CHAPTER EIGHT

Bioinformatics: Its Role in Drug Discovery

DAVID J. PARRY-SMITH
ChiBio Informatics
Cambridge, United Kingdom

Contents

1 Introduction, 334
2 Bioinformatics in Drug Discovery, 335
3 What is Bioinformatics?, 336
  3.1 Definitions, 336
  3.2 Integration of Information, 336
  3.3 Bioinformatics and Skills, 336
  3.4 Standardization, 337
4 Bioinformatics and Target Discovery, 338
  4.1 Functional Genomics and Target Discovery. 338
  4.2 Expression Profiling for Target Discovery, 340
    4.2.1 EST Profiling, 341
    4.2.2 Sequence Assembly for cDNA Cloning. 342
    4.2.3 Comparing ESTs with Databases, 342
    4.2.4 Statistics for Assessing Expression Level Significance, 342
    4.2.5 Genome-Wide Expression Analysis, 344
5 Databases, Tools, and Applications, 345
  5.1 Databases, 345
  5.2 Sequence Comparison, 346
  5.3 Phylogenomics and Gene Family Databases, 347
6 The Bioinformatics Knowledge Model, 349
  6.1 Data, Information, and Knowledge, 350
  6.2 The Hierarchical View, 350
  6.3 The Scientific View, 350
  6.4 Data is Not Knowledge, 350
  6.5 Drug Targets, 351
7 Structural Genomics, 352
  7.1 Predicting Protein Function from Structure, 353
  7.2 Neural Networks and Protease Function, 353
  7.3 Fold Compatibility Methods, 353
  7.4 CASP and the State-of-the-Art, 353
8 The Future, 354
9 Acknowledgments, 354

Burger's Medicinal Chemistry and Drug Discovery
Sixth Edition, Volume 1: Drug Discovery
Edited by Donald J. Abraham
ISBN 0-471-27090-3 © 2003 John Wiley & Sons, Inc.

1 INTRODUCTION

In January **2001,** the biopharmaceutical company Millennium announced that, as part of a multimillion-dollar research collaboration with Bayer, an anticancer agent had proceeded to clinical trials (1, 2). The remarkable achievement was not that the collaboration between a pharmaceutical company and a high technology genomics company had been so successful in terms of a product, but that it had ostensibly lopped **2** years from the discovery lifecycle in the process. **As** a result of perceived improvements in research efficiency such as this, more effort than ever is being placed in development and implementation of genomics and screening technology automation. The substantial volumes of data thus generated are used to pan for innovative lead compounds for novel therapeutic targets to feed ever more voracious development pipelines **(3).**

The genomics tools of rapid sequence screening, microarray chips, expression analysis, protein interactions, macromolecular structure determination, sequence comparison, and a host more are all biological techniques that generate different types of raw data. Moreover, the data are produced in far greater quantity than has been seen in biology before and the sheer speed with which the data are produced is unprecedented. The improvements in process throughput, which are so exciting to the financial markets and so critical to the alleviation of pain and suffering in human populations, are readily achievable because of the science of information integration and knowledge transformation that are the hallmarks of bioinformatics. It is not enough simply to produce data, even from the most leading edge of techniques; we must be able to manage it effectively and extract useful information that leads to that critical knowledge on which realistic drug discovery decisions can be made.

The opening example illustrates how genomic technologies, including bioinformatics, are making an impact at the drug discovery level. The purpose of this article is to provide some background to the tools and technologies that are used on a daily basis by bioinformatics scientists in an effort to make the subject more accessible to readers with a non-biological training. This article is not, however, a training manual in running programs nor is it an index to the latest resources on the Internet. Both bioinformatics and the World Wide Web grow at such a pace that any journal article is likely to be out-of-date before it goes to press. Nevertheless, the Internet is a crucial resource for the practicing bioinformatics researcher. All of the core uses for executing projects are available there both in terms of databases and computer programs. Lists and descriptions would fill many volumes. Certain resources are mentioned where necessary to illustrate specific examples. Some useful starting points for an exploration of bioinformatics on the Internet are shown in Table 8.1. **An** introduction to tools and techniques is available in Ref. 4, while Ref. **5** contains a more technical approach based on understanding of machine learning. For a more mathematical treatment, readers are referred to Ref. ***6*.** Reference may be made to the journal ***Briefings in Bioinformatics*** for reasonably. accessible descriptions of systems and processes. The journal ***Bioinfomatics*** is aimed at a more technical audience. The annual database issue of *Nucleic* ***Acids Research*** is the generally accepted place for publication of bioinformatics databases. Some fundamental papers referred to are relatively old for a young science.

Rather than delve into minutiae here, the aim has been to present an overview of the way bioinformatics is being used in the process of drug discovery in **2002.** Many important and exciting aspects of the academic research that is being carried out are passed over in silence or referred to only by a brief comment. Computer technologies change rapidly and bioinformatics has always been at the forefront in applying new computing paradigms to biological problems (e.g., use of the Internet, object-oriented programming technologies, neural networks, parallel computing). Some of the molecular biology or bio-

Table 8.1 A Selected List of Key Websites for Further Exploration of Online Bioinformatics Resources

Internet Site	Brief Description
http:llwww.ncbi.nlm.nih.gov	The National Center for Biotechnology Information (NCBI). Located at the National Library of Medicine in Bethesda, MD, USA. The home of the GenBank DNA sequence database; PubMed literature search engine; sequence search tools (e.g., PSI-BLAST); genomic sequence navigation tools. A substantial repository of resources in all areas of bioinformatics.
http:llwww.ebi.ac.uk	The European Bioinformatics Institute (EBI). This site is located at Hinxton Hall, Cambridge, UK. The home of the EMBL Nucleotide Sequence Database; data management tools [including publicly accessible version of SRS—the Sequence Retrieval System (7)]; protein family databases; microarray tools; etc. An extensive repository of resources for bioinformatics.
http:llwww.expasy.ch	The Expert Protein Analysis System. Dedicated to the analysis of protein sequence and structure as well as two-dimensional PAGE. Home of the SWISS-PROT protein knowledgebase and TrEMBL computer annotated supplement (8).
http:llwww.man.ac.uk	The University of Manchester Bioinformatics Education and Research site (UMBER). Useful because it is the home of the PRINTS (9,10) resource for protein fingerprint analysis and a valuable teaching site for bioinformatics.
http:l/www.ensembl.org	Site developed by the Sanger Centre, Hinxton Hall, Cambridge, UK and the EMBL-EBI, presenting tools for browsing and researching the human genome sequence (11).This is a public access server providing data and access at no charge. Commercial sites are also available for working on commercially produced human genome sequence.
http://www.mips.biochem.mpg.de	The Munich Information Center for Protein Sequences (MIPS). Provides a different view of several model organism genomes with tools for analysis.

physical aspects of techniques used to generate data are outlined from the point of view of a bioinformatician and not that of a practicing molecular biologist (although it has been reviewed by one, see the acknowledgments) to give a flavor of the kinds of experiments that are performed.

2 BIOINFORMATICS IN DRUG DISCOVERY

Bioinformatics in drug discovery has traditionally been used as a tool for finding new drug targets ("target selection"). This technique of target discovery is an important contribution that bioinformatics has been able to make to the drug discovery process. However, bioinformatics alone without the background of molecular biological and biophysical experiments is sterile. To understand even a bird's-eye overview such as this, some basic material has to be covered to enable comprehension both of the data and the manner of its analysis.

The second use to which bioinformatics has been put in the drug discovery process is more fundamental. It is concerned with the use of techniques in molecular sequence analysis to generate relationships between sequences that are themselves used to provide fundamental structures for databases of drug discovery information. Relationships between data elements are important because they help to place individual elements in a context that can be readily assimilated by the user of the system. In many situations, observers approach data from different points of view and bring to bear the richness of differing scientific experiences. Whether we care to admit it or not, "biologists" and "chemists" have dif-

ferent training and background and thus offer a range of opinions on similar pieces of information. Even the word "activity" means different things to a chemist who has synthesized a group of compounds or a biologist who has developed an assay to test the compounds. Both aspects are necessary for the discovery of new drugs, but they are different viewpoints that need to be supported by appropriate relationship mining in the data. If the bioinformatics job is done well, both views can be accommodated in the data structures and user interfaces used by both sets of users.

Throughout the pharmaceutical industry, bioinformatics and chemoinformatics groups are working closer together than has been the case hitherto. This is a consequence of the realization that managing data effectively requires integration of thinking (about definitions of common attributes of molecules both small and large), integration of processes, and integration of implementation. The recent rise in popularity in bioinformatics of the ontology is an example of the application of a computer science paradigm to the issue of redundancy in nomenclature in many areas of biology. Application across the chemistry–biology domain interface could well be beneficial for drug discovery effectiveness. The ontology is simply a means to an end, in this instance, that end is improved communication and understanding of basic concepts within and across the boundaries of major scientific disciplines. There may, of course, be a variety of other means to reach that goal.

3 WHAT IS BIOINFORMATICS?

3.1 Definitions

Concisely, bioinformatics is our ability to organize biological data. From another perspective, bioinformatics is our ability to understand how biological information is organized. From this understanding should spring an enhanced view of the interactions between biological molecules. This should, in turn, inform our search strategies for new small molecules that will modulate the behaviour of biological molecules to give a beneficial therapeutic effect. These definitions arise from observation of the way diverse skills are brought to bear in attempting to answer biological questions. They also stress the importance of organizing and understanding biological data, rather than linking these aspects strongly to specific hardware or software implementations. Use of computers may be involved in the process but the definitions are not limited by the application of any particular technology.

Bioinformatics has also been defined as the application of computer technology to solving biological problems. This definition, perhaps what some would consider to be the canonical one, is broad but restricts the scope of the definition to problems to which computer technology can be applied.

3.2 Integration of Information

Bioinformatics has become a byword for integration; specifically the integration of data across different data resources to generate integrated information resources. Linking data and information in this way is fundamental to bioinformatics activities and so some discussion of the meaning of data, information, and knowledge in the context of bioinformatics for drug discovery is provided in Section 6. Integration is important because it provides context, or at least a background, against which computational analyses are performed. In the past, for single molecule experiments, this background was achieved through reading the literature. Now that multiple molecule experiments are common, even genome-wide or inter-genome analyses, it is simply not practical any longer to rely on the literature in its raw form, unless it is part of an integrated knowledge-based approach that provides connections between disparate pieces of information, backed up by experimental evidence from which to draw conclusions (12).

