Using bioinformatics in gene and drug discovery

David B. Searls

Bioinformatics has, out of necessity, become a key aspect of drug discovery in the genomic revolution, contributing to both target discovery and target validation. The author describes the role that bioinformatics has played and will continue to play in response to the waves of genome-wide data sources that have become available to the industry, including expressed sequence tags, microbial genome sequences, model organism sequences, polymorphisms, gene expression data and proteomics. However, these knowledge sources must be intelligently integrated.

The pharmaceutical industry has embraced genomics as a source of drug targets and, as a corollary, has recognized that bioinformatics is crucial to exploiting data produced on a genome-wide scale. Bioinformatics is essentially a cross-disciplinary activity, and includes aspects of computer science, software engineering, mathematics and molecular biology[1–4]. It is far more than data management and has attained the status of a new scientific discipline, much as computational physics has become integral to progress in that field[5]. The lesson of the early stages of the genomic era is that bioinformatics is easily underestimated in both its cruciality and its resource requirements.

Bioinformatics

In target validation

In part, this is a consequence of the new paradigm of drug discovery, which is characterized more than anything else by huge leaps in scale. On the small-molecule side, combinatorial chemistry has vastly increased the diversity available, while laboratory automation has enabled high-throughput screening at unprecedented rates. Genomics has provided a corresponding increase in large-molecule targets. In fact, the bottlenecks in drug discovery have shifted remarkably to the point where the number of attractive targets available is no longer rate limiting – it has actually created the new problem of how to select the targets most likely to succeed from an embarrassment of riches. This has shifted the focus of bioinformatics from target identification to target validation.

The challenge to bioinformatics is evolving from that of creating long lists of genes to that of creating short lists of the targets most likely to be crucial in disease and least likely to fail for 'developability' reasons. There is a fundamental tension in the earliest stages of target selection between the desire to study targets that already have a well-understood role in disease and the desire to study those that might offer a completely novel mode of action (with the competitive advantage that this implies). Ideally, bioinformatics should provide the bridge that reconciles these goals, primarily by providing as many clues as possible to function and role.

In target selection

In addition to better filtration of targets in early discovery, bioinformatics can also help with three aspects of target selection:

- The characterization of targets, such as the classification and subclassification of protein families
- The understanding of targets, such as their behavior in a larger biochemical and/or cellular context
- The development of targets, such as making predictions about uptake or reuptake, detoxification, the stratification of patient populations and other gene-based variations.

David B. Searls, SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, PO Box 1539, King of Prussia, PA 19406, USA. tel: +1 610 270 4551, fax: +1 610 270 5580, e-mail: David_B_Searls@sbphrd.com

 1359-6446/00/$ – see front matter PII: S1359-6446(99)01457-9

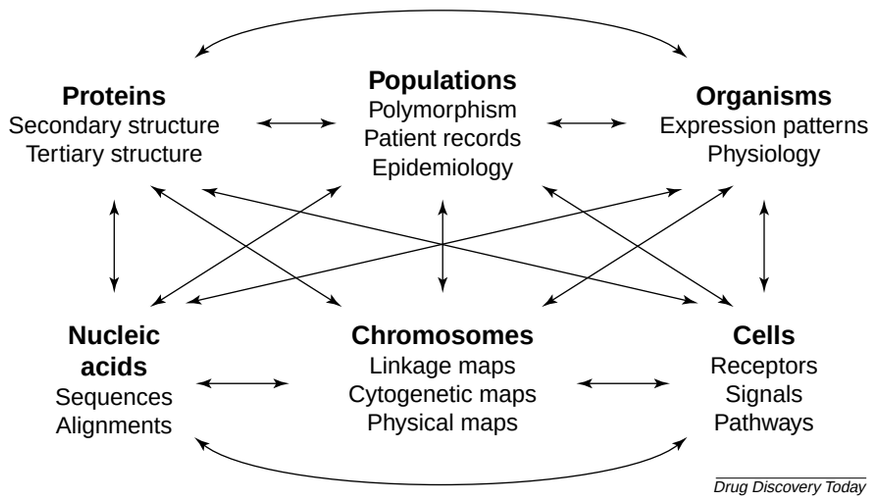

Figure 1. *The ultimate challenge to the bioinformatics community is the intelligent integration of data from many interrelated sources, which will be necessary to take greatest advantage of the knowledge available in the data. This figure attempts to capture the complexity of this challenge by displaying the various domains of bioinformatics that occur at many scales and embody many forms of data, suggesting that every domain relates to every other. Thus, for example, it could be used to examine how the expression pattern of a certain receptor varies with regard to a coding-region polymorphism that might affect the binding site of the protein and maps near a certain disease locus. How can this knowledge best be extracted from the myriad data available?*

To the extent that this can be carried out far in advance of key investment decisions, indeed before any significant bench effort is expended, the reward of bioinformatics will probably be enormous. In essence, this is the role that bioinformatics must play in target validation.

