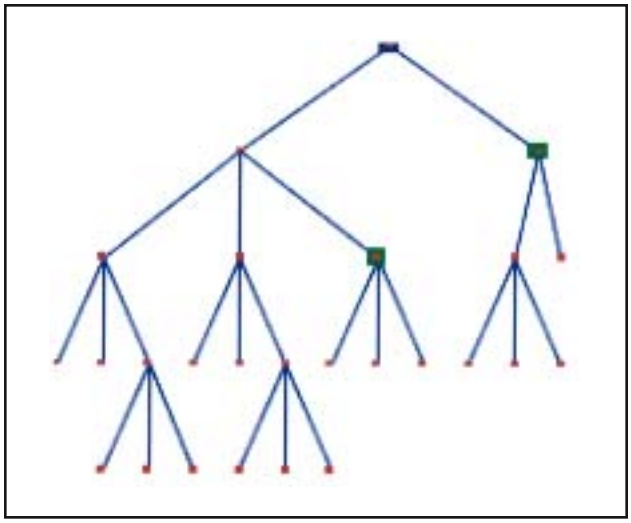

Figure 7. Dendogram output from the hierarchical clustering of gene expression data. From http://www.cwi.nl/InfoVisu/Examples.

Figure 8. Anthrax toxin. The blue-purple region is cleaved, and 7 copies of the remainder of this protein (the so-called protective antigen) form a ring that opens a hole in the cell membrane, thus allowing the other 2 associated proteins, edema factor and lethal factor, to enter the cell. Edema factor works by preventing macrophages from attacking bacteria, and lethal factor causes macrophages to swell and release their toxic chemicals (the chemicals that macrophages use to kill bacteria), which can result in an untreatable toxic shock phenomenon. From http://www.rcsb.org/pdb/. Images by David Goodsell, The Scripps Research Institute.

analysis of these sequences. Swiss Prot (http://expasy/ch/sprot/) is a highly curated protein sequence database with a high level of annotation (function, structural domain, posttranscriptional modification, variants). Pfam (http://www.sanger.ac.uk/Software/Pfam/) contains information on protein families. Orthologs are sequences that are homologues and perform the same function in different organisms. Paralogs are sequences that are homologous in different organisms, which have different functions. Both associations can provide researchers substantial information on the evolution of human genetic and proteomic makeup. Pfam also contains structural information, including prediction of 3-dimensional protein structure. Hidden Markov models can be built from Pfam alignments and are useful for identifying whether a new amino acid sequence contains an existing protein domain.[6] One such program is HMMER, which can be accessed at http://pfam.wustl.edu/hmmsearch.shtml. Another useful site for protein sequence information and analysis tooling is PROSITE (http://www.expasy.ch/prosite/). The Protein Data Bank is a useful site for identifying the 3-dimensional structure of proteins (http://www.rcsb.org/pdb/). It has 17,963 structures as of April 30, 2002. An example of the reconstructions that can be modeled by using the Protein Data Bank is the anthrax toxin (Figure 8).

Protein Fold Prediction

Comparisons with existing fold libraries are made based on the interaction energies needed to create a fold given a known amino acid sequence. The University College London (http://bioinf.cs.ucl.ac.uk/) provides the program THREADER, which can be used to score a candidate amino acid sequence as to its compatibility with known protein folds. These threadings are computed as pairwise interaction energies and potential energies. They are provided as a set of z scores ([Energy – Mean]/SD) for different fold configurations.[6] These tools can assist researchers in identifying analogous proteins. Analogous proteins have different sequences (nonhomologous) but the same or similar folds or functional sites. These proteins are believed to have arisen through convergent evolutionary lines.[7]

Databases of Transcription Factors

These databases allow one to search for protein-binding sites in DNA sequences. Transcription factor databases are particularly useful for identifying transcription factor binding sites. The BioInformatics & Molecular Analysis Section site (http://bimas.dcrt.nih.gov/molbio/index.html) provides the PROSCAN (Promoter Scan) and SIGSCAN (Signal Scan) analysis packages.

Databases of Signaling Pathways

Signaling pathway databases provide information on enzymes, reactions, and substrates. They also contain knowledge regarding regulatory mechanisms. These databases can cluster proteins together, providing a context for the investigation of signaling relevant gene expression at

the level of proteins or RNA signaling.[8] An example of a database of cell signaling networks can be found at the National Institute of Health Sciences in Japan Web site (http://geo.nihs.go.jp/csndb/).

General Information Databases

The Online Mendelian Inheritance in Man catalog (http://www.ncbi.nlm.nih.gov/Omim/) contains the human genes associated with genetic disorders. It is linked to MEDLINE articles and has textual descriptions of genetic disorders, images, and appropriate reference information. This is perhaps the most useful data set for clinicians who wish to learn about genetic disorders or to advise patients about their risk for one or more genetic disorders.

The Kyoto Encyclopedia of Genes and Genomes database holds current knowledge of molecular and cellular biology in terms of the information pathways. These pathways consist of interacting molecules or genes, and the databases also provide links from their encyclopedia to gene catalogs produced by genome sequencing projects (http://www.genome.ad.jp/kegg/). Two other general information databases are the International Society for Computational Biology database (http://www.iscb.org/) and GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). In addition, the Ensembl Genome Browser at the Sanger Institute is a valuable and user-friendly resource and can be accessed at http://www.ensembl.org/.