3.3 Bioinformatics and Skills

The pursuit of bioinformatics involves a number of different skills. Organizing, storing, retrieving, and querying sets of biological data are techniques that lie at the heart of the subject. An ability to analyze the characteristics of particular sets of biological data is fundamental. The translation of those characteristics into electronic representations that can be organized on a large scale is the domain of the

bioinformatics software and database developer. The process of analyzing and understanding biological data using the tools available is the domain of the bioinformnatics analyst. When new tools are in the course of development, substantial interaction between the two skill sets is essential.

In the pharmaceutical environment, both developer and analyst skills are necessary. This is so even where commercial software is in use, because there is no single system available commercially that provides the level of integration between the worlds of bio- and chemoinformatics necessary to effectively enhance the drug discovery process. Some interfacing of different systems is required and the warehousing of proprietary data is always an issue.

This broad description of bioinformatics and of the two types of bioinformatics scientist is quite abstract. It does not detail the characteristics of the data with which the bioinformatics scientist has to work. Neither does it define the set of tools that the developer should work with or implement. There was a time, in the late 1980s and early 1990s, when the type of data was well defined. Molecular sequence data, the stream of bases in DNA, and the stream of residues at the protein level were the main types of data. Programmers developed code in FORTRAN or C and scripting languages were immature.

Now, as science moves forward into a new millennium, additional types of data have become important; for example, protein–protein interactions and three-dimensional structure, high density gene expression chips, cell imaging, etc. Developers have a wide range of tools to call on, including high performance C and C++ compilers, rich scripting languages (Perl, Python, etc.), and efficient, easily accessible operating systems (particularly Linux) that make porting software to different hardware platforms less of an issue than it was.

Of equal importance to the medicinal chemist, to whom this review is principally directed, is the impact of bioinformatics on the discovery of new medicines. Rather than explain comprehensively all the popular tools and their underlying algorithms, this review focuses on the points in the discovery research process where bioinformatics is making an impact. Technologies will be described to the extent that such understanding is necessary to grasp the relevance of the data being generated and its significance.

3.4 Standardization

Progress in linking items of relevant data and generating integrated information resources would be very limited were it not for efforts in standardization that have been brought about by international collaboration. There is still a long way to go, however. While it is becoming cheaper to obtain each piece of individual data, the proportion of automated experiments is increasing, at least in the life sciences, because of the ready availability of new technologies. It may seem a simple matter to create resources that store and manage streams of DNA bases—represented by the four alphabetic characters A, C, T, and G. However, when we also wish to integrate information on experimentally or computationally determined annotation and cross-reference to other resources using gene names, there are significant problems. The literature abounds with synonyms for gene names and functions; even the labels given to specific cellular functions are not always clearly defined (13).

To be able to process data automatically, it has to be presented in a form that can be parsed by a computer program and must also include all the elements necessary to an understanding of the biological system under study. Reliable information systems should have source data of a consistently high quality to prevent application errors and enable integration into other biological information systems. Some progress is now being made towards consensus in gene naming through the work of the HUGO Gene Nomenclature Committee (see http://www.gene.ucl.ac.uk/nomenclature/). Many researchers now use this system as a source of unique gene names and descriptions in the published literature (14) and in commercial products (e.g., see http://www.biowisdom.com/). Standardizing vocabulary expressing the relationships between the complex network of gene functions is the work of the Gene Ontology (GO) project (see http://www.geneontology.org/).

4 BIOINFORMATICS AND TARGET DISCOVERY

The desire to find new drug targets is grounded in the need of pharmaceutical companies to address the requirements of different disease markets. The literature is full of papers detailing sequence determinations of newly cloned receptors and enzymes along with research on their functional properties. The realization that the sequences of genes could be acquired relatively cheaply through the use of automated sequencing machines using fluorescent base technology, rather than the previous generation of radioactive sequencing gels, meant that sequence data began to flood the public DNA sequence databases. The growth of these databases is reviewed in Fig. 8.1.

Because the translation of DNA coding sequence to protein sequence is straightforward, given the understanding of the genetic code, it is a trivial task to implement software to provide translations of open reading frames (ORFs[1]) to be housed in the annotation sections of the DNA databases. Consequently, protein sequence databases have become swamped with hypothetical proteins—those proteins assumed to exist because open reading frames have been discovered and from which hypothetical protein sequences had been computed. The function of these sequences has been assigned through a comparison of their amino acid residue similarity with that of known sequences (for example, those that have had their biochemical function demonstrated through heterologous expression). In this way, very large numbers of sequences have been processed into the databases and annotated using such sequence comparison techniques.

When it comes to the practical details of how bioinformatics can speed the process of drug discovery, it is reasonable to ask what sorts of data could be valuable in that process. The stages of the drug discovery process where bioinformatics makes an impact are target identification, assay selectivity panel selection, and integration throughout the assay development and screening process. Target identification makes use of sequence data for functional assignment by inference from similarity with known sequences (Fig. 8.2). It also benefits from assessments of differential levels of expression in different cellular contexts and at various stages in the expression process (transcriptome or proteome, see Fig. 8.3 for definitions). Selectivity panel selection relies on a thorough mining of the related gene family and may benefit from phylogenomic analysis (see Section 5.3). Use of bioinformatics for integration of data capitalizes on the generation of relationship information between known genes and the ability to use hyperlinking to create navigational tools and usable interfaces.

A further development, the production of millions of short expressed sequence tags (ESTs), encouraged the focus on target discovery during the 1990s. The development of EST technology itself spawned the genesis of several new genomics companies, including Human Genome Sciences and Incyte, which have worked in collaboration with the pharmaceutical industry to hunt for new disease related targets.

4.1 Functional Genomics and Target Discovery

The collation of gene sequence data, from whatever source, is in effect simply a matter of transferring data from one place to another; for example, from a sequence chromatogram to a computer database. Learning the sequence of a genome, or any of its constituent parts, is a long way from understanding its biological function. The sequencing of genomes has resulted in a technical genome description (at a particular level of detail) through the process of cataloguing an organism's genes. This level of detail is often called the physical map of the genome. There are other ways of mapping the genome that provide different levels (or we may think of it as resolution) of genomic detail; genetic maps indicating the location of genes for specific traits have been known for some time, while single nucleotide polymorphism (SNP) maps can be used to highlight the positions of differences between populations through study of genetic polymorphisms (15). Indeed, the identification

[1] ORFs are contiguous strings of residues, uninterrupted by the genetic code's "stop" signal.

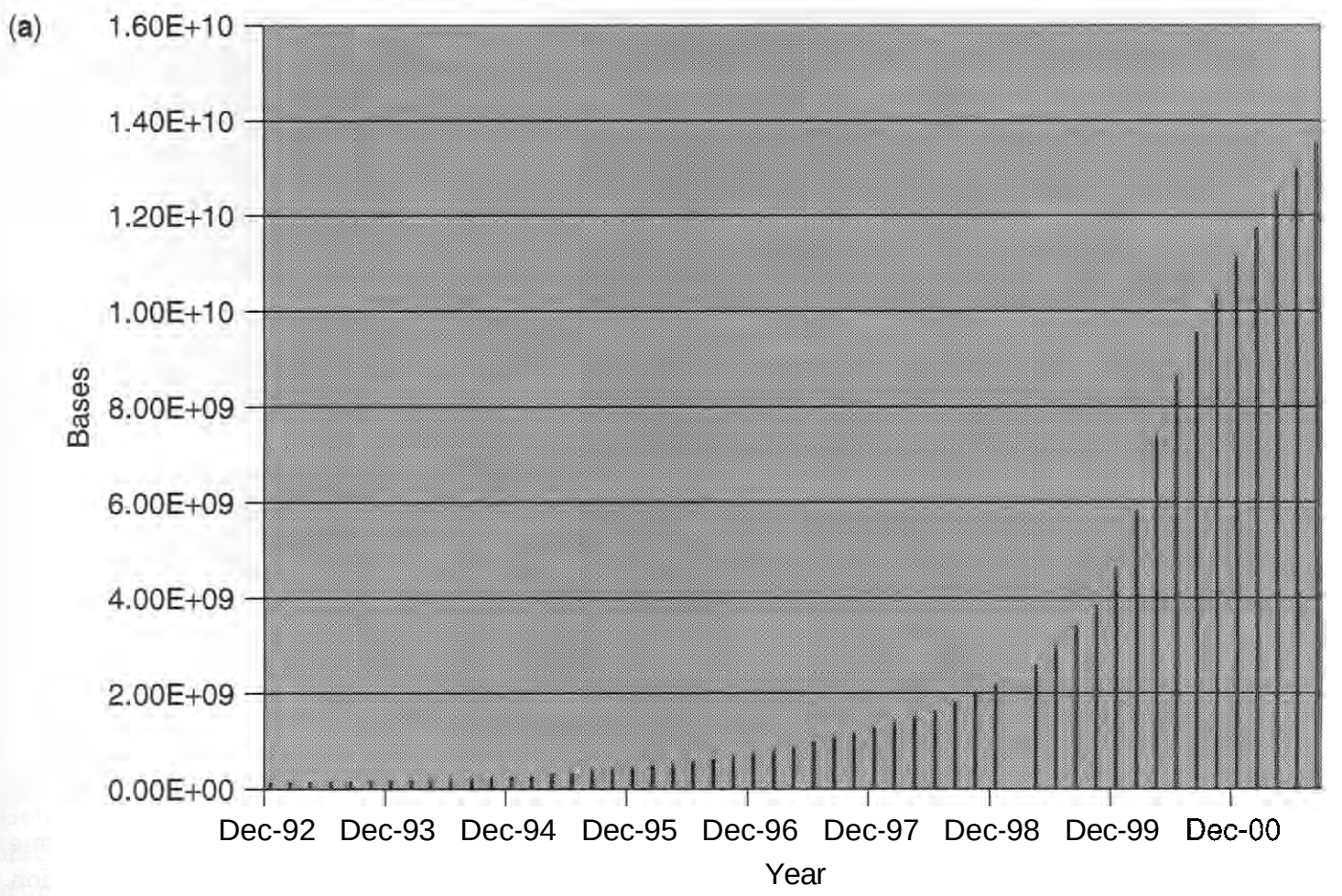


Figure 8.1. Bar charts indicating the growth of GenBank from December 1992 to August 2001 in terms of (a) bases and (b) sequence entries. The release files indicate no release in February 1999. It is evident from the trends in both charts that while there has been explosive growth, particularly from December 1999 until about August 2000, growth is slowing. The base entry curve is showing a distinctly sigmoid shape.

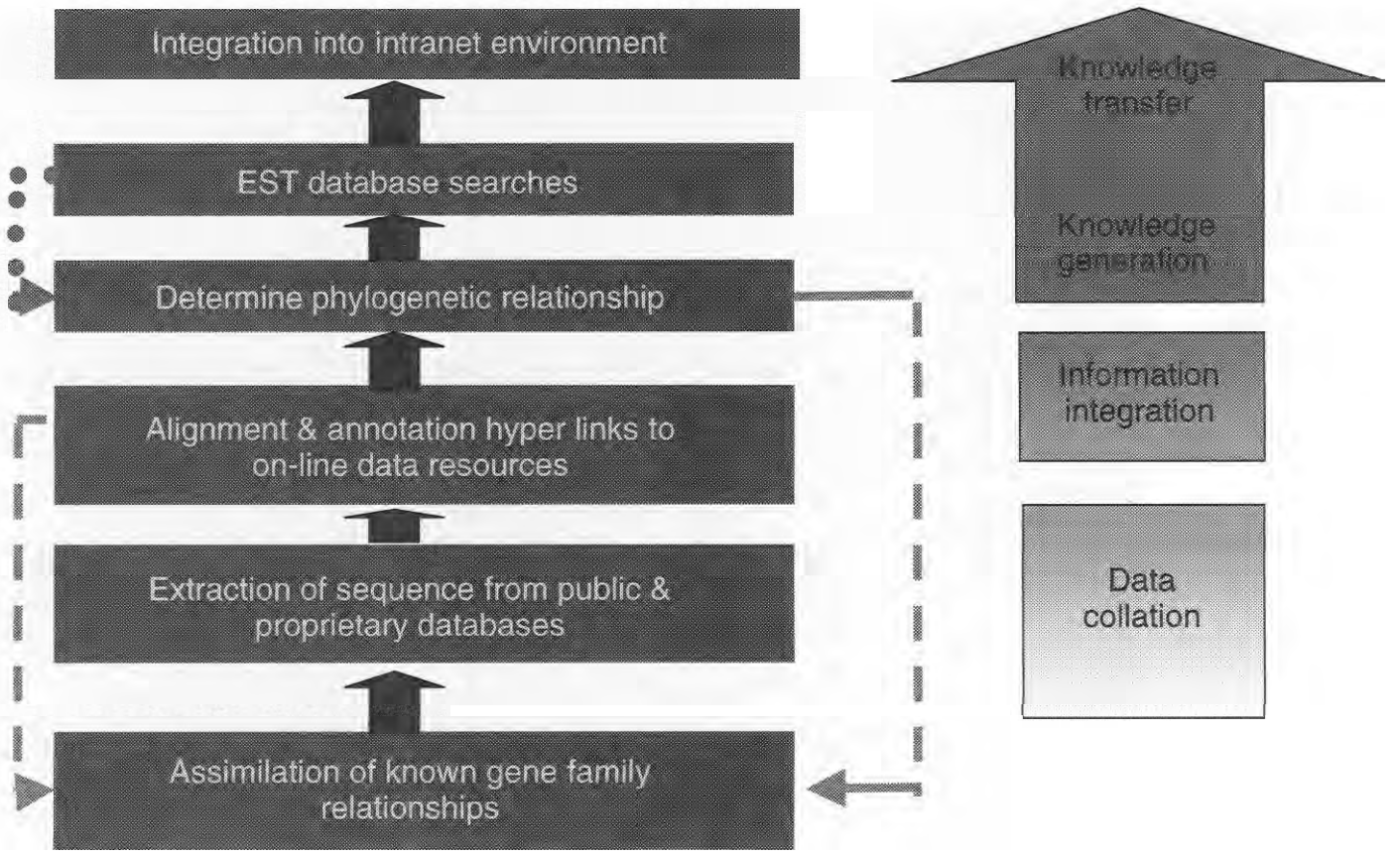


Figure 8.2. **A** schematic illustrating the bioinformatics process required to create an online gene index by collating data and then integrating related elements to generate value added information through hyperlinking to online resources. Determination of phylogenetic relationships is a relatively late stage in the process. EST analysis is only performed after phylogenetic relationships have been determined because EST data does not cover the whole expressed sequence and may not therefore cover regions that were included in the phylogenetic determination. This is a knowledge generation phase because it is allowing placement of potential new targets within the context of a carefully researched phylogenetic tree. Transfer of knowledge is intimately related to the environment in which the results of analysis are made available, in this case as an online resource.

of genes themselves from genomic sequence is itself a non-trivial matter, especially where those genes are interrupted by non-coding regions (introns) and control regions (expression promoter sites). Functional genomics is the process of creating an understanding of the way genomes function through gene expression. Genes are expressed by a variety of mechanisms, not all of which are fully understood. We can, however, make some measurements of the results of gene expression at the transcript level, mRNA, and at the protein level. Several of the techniques that have been used to assist drug target discovery are presented in the following sections.

4.2 Expression Profiling for Target Discovery

Bioinformatics spans analysis in depth on small quantities of data through to expansive genomic scale analyses, which may be at a lesser level of detail. Historically, the expression of genes at the mRNA level or at the protein level has been a crucial tool for assessing the significance of specific classes of cells as targets. With the advent of fluorescence-based sequencing techniques and automated sequencing technology, it is now much quicker to generate sequence data on specific molecular targets than ever. Many researchers spend entire careers working on one target type or a restricted part of a target gene family. This approach has yielded many valuable targets for drug discovery. With the new technologies of molecular biology, it is now possible to survey targets in a variety of contexts: perhaps within different types of cells, cells treated with different agents, or even across entire genomes using chip technologies.

There are issues of interpretation of experimental design and results. Does mRNA expression mean anything at a quantitative level? Perhaps even a qualitative view of mRNA expression can be misleading. How is mRNA expression correlated with protein ex-

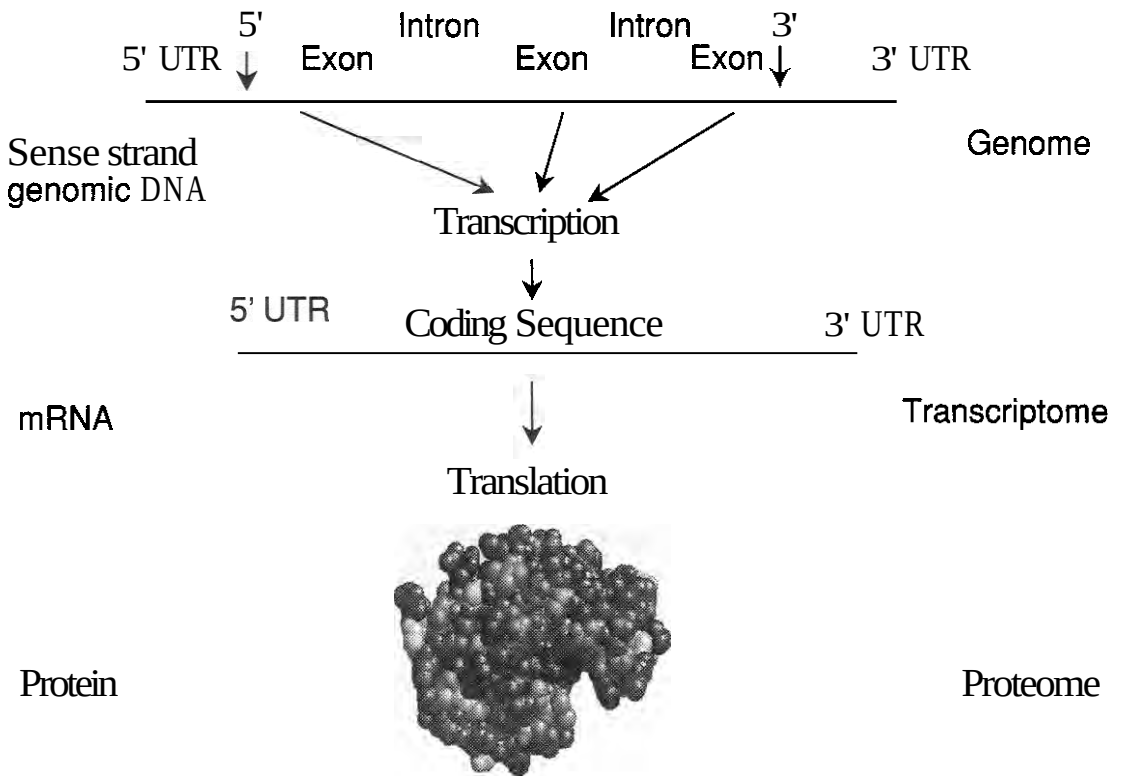


Figure 8.3. A schematic illustration of the relationships between levels of genomic information. Genomic DNA is contained in the nucleus of eukaryotic cells. In many species, including humans, information required to make up the coding sequence of a gene is split into exons (regions that are expressed) interrupted by introns (regions that are not expressed and are edited out of the message at the transcription step). At either end of the gene sequence are untranslated regions (UTRs). **5'** and **3'** refer to the orientation of the strand of DNA as defined by the sugar-phosphate backbone. The **mRNA** is the messenger RNA molecule generated by the process of transcription, which is itself mediated by a number of enzymes. The collection of mRNA transcripts that make up the mRNA expression profile of a cell is known as the transcriptome, although the term could also refer to the total possible mRNA transcripts achievable from a genome. Finally, translation of the mRNA occurs on the ribosome and protein sequence is produced, which folds into its final three-dimensional shape—a process that may be assisted by a number of different chaperone proteins. Any post-translational modifications are all part of the proteome, the collection of proteins that represent the expressed genome.

pression? In general, most drugs we discover are likely to interact with proteins and not mRNA, so some understanding of protein expression is an essential adjunct to our genomic knowledge. Hence the proteomics approaches described later. Our exploration of expression profiling begins with a study of mRNA transcript profiling using expressed sequence tags because this technique has led to rapid gene discovery that has, in turn, been able to assist with the annotation of genomic sequences. Then, we consider how whole genome expression profiles can provide a rich new source of data for bioinformatics analysis.