MOLECULAR DRUG TARGETS

To date, large pharmaceutical companies are concentrating on investigations of drug targets that are of the highest likelihood of producing an immediate yield. These targets include but are not limited to proteases, kinases, nuclear hormone receptors, 7-transmembrane proteins, chemokines, cytokines, and adhesion molecules. Both structure and function information are merged with metabolic and regulatory network information to generate extracted pathways. These pathways are combined with gene expression data to create a scoring of each pathway. Selected candidates are modeled to produce a predicted protein structure that substantially accelerates the process of drug target identification.[9]

This process points the way to rational drug design. One method used in this process begins with isolating a human DNA sequence. This sequence is translated into amino acids. Next, bioinformatics tools are used to search for similar sequences in model organisms (eg, mice, yeast). Modeling the protein structure in the model organism can create a model of the human protein structure. One can identify a drug that binds to the model protein by using docking algorithms. An example of this is the *MLH1* gene, a human gene encoding a mismatch repair protein situated on the short arm of chromosome 3 (drug target). Abnormalities at this locus have been implicated in nonpolyposis colorectal cancer type 2. This was identified by using a yeast model that has an orthologous sequence. Another example is imatinib mesylate, which has been used successfully in the treatment of chronic myelogenous leukemia (CML). Imatinib works by interfering with an abnormal protein (tyrosine kinase) made in CML. Although the molecular target (the bcr-abl breakpoint cluster region on chromosome 22 accepts the translocation [t:9;22] of the c-*abl* oncogene, which encodes the Abl protein, which causes CML) has been known for many years, the drug imatinib (originally STI571) was identified by using a high-throughput screen for tyrosine kinase inhibitors while optimizing its activity for specific kinases. The ability to identify and target specific genetic markers by using bioinformatics tools facilitated the discovery of this drug.

WHOLE GENOME ANALYSIS

The analysis of entire genomes allows the detection of unique sequences (genes) that are not present anywhere in the database or in close evolutionary relatives of a species under investigation.[10] Whole genome analysis can provide key determinants for species-specific phenotypic properties such as pathogenicity in bacterium. The products of these analyses can yield interesting drug targets. Alignments identified can facilitate comparisons of genome organization. Data can be obtained to functionally couple gene clusters among organisms. Whole genome analysis can assist with fusion analysis, the investigation of evolutionary patterns along functional lines. For example, homologues of certain genes appear to fuse during evolution, the identification of which can also assist in prediction of gene function.

CHALLENGES IN BIOINFORMATICS

Many challenges remain for bioinformaticians, some of which include creating simulations from molecules to populations. Creating gene networks and modeling signaling pathways remain a challenge. There is a need for improved data structures and database support for biomedical investigation in this increasingly data-intensive environment. Another area of active research is whole cell modeling, in which bioinformaticians are working on developing models to simulate the wide variety of intracellular interactions.[11]

On the horizon is the field of phenomics, the study of the expressed clinical state and its relationships to the genomic and proteomic data being accumulated. To accomplish this linkage, groups have started work on formal representations of gene ontologies, which will serve to improve the

reliability of the data used in these linkage studies.[12] This requires a detailed compositional understanding of the information in our health records.[13] Health records are primarily stored in free text. The information that is hidden (with respect to automated inquiry) in our health records needs to be converted into data if we are to accomplish the goals of phenomics. An intelligent data design can facilitate our ability to obtain answers to clinical genomics and proteomics questions.

Groups are now working on biological computing, in which data are stored as nucleotide sequences. This approach effectively uses genes as an information storage media. This novel approach is touted by some to be a much more efficient and responsive method for data storage and computing. Adleman et al[14] described one such effort in which they used biological computing to perform data encryption.

CONCLUSIONS

Bioinformatics is a rapidly developing field that brings the world of genomics and proteomics closer to the clinical desktop. The work of bioinformatics has contributed substantially to the rapid mapping and current understanding of the human genome. Current work in bioinformatics is assisting researchers in their proteomic investigations. Bioinformatics research is now focusing on simulations and modeling of cellular and subcellular messaging, which has the promise of improving our understanding of intracellular signaling. Future work includes a focus on an improved understanding of the semantics of our clinical record to serve the needs of the phenomics community.

REFERENCES

1. Pallen MJ. Microbial genomes. *Mol Microbiol*. 1999;32:907-912.
2. Gandelman KY, Harmon S, Todd RD, O'Malley KL. Analysis of the structure and expression of the human dopamine D2A receptor gene. *J Neurochem*. 1991;56:1024-1029.
3. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genet*. 1999;21(1, suppl):33-37.
4. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*. 1999;96: 2907-2912.
5. Wu TD. Analysing gene expression data from DNA microarrays to identify candidate genes. *J Pathol*. 2001;195:53-65.
6. Starkey MP, Elaswarapu R, eds. *Genomics Protocols*. Totowa, NJ: Humana Press; 2001. Methods in Molecular Biology series; vol 175.
7. Aravind L. Guilt by association: contextual information in genome analysis. *Genome Res*. 2000;10:1074-1077.
8. Kaminski N. Bioinformatics: a user's perspective. *Am J Respir Cell Mol Biol*. 2000;23:705-711.
9. Fagan R, Swindells M. Bioinformatics, target discovery and the pharmaceutical/biotechnology industry. *Curr Opin Mol Ther*. 2000;2:655-661.
10. Searls DB. Bioinformatics tools for whole genomes. *Annu Rev Genomics Hum Genet*. 2000;1:251-279.
11. Tsoka S, Ouzounis CA. Recent developments and future directions in computational genomics. *FEBS Lett*. 2000;480:42-48.
12. Ashburner, M, Ball CA, Blake JA, Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000; 25:25-29.
13. Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. *Proc AMIA Symp*. 1998;765-769.
14. Adleman LM, Rothemund PW, Roweis S, Winfree E. On applying molecular computation to the data encryption standard. *J Comput Biol*. 1999;6:53-63.

Primer on Medical Genomics Part VI will appear in the March issue.