4.2.1 EST **Profiling.** An EST is a short, single sequence run collecting data over about 200–400 bases from a clone selected from a cDNA library. Typically, cDNA clone libraries contain **1–3** million clones. The library itself is created from mRNA extracted from tissue or refined cell populations. By making a random selection of several thousand clones from a cDNA library it is possible by sequencing ESTs to generate a rapid, if somewhat low resolution, survey of the types of genes represented by the library. The library in turn reflects the composition of genes that are expressed in the tissue or cell line from which it was constructed. Thus, we have a qualitative link between gene expression, at the mRNA level, and the sequence level analysis required for target identification, without the need to go through the full sequencing and validation process across the whole length of each clone. This is a very significant time and cost saving. One of the major issues of EST profiling has been the significance that can be

ascribed to expression levels through counting copies of ESTs. This issue is dealt with in some detail in Section 4.2.4.

4.2.2 Sequence Assembly for cDNA Cloning. To appreciate fully the speed advantage of sequencing tags, rather than fully validating the sequence of an entire clone of a gene, it is useful to step through a brief description of the cloning process from the point of view of a practitioner of bioinformatics.

Sequence assembly is the process of dealing with the bioinformatics of cloning and genomic sequencing (16, 17). When a gene is cloned, it is selected from a set of potential clones in a cDNA library. The gene is present as a piece of cDNA inserted into a cloning vector (a piece of circular DNA) that has been designed for the purpose of cloning. It is necessary to check that the cDNA indeed represents the sequence of the gene that has been cloned. To do this, DNA oligonucleotides are designed that will bind in a complementary fashion (hybridize) to the DNA of the cloning vector and also at 150- to 200-base intervals along the cDNA itself. These oligonucleotides are then extended by adding a base that is complementary to the cDNA insert by using a DNA polymerase. Different polymerases are available commercially that provide high fidelity reproduction of the cDNA insert. In fluorescence-based sequencing, a small proportion of the nucleotides available to the polymerase are fluorescent analogues. Incorporation of one of these into the oligonucleotide terminates the extension, resulting in a population of oligonucleotides of different lengths. These are separated by electrophoresis and the sequence determined.

When the sequence of each oligonucleotide has been determined, the strings of letters that represent the bases are assembled together to generate a full-length sequence of the cDNA that has actually been cloned. Errors in the base sequence can be resolved at this stage, and if necessary, mutagenesis experiments designed to correct any mistakes. The bioinformatics process is intimately linked with the molecular biology techniques of cloning and sequencing. For target discovery, a very high degree of confidence in the sequence of the cDNA clone is required before the clone can be expressed and used in assay development. It can also be seen that this process is much lengthier than taking a single sequence read (a single oligonucleotide string) without correcting errors or considering coverage of the complete gene sequence.

4.2.3 Comparing ESTs with Databases. Bioinformatics provides the tools necessary to compare each EST with the databases of known genes and a hypothetical functional assignment may be made to a proportion (typically 40–50%) of all the ESTs from a sequencing run.

In this way a rich resource of tags for many clones from many diverse libraries has been built up in the public domain and in commercially available, proprietary databases. One particular approach that generated much interest in the 1990s was that advocated by Incyte. Here, the simple identification of a gene expressed through identification of its EST was not the primary goal. Instead, the approach was based on comparative transcript expression (the so-called "digital Northern"). Here the number of copies of each EST identified was calculated, giving counts for the numbers of each type of EST found in comparing normal with diseased tissue, for example. Subsequent techniques have focussed on more controlled experiments in which specific cell lines are treated with an agent and the expression of genes before administration is compared with the profile afterward. This is the basis of pharmacogenomics (18).

4.2.4 Statistics for Assessing Expression Level Significance. There are issues with approaches based on counting the number of copies of an EST observed in the output from a sequencing machine. First, the tissue or cell line must be of very high quality and the mRNA harvested in a timely manner because it degrades very quickly. Second, the process of preparing the cDNA library should enable the numbers of clones to be estimated as accurately as possible. Third, the random sampling for the sequencing runs must be controlled carefully so as not to introduce bias into the experiment. The mathematical model for evaluating the meaning of data from such experiments is not well worked out.

Comparison of the differences between the

Table 8.2 Comparative EST Counts for Five Genes Sequenced from Normal Prostate, Stage B2 Cancer, Stage C Cancer, and Benign Prostatic Hyperplasia (BPH) cDNA Libraries

Gene	Normal Prostate Total	Stage B2 Cancer		Stage C Cancer		BPH		All Other Tissue
		Tags	P	Tags	P	Tags	P	
PSA	13	7	0.7–0.8	14	0.6–0.7	22	0.8–0.9	0
PAP	4	1	0.1–0.2	34	>0.999	9	0.7–0.8	1
HGK	1	7	>0.999	6	0.97–0.98	5	0.8–0.9	0
PS1	0	3	0.993–0.994	7	0.997–0.998	1	0.4–0.5	0
PS2	0	2	0.97–0.98	7	0.997–0.998	0	0–<0.1	0
Total clones	4500	1400		3400		4800		732,000

The tag counts are from Ref. 21. The P values are calculated according to Equation 8.1, modified for use with different total EST counts from the source libraries. The web URL http://igs-server.enrs-mrs.fr/~audic/egi-bin/winflat.pl was used to calculate the probability intervals. A P value nearer to 1 indicates that the differential expression is likely to be significant. While prostate specific antigen (PSA) and glandular kallikrein (HGK) have been proposed as prostate cancer markers, both PS1 and PS2 are prostate specific. Thus, the down-regulation of PAP in stage B2 cancer is not significant using this test, whereas, the test shows its up-regulation in the BPH sample to be more significant. So, for lower changes in copy number, where more sensitivity is expected, this test of significance is a valuable tool.

overall profiles obtained from tag counting experiments could be performed using the traditional χ^2 test. However, this is the wrong approach for experiments where the significance of differences between expression levels (i.e., tag counts) of individual genes is to be determined, for example, in diseased and normal tissue states (19). One of the issues in performing tag-sampling experiments is that the experiments themselves are usually not replicated. Thus, the dispersion of results cannot be used to estimate the SEs associated with each expression measurement. This eliminates the possibility of using standard tests of variance. Instead the Poisson distribution, which includes an implicit estimate of standard error, approximates random sampling of tags very well. Audic and Claverie (20) have proposed a significance test (see Equation 8.1) in which the sample size plays no part, so long as it is the same for both experiments, but only depends on the observed tag counts of the same gene from diseased, g_A, and normal, g_B, states:

$$p(g_B|g_A) = \frac{(g_A + g_B)!}{g_A!g_B!2^{(g_A+g_B+1}} \qquad (8.1)$$

The equation has also been extended to cover the more practical case of different total numbers of tags. Thus, taking some data from Fannon (21) as an example, we can calculate values for the probability of certain genes expressed at different levels in normal prostate, stage B2 cancer, stage C cancer, and tissue from a benign prostatic hyperplasia (BPH) sample as shown in Table 8.2.

The relationship between gene expression, mRNA level, and protein expression is complex and not one that can be gleaned from collecting copy number information in this type of experiment. Even with careful statistical analysis, such as that described above, the assumption that increases or decreases in copy number reflect real biologically significant events relies on the confidence with which we can compare a library made from one set of cells to a library made from a different set of cells. Thus, most transcript analysis experiments setting out to be quantitative end up simply as target identification exercises. A major goal of proteomics is to generate a factory-type approach to profiling protein level expression that more closely reflects the biological reality. The EST approach has been turned into an industrial scale process but has not been able to impact the drug discovery process significantly because of the biological limitations described and the lack of sound mathematical modeling of the whole process.

Expression experiments are measures of cell population averages, not the contents of individual cells, so it is important to consider to what extent all cells in the candidate popu-

Table 8.3 Brief Descriptions of Three Technologies for Genomic Scale Transcript Profiling

Expression Profiling Technology	Brief Description	Form of Data Generated
cDNA array chip	Tens of thousands of cDNA clones of genes are placed onto a glass slide in a grid formation. Hybridisation of molecular probes (RNA extracts) to the clones is detected using a fluorescence system. By using two sets of probes, labelled with differently coloured fluorescent dyes, it is possible to assess expression differences.	Fluorescence intensities and colours for each spot on the chip. The nature of the clones on the chip is known.
High-density oligonucleotide arrays	Arrays of oligonucleotides are synthesised directly onto the glass chip using special chemistries and light sensitive masking. This generates arrays of known sequences of fixed length. Probes are hybridised to the arrays and computational analysis is necessary to interpret the resulting patterns.	An image of the entire chip is processed using specialised chip scanning software.
Serial analysis of gene expression	A sequence-based approach to the identification of differentially expressed genes through comparative analysis. Allows simultaneous analysis of sequences that derive from different cell populations or tissues. This is not a chip-based method. Identification of sequences relies on completeness of public sequence databases and, therefore, can only be used to analyse known genes.	Sequence data for SAGE tags allows profiling of gene expression.

lation are in the same state (22). Whereas work in single-celled organisms may be more straightforward to control, work in multi-cellular organisms has the added complexity that expression measurements may involve contributions from cells derived from a variety of tissues. Furthermore, when taking into consideration mRNA copy number, it should be understood that absolute transcript abundance measurements do not completely measure mRNA concentration.

Although there was initially some concern that the use of ESTs was a shortcut to discovery of genes for the purposes of patenting and ring-fencing areas of research for profit, in fact, the substantial numbers of quality ESTs in the public domain have helped in the pinpointing of genes in genomic data and have contributed to the speed with which the human genome sequence was completed.

4.2.5 Genome-Wide Expression Analysis. A major step towards understanding how organisms work is the determination of the complete sequence of all genes in the genome. This remarkable goal has been achieved for a number of organisms, including Homo sapiens, the flowering plant Arabidopsis thaliana, the single celled yeast Saccharomyces cerevisiae, and a large number of bacteria. The analysis of the sequence data then becomes the issue. It is no trivial task even to locate the positions of all the genes in the human genome. Genes for which there are no homologs in the current sequence databases will take some time to elucidate. See Ref. 23 for a detailed analysis of this topic and then Refs. 24 and 25 for detailed studies on the human genomic sequence.

The three basic technologies for generation of genome-wide expression information are cDNA microarrays, high-density oligonucleotide arrays ("GeneChips"), and serial analysis of gene expression (SAGE) (22). These technologies are outlined in Table 8.3.

In terms of quantities of data, a single microarray experiment looking at 40,000 genes from 10 different samples, under 20 different conditions, produces at least 8,000,000 pieces of data (26). Chip technologies, though originally expensive because of the costs of chip fabrication, are now being used to contribute data to public domain databases and are

widely used in industrial applications. A recent comparison of array databases available for local installation, public submission of data or public query, listed 13 different systems from sources worldwide (27). One such repository, ArrayExpress, is now being funded by the European Union at the EBI and is in the early stages of development (see http://www.ebi.ac.uk/arrayexpress). It is intended to be compliant with the microarray gene expression database (MGED) standard (see http://www.mged.org/).

The process of gene expression by microarray is shown below, based on (13).

1. Construct array
2. Prepare biological samples for investigation
3. Extract and label sample RNA
4. Hybridize samples to array
5. Image the array
6. Locate spots and evaluate fluorescent intensities
7. Construct gene expression matrix from spot intensities
8. Analyze gene expression matrix

As with any biological experiment, the result of this process should be the accumulation of knowledge concerning the biological processes under study. Interestingly, the first five steps are material handling processes, while the remaining steps only involve information processing.

5 DATABASES, TOOLS, AND APPLICATIONS

Because bioinformatics is about the management of information in the domain of biology, databases play a significant part in acting as repositories for a wide variety of different types of data. The main focus of this section is to give a flavor of the breadth of databases available and to highlight the role of the primary sequence databases and the secondary pattern (or family) databases in assisting with protein functional assignment.

5.1 Databases

Data repositories of DNA sequence, protein sequence, and higher-level resources of integrated information pertaining to the relationships between sequences (for example, pattern and profile databases) are the core tools for performing a wide variety of bioinformatics analyses. Many of these databases are in the public domain and are freely accessible through links available at a range of websites (including those listed in Table 8.1), although copyright is claimed in the annotation sections of some databases. The January 2002 issue of *Nucleic Acids Research* is a special annual database issue. It contains 112 articles describing in some detail different databases in use in the field. These are a subset of the 339 databases (up from 281 in the previous year) listed and briefly described in the Molecular Biology Database Collection, which constitutes an additional article in itself (28, 29). The complete list can also be found at http://www.nar.oupjournals.org. While the list was being prepared for the 2001 edition, 55 new databases were added to the previous total. In 2002, 58 additions were made. This rapid expansion in the number of databases available is indicative of the recognition by the community of the need for accessible, carefully designed databases to meet the needs of a wide diversity of research programs.

Much of the value of databases, assuming the provision of accurate sequence data, arises from the quality of the annotation that is available. This normally includes at least a brief description of the function of the sequence and essential references to the literature. Many databases include a lot more than this. In particular, SWISS-PROT (8) is viewed as the most reliable source for annotation information. SWISS-PROT emerged in the 1980s out of a need to have high quality, robust annotations for the protein sequences that made up its core content. However, the process of annotation is labor intensive and not one that is easily automated. Although the content of SWISS-PROT is well regarded, it lacks the completeness of the source DNA databases because of the necessary delay in incorporating newly annotated sequences. Indeed, a team of annotators is employed at the

EBI solely to perform this task. A computationally annotated supplement, TrEMBL (8), has been made available to make up for this deficiency. Nevertheless, computer annotation still has some way to go before it comes close to the level of competence of skilled human annotators. This is an area of active research (30).

Nevertheless, with the rapid generation of sequence data from genome scale experiments more effective means of characterizing protein sequences and annotation are now required. The database has responded by improving labeling of annotation in both SWISS-PROT and TrEMBL and by adding more advanced and rigorous tagging of evidence for functional statements that have been made (31).

Whereas most patent sequences are available in the public domain for use in research and for commercial exdoitation, there is a substantial body that are the subject of patent protection. It is often useful when conducting searches of sequence databases to be aware of the sequences that are patented because this may imply certain restrictions on the use to which these sequences can be put in a commercial context. The commercial repository is maintained by Derwent (Thomson Scientific), which generates the Geneseq database of patented sequences. This is a useful collection because it contains a broad historical collection as well as more recent examples, although the terms for a commercial license to use the database may be off-putting to some potential users. There are also patent sections of GenBank/EMBL DNA databases too. but these are of limited value because they contain only more recent sequence data.

5.2 Sequence Comparison

When dealing with the output of most experiments in target discovery the question "has this gene been seen before?" arises. The answer is, at first sight, straightforward: Compare the sequence obtained from the experimental output with all the known sequences and print the result.

Sequence comparison makes up a major part of the work of the bioinformatics analyst. It demands skill in operating the tools; for example, choosing the appropriate databases to search and selecting the appropriate search method, followed by insight and experience in assessing the meaning of the results of the search. A search query with a single previously known sequence is likely to return not only the match with itself but also a host of other matches at varying levels of similarity with the query sequence. This extra information can be very valuable in placing the query sequence in the context of many closely related sequences that make up the family of genes to which the query belongs. More distantly related sequence matches can potentially indicate genes with similar function, even if the match is relatively short and of low score.

The experienced analyst should be able to sort the significant matches from the uninteresting ones. Often, this type of experience is difficult, if not impossible with current technologies, to capture in a computer program. Rules that seem to work under some circumstances produce nonsensical results in others. As a result, many of the techniques used for current sequence comparison engines are heuristic rather than strictly algorithmic, that is, the rules that are implemented as part of the process for returning significant hits from the query database tend to produce the correct result but cannot be guaranteed to do so in all circumstances. For a fuller discussion of algorithms and heuristics, albeit outside the context of bioinformatics, see Ref. **32.**

One of the key aspects of sequence comparison is the understanding of similarity when applied to molecular sequences. There are essentially two ways of considering this: simple residue identity and residue substitution. In this discussion, we consider the comparison of two protein sequences, but the process is the same for comparison of DNA or RNA sequences. The alphabet used in the comparison is just different because it is 20 for protein sequences and 4 for DNA and RNA. By comparing residues at the same position in each sequence and counting up the number of identities we arrive at, a score that can be expressed as a percentage match for the pair of sequences. The alternative method compares each pair of residues and looks up a score for that pair in a substitution table or scoring matrix. The summed score across the whole se-

quence length can again be expressed as a percentage match. The two sequences under comparison are, however, likely to be sufficiently different that equivalent residue positions are not in register when the two sequences are laid out, one on top of the other. In this situation, the sequences must be aligned with each other so that equivalent residue positions are in register to make the score meaningful. This may involve insertion of gaps into one or both sequences. The skill here is to create an alignment between the two sequences that reflects some biological reality; it is from this biological reality that we derive the notion of equivalent residue positions. These positions can be deduced from manual manipulation of the alignment on the basis of mutation data or other functional information using a suitable sequence editor (33), or perhaps from understanding the spatial layout of residues if structural data is available. In each of these cases, the resulting sequence alignment will reflect the manner in which equivalent residue positions have been determined—both methods have their place. A variety of methods have been developed for comparing pairs of sequences, including the basic classical methods of Needleman and Wunsch (34) and Smith and Waterman (35).

Extending these pairwise comparison methods to database searching has been carried out, and a plethora of hybrid methods and improvements have been made. The manner in which significant alignments are reported varies from implementation to implementation. Database searching by alignment in this way is computationally intensive and specialized computer hardware is often used to gain speed increases. Because the comparison of pairs of sequence takes place in an exhaustive manner, these types of database searching methods are considered to be the most sensitive. More modern methods of database searching look for shorter matches spread over the lengths of the query and database sequences, and then extend these matches until the score for the match falls below a threshold level. Lists of sequence matches returned are then aligned using a pairwise alignment technique to provide a match and score over the whole length of the comparison sequences. For an example of this type of approach, see FASTA (36). Such methods are readily implemented on standard computer hardware and thus are accessible as Internet resources or as local implementations on UNIX or Linux servers.

The most popular tool currently in use is BLAST (Basic Local Alignment Search Tool) (37) from the NCBI. BLAST is an example of a heuristic that attempts to optimize a specific similarity measure. The most recent revisions to the algorithm are gapped BLAST and PSI-BLAST (38), with improved accuracy for PSI-BLAST using composition-based statistics (39).

5.3 Phylogenomics and Gene Family Databases

Determining protein function from genomic sequences is a central goal of bioinformatics (40), and to achieve this goal, comparing single sequences against databases of DNA or protein sequences is a necessary bioinformatics skill. However, many such searches have already been carried out, and the results are available to analyze at a higher level of abstraction in the protein and gene family databases (9, 10, 14, 41–43). It is the relationships between sequences that form the basis of any gene family database. Many of the current databases did not set out to become gene family databases. However, application of the underlying methodology for defining gene families (whether based on blocks of conserved sequence alignment or on profiles representing entire sequences, or simple regular expressions) has resulted in a number of resources that are particularly valuable in placing drug discovery targets in their biological context.

The processes of evolution by natural selection imply that species are related to each other in a tree-structured hierarchy; but more than this, the history of sequence relationships during evolution is also significant. Organisms are defined by their genes, and their behavior is modified through environmental experience. The relationships between genes within a single organism indicate that genes and their protein products also fall into well-defined families. Protein phylogenetic profiling (40) and phylogenomic analysis (44) are methods that are valuable where functional assignment by sequence similarity alone is

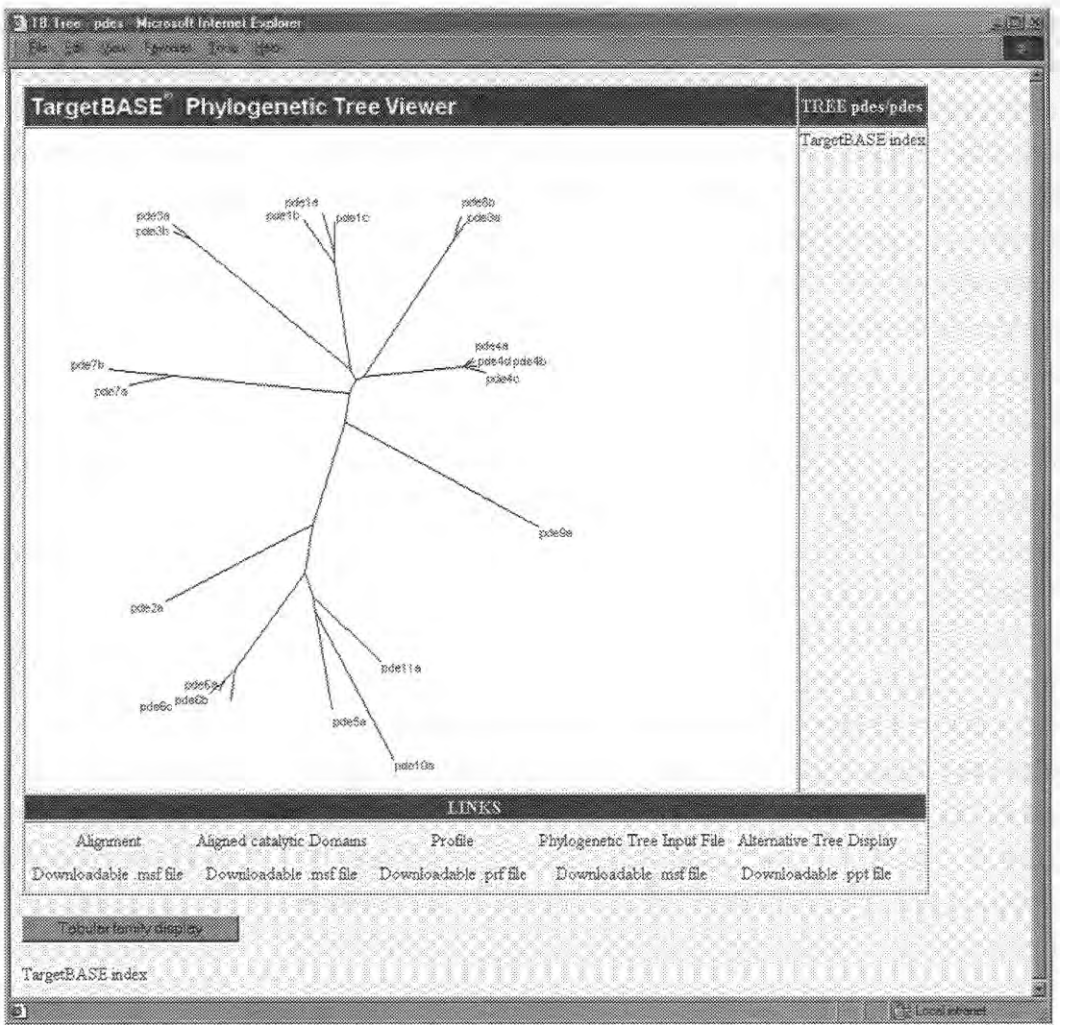


Figure 8.4. An example of a user interface to a phylogenomic-oriented database (48). Relative distances, following black line paths, between nodes on the tree of phosphodiesterases indicate the similarity level between members of the family, based on the regions of the sequences selected for the phylogenetic analysis. Links to aligned domains permit the alignments themselves to be explored. The order in which the genes appear in the tree (the branching order) gives an indication of the homology relationship between members of the family. See Section **5.3.**

problematic. This is because phylogenomic analysis is based on understanding the process by which sequences have diverged from common ancestors rather than focusing on the sequence similarity itself, which is an evolutionary endpoint. The approach is to determine the phylogenetic tree of a gene family and then to overlay any known functions of the genes on the tree. The functions of uncharacterized genes are predicted by their phylogenetic positions relative to those of the previously characterized genes. Importantly, depending on the manner of their construction, the trees may indicate similarity distances along connecting branches but it is the order of branching that reflects evolutionary relatedness (otherwise known as homology). For an interesting discussion of the correct use of the terms homology and similarity see Ref. 45.

This approach is illustrated in the phosphodiesterase gene family tree presented in Fig. 8.4. The set of relationships has been determined by comparing not just two sequences with each other, or a database of sequences to one sequence (as in a sequence database search), but by comparing a set of phosphodiesterase sequences to each other in the form of a multiple sequence alignment (see Fig. 8.5). Conserved regions of un-gapped sequence were chosen from this alignment to use as input to a phylogenetic analysis method (46, 47) and an evolutionary tree was eventually reconstructed. This tree represents aview of the relationships between genes in the phosphodiesterase gene family: more closely related genes are closer together in the diagram (i.e., they are connected by shorter paths); those further away are less closely related to each other. Figure 8.4 is based on an entry in TargetBASE (48), which adopts the phylogenomic paradigm. The relationships between members of gene families are used as the structure for an object-oriented database and associated user interface that provides a navigation tool for a curated gene index.

There are other approaches to family databases that rely more extensively on sequence similarity to define classes of genes or proteins. For example, PROSITE (49) is a resource that uses regular expressions to define patterns of residues that represent biologically significant sequence motifs. Recent versions have incorporated profiles, weight matrices that express the characteristics of a gene

```
                  251                                                    300
pde1a_human       KLHYRWTMAL MEEFFLQGDK EAELGLP.FS PLCDRKSTM. VAQSQIGFID
pde1b_human       LVHSRWTKAL MEEFFRQGDK EAELGLP.FS PLCDRTSTL. VAQSQIGFID
pde1c_human       DLHHRWTMSL LEEFFRQGDR EAELGLP.FS PLCDRKSTM. VAQSQVGFID
pde2a_human       KTTRKIAELI YKEFFSQGDL E.KAMGNRPM EMMDREKA.Y IPELQISFME
pde3a-human       ELHLQWTDGI VNEFYEQGDE EASLGLP.IS PFMDR.SAPQ LANLQESFIS
pde3b_human       DLHLKWTEGI VNEFYEQGDE EANLGLP.IS PFMDR.SSPQ LAKLQESFIT
pde4a_human       ELYRQWTDRI MAEFFQQGDR ERERGME.IS PMCDKHTAS. VEKSQVGFID
pde4b_human       ELYRQWTDRI MEEFFQQGDK ERERGME.IS PMCDKHTAS. VEKSQVGFID
pde4c_human       PLYRQWTDRI MAEFFQQGDR ERESGLD.IS PMCDKHTAS. VEKSQVGFID
pde4d_human       QLYRQWTDRI MEEFFRQGDR ERERGME.IS PMCDKHNAS. VEKSQVGFID
pde5a_human       PIQQRIAELV ATEFFDQGDR ERKELNIEPT DLMNREKKNK IPSMQVGFID
pde6a-human       EVQSQVALLV AAEFWEQGDL ERTVLQQNPI PMMDRNKADE LPKLQVGFID
pde6b_human       EVQSKVALLV AAEFWEQGDL ERTVLDQQPI PMMDRNKAAE LPKLQVGFID
pde6c_human       EVQSQVALMV ANEFWEQGDL ERTVLQQQPI PMMDRNKRDE LPKLQVGFID
pde7a_human       ELSKQWSEKV TEEFFHQGDI EKKYHLG.VS PLCDRHTES. IANIQIGFMT
pde7b_human       EMSKQWSERV CEEFYRQGEL EQKFELE.IS PLCNQQKDS. IPSIQIGFMS
pde8a_human       QYCIEWAARI SEEYFSQTDE EKQQGLPVVM PVFDRNTCS. IPKSQISFID
pde8b_human       DLCIEWAGRI SEEYFAQTDE EKRQGLPV'VM PVFDRNTCS. IPKSQISFID
pde9a_human       EVAEPWVDCL LEEYFMQSDR EKSEGLP.VA PFMDRDKVT. KATAQIGFIK
pde10a_human      PVTKLTANDI YAEFWAEGD. EMKKLGIQPI PMMDRDKKDE VPQGQLGFYN
pde11a_human      EISRQVAELV TSEFFEQGDR ERLELKLTPS AIFDRNRKDE LPRLQLEWID
```

Figure 8.5. Part of an alignment of catalytic domains of the human phosphodiesterase gene family. Positions in the alignment where gaps have been introduced into a sequence to bring it into alignment with other sequences are indicated by "." characters.

family using all the sequence information available. The principal value of this resource is that it presents patterns for recognition of gene families that are relatively simple to understand. The downside is that the use of such patterns can produce both true positive hits (members correctly predicted) and false positive hits (members incorrectly predicted). PROSITE lists true and false positives for searches performed in the production of a release of the database, but it is as well to be aware that when the patterns are used in isolation, there is often a false positive hit rate that must be taken into account by reconciling the results of a pattern search with the results of database annotations or other pattern recognition methods.

The PRINTS system **(9,** 10, 50, 51) is an approach based on an examination of core regions of un-gapped sequence conservation within a set of aligned sequences (multiple sequence alignment). The method rigorously builds up fingerprints for a gene family through use of an iterative database searching technique allied to intelligently applied sequence alignment. The fingerprints themselves can then be used to diagnose new gene family members in novel sequence data or can be used to identify modules of functional sequence across different gene families.

One of the issues in using different databases of gene family information is that definitions of which genes belong to which gene families can vary depending on the method used. Apweiler et al. have undertaken a useful effort at rationalizing and integrating family database annotation at the EBI in the InterPro resource (52). The databases that make up the membership of the InterPro consortium are PROSITE (49), PRINTS (9), Pfam (53), ProDom (54), and SMART (55). InterProScan is a tool that enables scanning of individual protein sequences against the InterPro member databases **(56).**

6 THE BIOINFORMATICS KNOWLEDGE MODEL

Up to this point, we have discussed sources of data and means of manipulating and comparing data elements (in terms of sequences, alignments, gene families, etc.), but the end point of all this analytical process must be the acquisition of knowledge. It is through increased understanding that sound decisions

can be made in applying the results of bioinformatics analyses to application areas, such as drug discovery. So, in this section we consider the relationships between data, information and knowledge, which are frequently regarded as poor relations to laboratory-based experimental data acquisition. However, as drug discovery organizations, including large pharmaceutical and smaller biotechnical companies, develop a significant history of assays, screens, and leads, it is vital to have strong internal support for managing data flows, integrating related data into information systems, and transforming knowledge thus gleaned into tangible benefits.

6.1 Data, Information, and Knowledge

According to the University of California at Berkeley (57), it has taken 300,000 years for humankind to accumulate 12 exabytes of data.[2] It will take just **2.5** more years to create the next 12 exabytes. **(An** exabyte is 1,000,000,000,000,000,000 bytes or a billion gigabytes.) This is a truly unimaginable amount of data, equivalent to the data stored on a pile of floppy disks 24 miles high. It is the rate of accumulation of data that is the key point of interest, however, and the fact that it is accelerating.

It is crucial to distinguish between the terms data, information, and knowledge so that we can think clearly about the goal of data accumulation in our own industry sector. There are two views: a tiered hierarchical view and a more formally correct scientific view.

6.2 The Hierarchical View

In the hierarchical view, data is the bottom rung of a ladder leading to the accumulation of information that leads, ultimately, to an increase in knowledge. Apply this hierarchical principal to an everyday example of taking this article to the photocopier: the data represented by the article is the sequence of strokes and dots on the page that make up the page image. The page image is the information represented by the article. The knowledge element only comes later when the observer actually reads and understands the article. Compare this with the act of photocopying a research article, a process that does not in itself add to understanding on the part of either the photocopier or of the researcher. The acquisition of knowledge implies an active relationship between author and recipient of the information. In this, intuitive sense, we know that the hierarchical view works to some extent as a model of the way in which some knowledge is acquired.

6.3 The Scientific View

The second view is the scientific one (58). Here, we start with the piece of information that we are trying to understand, perhaps a gene whose function we plan to determine. Experiments are designed and performed to determine the characteristics of the function of the gene; such experiments yield data that describe aspects of the information. Knowledge comes from understanding and interpreting the results of the experiments. Again, knowledge is accumulated as part of an active relationship between the data describing the information and the investigator reviewing the data and drawing conclusions about the state of the information. Gene function is itself a complicated concept because the functions of gene products can rarely be assessed in isolation, owing to the network of interactions in which most genes are involved. A collection of sequence data, collected at the DNA or protein level, describe the molecular structure of a gene or its product at a primary level—it is not, however, a complete description. There are other biochemical factors to be considered; for example, proteins that assist in the folding process to create an active three-dimensional molecule, post-translational modifications, glycosylation, interactions with other molecules to generate a higher-level function, etc.

6.4 Data is Not Knowledge

Simply increasing the amount of data in the genomic universe does not necessarily increase the speed of knowledge acquisition. In short, data is not knowledge. Knowledge itself requires understanding and demands the active participation of the one acquiring the knowledge.

[2] In fact, the referred study uses the word "information." However, within the usage of this article "data" is a more appropriate term.

Most pharmaceutical organizations have in place the means for collating data from a wide variety of sources. Genomic information is available freely in the public domain as well as in proprietary databases. Some successful companies have used the multiple subscription database model to generate revenue to create more data to return to their customers. In this model, data tends to be available on a non-exclusive basis but it is up to the licensee to determine how best to interpret the data in its own research environment. Some information linking is available in such models, more especially those that use Internet portals as a user interface and results delivery mechanism. This has been a valuable means of acquiring data in gene and protein expression. The model is, however, showing signs of age. Pharmaceutical and biotechnology companies have needed to make substantial investments in technology and specialist skills (particularly bioinformatics) just to warehouse the data and make the analytical results available to drug discovery program scientists. Yet, all this effort still remains the heartland of genomics: target discovery groups have sprung up in the pharmaceutical and biotechnology companies to create a process by means of which targets can be gleaned from the genomic morass.

6.5 Drug Targets

When considering drug targets, as opposed to simply gene products, there are a host of characteristics that must be taken into consideration. For all the drugs that are currently available on the commercial market, there are only about 500 drug targets on which they individually act. We now understand the human genome to contain about 30,000 genes, which give rise to a still incalculable number of protein products (59). How many of these products represent tractable drug targets? The gold standard for assessing whether a protein is or is not a target is target validation. In this process, additional biochemical or molecular genetic data are required to determine whether a protein is truly involved in a disease state and thus whether it could be considered to be a suitable target molecule for drug discovery. The prospect of performing such target validation experiments, always assuming the results could be interpreted unambiguously, is a daunting one. Structural determination, by X-ray crystallography or nuclear magnetic resonance, has deepened our understanding of some biological processes immeasurably—particularly in the realm of certain proteases, DNA binding proteins, and some other soluble enzymes. The majority of drug targets are, however, membrane bound (for example, the plethora of ion channels and G-protein–coupled receptors), making structural determination to any degree of critical confidence impossible. Molecular modeling can assist in this process and has been a valuable tool for many years in thinking through possibilities and providing a framework for interpreting other biochemical results. The further away we move from rigorously determined experimental data, however, the less likely is a pharmaceutical company to embark on the commitment of expense to exploratory studies in drug discovery. Finally, in considering the suitability of a gene as a drug discovery target, we must take into consideration the temporal nature of gene expression, an area of research that has not yet been adequately addressed in the analysis of genomic data.

The alternative approach for dealing with the wealth of genomic data, in the context of drug discovery, is to consider the genomic data as a background landscape against which to pick out the currently known and validated drug targets. These targets fall into families whose relationships can be rigorously determined at the sequence level. The related members of these families can then be assessed by analogy to determine their appropriateness as drug targets. These phylogenetic (gene family) relationships then form the basis of the structure for a database and become both a tool for navigating and exploring the relationships themselves as well as a mechanism for integration of other types of drug discovery data—for example, high-throughput screening results and structure activity relationships of bound ligands. Thus, we can see that an integration of information across the realms of genomics, target discovery, screening and lead optimization becomes possible and an achievable goal.

Target analysis belongs in the domain of knowledge for drug discovery. Such knowl-

edge is a representation of our understanding of the function of a gene product in a disease state. This functional understanding is derived from the analysis of experimental data linked to our experience of the functional information, itself defined (if only partially) by the data that characterize that functional information. Setting such knowledge within the context of the genomic universe enhances our ability to select new targets for future validation studies, either through molecular genetic techniques (gene knock-out, anti-sense, etc.) or through mechanistic validation using small molecule tools discovered as part of the assay development and screening process.

There are attempts at capturing this type of knowledge within databases (known as knowledge-bases). Much valuable research is going on in the allied areas of knowledge representation. At present, active human participation is required in accumulating knowledge and deriving ultimate benefit from it in the form of new drugs that fulfill unmet human needs in therapeutic situations.

7 STRUCTURAL GENOMICS

Structural genomics is the process of determining the three-dimensional structures of an organism's proteome (60). Predictions of protein function can be attempted from knowledge of the structure alone, or the additional information gained can be used to inform sequence-based methods of functional prediction (61).

The traditional paradigm of classical structural biology has been to select a protein based on its known biological function, ascertain its molecular structure, and use the data thus gleaned to understand how its biological function is carried out at the molecular level. To this end, more than 12,000 structures have been determined, to varying degrees of resolution and confidence. Much has been learned, as a result, of the complexity of protein structure and the manner of interaction of proteins and their native ligands, or of proteins and small molecule drugs.

The essence of structural genomics is to start from a gene sequence, produce the functional protein, and then determine its three-dimensional structure. The biological function of the protein *in vivo* is then deduced from an understanding of the structure. In this paradigm, there is no limitation on the number of structures that can be determined except the ability to purify sufficient protein for crystallization trials. The usual caveats apply regarding the solution of structures of membrane-bound proteins, for example, G-protein–coupled receptors, ion channels, certain classes of kinases, etc.

Often, the most useful functional information is derived from the structures of protein-ligand complexes because they reveal the nature of the bound ligand and its location in the protein. In the case of enzymes, a catalytic mechanism can often be postulated taking into account the disposition of residues in the active site pocket. While such structures have traditionally been determined by design, the ligand is unknown in the structural genomics approach. Only in rare cases will a ligand be co-crystallized by serendipity from the cloning organism.

From the perspective of bioinformatics, it is important to appreciate that structural determination can only provide data that reflect the biochemical or biophysical properties of the protein. The biological role in the cell or organism is a complex of interactions including spatial and temporal dimensions. Sometimes information can be derived using other techniques—for example, cDNA expression analysis, two-dimensional gel electrophoresis, biochemical assay, etc. Bioinformatics techniques building on other types of data can also assist in providing biological context for the function of a protein—for example, phylogenetic, fingerprint or regular expression analyses, etc. All of these techniques yield data that, when reviewed as a whole, can direct the course of further experiments or influence experimental design.

We have seen that, purely using techniques of sequence comparison, the function of about 40% of genes sequenced from genome projects can be inferred from sequence identity or similarity measures or by motif comparison using a variety of techniques. It is known that proteins exhibiting insignificant sequence similarity often adopt similar tertiary structures, which themselves have similar (or at least re-

lated) molecular functions. In fact the variety of types of fold taken up by polypeptides is thought to be quite limited [SCOP (62) and CATH (63)]. Discovering a useful relationship between folding topology and sequence, which can be used to predict folding accurately is, however, not trivial. By comparing the structures newly determined from structural genomics initiatives with structures already deposited in the Protein Data Bank, it may be possible to extend the inference of molecular function further than that achieved from sequence comparisons alone. Once the molecular function has been characterized in this computational way, we may begin to postulate the cellular function of the protein under analysis.

7.1 Predicting Protein Function from Structure

Some consider the "Holy Grail" of computational biology to be the accurate prediction of a protein's function solely from knowledge of its primary sequence. We have already briefly mentioned the role of structural information in guiding and illuminating the process of molecular sequence alignment (Section 5.2.1). Structural data can be a truly effective means of understanding spatial relationships between amino acid side-chains, backbone donor and acceptor groups, and the means of interaction of natural and man-made drugs. For a thorough and insightful overview of these matters see Ref. 64—the entire volume is essential reading.

Many methods have been proposed for capitalizing on our understanding of protein structure by creating algorithms that attempt to predict function from structure, or place proteins in structural categories that may have implications for functional analysis.

7.2 Neural Networks and Protease Function

Stawiski et al. (60) recently performed a study of the unique structural features of proteases in which the authors noted consistent structural similarities among unrelated protease family members. They found that proteases tend to be more tightly packed than other proteins and they tend to have fewer α helical regions and more residues in loop structures. A neural network was trained to predict protease function with 86% accuracy in a test set. Neural networks are an example of a technique used in bioinformatics for generating a predictive program from a set of weights that can be applied in a learning tool. The tool is trained by using parameters that show discrimination between, in this case, proteases and non-proteases. In this example, 36 proteases were tested. Each protease in turn was used as a test example, the network being trained using the remaining 35 proteases. In 31 of 36 cases (86%), the network was able to identify the remaining protease. By performing the same test on 258 counter-examples, 87% were correctly classified as non-proteases.

7.3 Fold Compatibility Methods

The ability to recognize the way in which a protein sequence is folded in three dimensions should enable us to model the interactions of specific side-chains in a manner that is simply not possible when considering proteins entirely at the sequence level. This notion has resulted in sequence threading algorithms that assess the level of compatibility of a sequence with a database of fold patterns (65, 66). The principal downside to this approach is that novel structural types cannot be predicted, because at least one example of each fold type must be present in the fold pattern database. Structural genomics may be the means whereby fold pattern databases can be populated with sufficient data to make them useful as predictive tools.

7.4 CASP and the State-of-the-Art

Currently, methods of structure prediction from sequence perform poorly. The results of the biannual CASP experiment (Critical Assessment of techniques for protein Structure Prediction see http://predictioncenter.llnl.gov/) are equivocal to say the least. A recent report on the improvements of aligning target sequences to a structural template (67) indicates that over the last four CASP competitions there was no significant improvement in quality in this key step in the prediction process. Alignment remains the major source of error in all models based on less than 30% se-

quence identity. The subjective impression is that structure prediction is getting better year after year. This analysis, however, seems to suggest there is some way to go before reliable models can be generated for fold types not yet available in the structural databases.

8 THE FUTURE

Bioinformatics is a wide-ranging science that has developed over the last 50 years, since the discovery of the structure of DNA, a period that has resulted in the sequencing of the entire genomic material of major species. Techniques of sequence comparison, database management, design, and curation have resulted in a healthy base on which to build more automated systems. It is this author's view that experienced sequence analysts will always have a place in this process, guiding the design of new algorithms and better knowledge bases. Only in this way will the true synergy between analyst and developer be realized and contribute to the understanding of the fruits of genomic research.

Bioinforrnatics, allied to drug discovery, is used for discovering new potential drug targets through the use of standard bioinformatics techniques in assigning function to novel gene products at the sequence Ievel and by the informed use of structural, mutational and biochemical data—reflected in sequence level alignment models. Assessment of expression levels of genes and the statistical relevance of differences in levels of expression at the mRNA level has contributed to drug discovery programs in pharmaceutical and biotechnology companies globally. In many respects, it is still early days for seeing the fruits of this work in the products offered on the market by these companies. There should, however, be many clinical candidates on trial in which bioinformatics has contributed, albeit at a level of detail that is frequently far below the level of interest of industry publicists. The fact is that bioinformatics is now engrained in the discovery process for new drugs. The next stage in its development will be integration between chemoinformatics (chemical informatics) and bioinformatics, driven by the need to understand the ways drug interact with their targets rather than merely exploiting biological assay systems as tools for drug discovery.

9 ACKNOWLEDGMENTS

I thank Jeremy Packer of the BioFocus Plc Bioinformatics Group for casting a critical eye over the manuscript before it went for review and Sue Scott for her proofreading skills.

REFERENCES

1. *Millennium and* Bayer *Announce* Genome-derived *Oncology Drug Candidate Selected For First Human Studies, available online at* http://www.mlnm.com/news/2001/01-10--0.html, *accessed on September* 12, 2002.
2. J. *Owens, A. Hinde, B. Ramster, and R. N. Lawrence, Drug Discov. Today,* 6, 229-230 (2001).
3. *B. A. Kenny, M. Bushfield, D. J. Parry-Smith, S. Fogarty, and J. M. Treherne, Prog. Drug Res.,* 51, 245–269 (1998).
4. *T. K. Attwood and D. J. Parry-Smith, Introduction to Bioinformatics, Addison Wesley* Longman, *Harlow, UK,* 1999.
5. *P. Baldi and S.* Brunak, *Bioinformatics: The Machine Learning Approach, 2nd ed. MIT Press, London,* 2000.
6. *M. S. Waterman, Introduction to Computational Biology. Chapman & Hall, London,* 1995.
7. *T. Etzold and P.* Argos, *Comput. Appl. Biosci., 9,* 49-57 (1993).
8. *A. Bairoch and R. Apweiler, Nucleic Acids Res.,* 28, 45–48 (2000).
9. T. K. *Attwood, M. D. Croning, D. R. Flower, A. P. Lewis, J. E.* Mabey, *P. Scordis, J. N. Selley, and W. Wright, Nucleic Acids Res.,* 28, 225–227 (2000).
10. T. K. *Attwood, D. R. Flower, A. P. Lewis, J. E.* Mabey, *S. R. Morgan, P. Scordis, J. N. Selley, and W. Wright, Nucleic Acids Res.,* 27, 220–225 (1999).
11. *T.* Hubbard, *D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V.* Curwen, *T. Down, R. Durbin, E.* Eyras, *J. Gilbert, M.* Hammond, *L. Huminiecki, A.* Kasprzyk, *H.* Lehvaslaiho, *P. Lijnzaad, C. Melsopp, E.* Mongin, *R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A.* Ureta-Vidal, *I. Vastrik, and M. Clamp, Nucleic Acids Res.,* 30, 38–41 (2002).

. M. Gerstein, *Nat.* Struct. Biol., 7 Suppl, 960–963 (2000).
. A. Brazma, Bioinformatics, 17, 113–114 (2001).
14. J. Packer, E. Conley, N. Castle, D. Wray, C. January, and L. Patmore, Trends *Pharmacol.* Sci., 21,327–329 (2000).
15. B. Destenaves and F. Thomas, Curr. Opin. Chem. Biol., 4,440–444 (2000).
16. R. Staden, K. F. Beal, and J. K. Bonfield, Methods Mol. Biol., 132, 115–130 (2000).
17. R. Staden, D. P. Judge, and J. K. Bonfield, Methods Biochem. Anal., 43,303–322 (2001).
18. D. S. Bailey, A. Bondar, and L. M. Furness, Curr. Opin. Biotechnol., 9, 595–601 (1998).
19. J. M. Claverie, Hum. Mol. Genet., 8, 1821–1832 (1999).
20. S. Audic and J. M. Claverie, Genome Res., 7, 986–995 (1997).
21. M. R. Fannon, Trends Biotechnol., 14,294–298 (1996).
22. M. Gerstein and R. Jansen, Curr. Opin. Struct. Biol., 10, 574–584 (2000).
23. F. Sterky and J. Lundeberg, J. Biotechnol., 76, 1–31 (2000).
24. Science, 291,1145–1434 (2001).
25. Nature, 409,745–964 (2001).
26. A. Brazma, A. Robinson, G. Cameron, and M. Ashburner, Nature, 403, 699–700 (2000).
27. M. Gardiner-Garden and T. G. Littlejohn, Brief Bioinform., 2, 143–158 (2001).
28. A. D. Baxevanis, Nucleic Acids Res., 29, 1–10 (2001).
29. A. D. Baxevanis, Nucleic Acids Res., 30, 1–12 (2002).
30. A. G. Rust, E. Mongin, and E. Birney, Drug *Discov.* Today, 7, S70–S76 (2002).
31. R. Apweiler, Brief Bioinform., 2, 9–18 (2001).
32. W. D. Hillis, The Pattern on the Stone, Weidenfeld & Nicolson, London, 1998, pp. 77–90.
33. D. J. Parry-Smith, A. W. Payne, A. D. Michie, and T. K. Attwood, Gene, 221, GC57–GC63 (1998).
34. S. B. Needleman and C. D. Wunsch, *J.* Mol. Biol., 48,443–453 (1970).
35. T. F. Smith and M. S. Waterman, J. Mol. Biol., 147,195–197 (1981).
36. W. R. Pearson, Methods Mol. Biol., 132, 185–219 (2000).
37. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, J. Mol. Biol., 215, 403–410 (1990).
38. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Nucleic Acids Res., 25, 3389–3402 (1997).
39. A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul, Nucleic Acids Res., 29, 2994–3005 (2001).
40. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, Proc. *Natl.* Acad. Sci. USA, 96,4285–4288 (1999).
41. E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, and R. Durbin, Nucleic Acids Res., 26,320–322 (1998).
42. E. L. Sonnhammer, S. R. Eddy, and R. Durbin, Proteins, 28, 405–420 (1997).
43. J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff, Nucleic Acids Res., 28, 228–230 (2000).
44. J. A. Eisen, Genome Res., 8, 163–167 (1998).
45. G. Theissen, Nature, 415, 741 (2002).
46. J. Felsenstein, *Annu. Rev.* Genet., 22, 521–565 (1988).
47. J. Felsenstein, Methods Enzymol., 266, 418–427 (1996).
48. J. Packer and D. J. Parry-Smith, Curr. Drug Discov., March, 29–33 (2002).
49. L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A. Bairoch, Nucleic Acids Res., 30, 235–238 (2002).
50. W. Wright, P. Scordis, and T. K. Attwood, Bioinformatics, 15, 523–524 (1999).
51. T. K. Attwood, M. J. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Maudling, L. McGregor, A. L. Mitchell, G. Moulton, K. Paine, and P. Scordis, Nucleic Acids Res., 30,239–241 (2002).
52. R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov, Bioinformatics, 16, 1145–1150 (2000).
53. A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer, Nucleic Acids Res., 30,276–280 (2002).
54. F. Corpet, F. Servant, J. Gouzy, and D. Kahn, Nucleic Acids Res., 28, 267–269 (2000).
55. J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork, Nucleic Acids Res., 28, 231–234 (2000).

56. E. Zdobnov and R. Apweiler, Bioinformatics, **17**, 847–848 (2001).
57. P. Lyman and H. R. Varian, How much information?, available online at http://www.sims.berkeley.edu/research/projects/how-much-info, accessed on September 12, 2002.
58. D. E. Knuth, Selected Papers on Computer Science, Cambridge University Press, Cambridge, UK, 1996.
59. J. M. Claverie, Science, 291, 1255–1257 (2001).
60. E. W. Stawiski, A. E. Baucom, S. C. Lohr, and L. M. Gregoret, Proc. Natl. *Acad.* Sci. USA, 97, 3954–3958 (2000).
61. M. Weir, M. Swindells, and J. Overington, Trends Biotechnol., 19, 61–66 (2001).
62. C. L. Lo, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia, Nucleic Acids Res., 28, 257–259 (2000).
63. C. A. Orengo, F. M. Pearl, J. E. Bray, A. E. Todd, A. C. Martin, C. L. Lo, J. M. Thornton, Nucleic Acids Res., 27,275–279 (1999).
64. M. Perutz, Protein Structure: New Approaches to Disease and Therapy, Freeman, New York, 1992, pp. 119–137.
65. R. T. Miller, D. T. Jones, and J. M. Thornton, *FASEB* J.,10,171–178 (1996).
66. D. T. Jones, M. Tress, K. Bryson, and C. Hadley, Proteins, 37, 104–111 (1999).
67. C. Venclovas, A. Zemla, K. Fidelis, and J. Moult, Proteins, 45 (Suppl 5), 163–170 (2001).