Bioinformatics can also attempt to become integrated more intimately into the discovery process itself by establishing 'wet–dry cycles'. Such cycles occur whenever a computational model can be linked to a biological one, such that predictions from the former can be immediately tested at the bench, with the results being fed back for producing refinements of the model. Instead of simply handing candidate genes to an independent bench validation process, it is important for the bioinformatics function to follow targets through the pipeline by, for instance, modelling biological systems, suggesting experiments, and using the results from the bench to refine the models further. In this way, bioinformatics can add more value by shortening cycle times. Value can also be added to targets already in development by continuing the search for homologs, both orthologs (homologs in different species, presumed to have similar function) and paraloges (homologs in the same species, which might have diverged in function), that could provide additional models in other species and follow-on targets that can use existing assays and compound libraries.

In data integration

Where bioinformatics offers the greatest hope for the future, however, is in data integration. The first wave of genome-wide data was that arising from expressed sequence tags (ESTs); this required an infrastructure for high-throughput data management and basic analysis, as well as novel algorithms for sequence assembly, expression analysis and so on (Fig. 1). This wave was followed by whole microbial sequences, for which some reuse of the infrastructure was possible but which also enabled some unique forms of analysis by taking advantage of the availability of complete genomes and a wide phylogenetic distribution.

Even so, there was no pressing need to integrate EST data and microbial genomes. Now, however, the waves of genomic data are coming in quick succession including, in the immediate future, the human genomic sequence, a profusion of single-nucleotide polymorphisms (SNPs) and expression data from microarrays and related technology. Not only must these waves be handled from the perspective of high-volume data management and application-specific analysis algorithms but there is also a new element: the full value of all these technologies and data can only be realized by relating them to each other – that is, by integration. For example, the high quality of the genomic sequence will add tremendous value to the error-prone EST data; at the same time, the ESTs are the single most powerful agent for identifying genes in the genomic sequence. Similar correspondences are apparent in data on expression and polymorphism.

Reviews of bioinformatics are most often technology centred, focusing on the techniques that have evolved rapidly in this new discipline for ever-more-sophisticated analyses of sequences, structures and phylogenies[6]. As an alternative, this article will examine the field from a data-centred point of view, which also serves to recapitulate the major phases of the role of bioinformatics in industry.

Expressed sequence tags

The beginning of the genome era in the pharmaceutical industry can be traced back to the arrival of ESTs as a source of truly large-scale gene-sequence data in 1993 (Refs 7,8). Before this, the available human gene-sequence data had largely been the result of the academic investigation of individual genes – just a few extended genomic regions were available, awaiting the output of the nascent human genome initiative. EST technology promised a shortcut to most, if not all, genes, prompting a 'gold rush' by the pharmaceutical industry and giving rise to the novel business models of biotechnology companies such as Human Genome Sciences (Rockville, MD, USA) and Incyte (Palo Alto, CA, USA), which (in effect) marketed the human genome *en masse*. Largely in response to this, a public EST sequencing effort was undertaken with funding from Merck[9] (Whitehouse Station, NJ, USA).

The immediate challenge that faced the industry in making use of these data was that of data management. Because of the time-critical value of the information, it was necessary to build a computational infrastructure that could rapidly identify interesting homologues, generally by conventional BLAST (Basic Local Alignment Search Tool) database searches[10], and then present the results to biologists for evaluation. This was quickly followed by a recognition that, in part because of the sheer volume of ESTs and in part because they represented relatively short and error-prone sub-sequences, it would be necessary to process them further by clustering or (even better) by assembling overlapping fragments to create a dataset that more closely represented the actual underlying genes. This need gave rise to the public Unigene resource[11] at the US National Centre for Biotechnology Information, which provides clusters of ESTs likely to have arisen from the same genes, as well as to several proprietary efforts with significantly greater algorithmic and software engineering sophistication. In many companies, this corresponded to the transfer of responsibility for the data from generic corporate IT departments to new and growing bioinformatics groups with specialized skills and biological backgrounds.

Finally, in what might be called a scientific 'aftermarket' to the development of EST resources, a series of even more sophisticated bioinformatic analyses were layered onto the basic data and its clusters or assemblies. The most notable such value-added analyses arose from the availability of ESTs derived from a wide variety of cDNA libraries representing different tissues and states of disease and development. This afforded the opportunity to view ESTs as samplings of populations of transcripts and, by simply counting the representation of different sources in any given cluster or assembly, to infer relative levels of expression by tissue or stage[12,13]. Other algorithmic extrapolations from the EST data have included efforts at error correction, the analysis of alternative splicing and the detection of putative SNPs (Refs 14,15). Although much of this phase of computational development occurred in academia, a few companies established bioinformatics research groups at this time to develop such approaches to extracting maximum value from the data.

Microbial genomes

Beginning with the first publications of complete microbial genome sequences[16], pharmaceutical and biotechnology companies with an interest in antibiotics began maneuvering for access to genomic sequences from relevant pathogens. These data had a much lower error rate than EST sequences but did produce a fresh set of challenges to the bioinformatics community. Once again, the first phase was primarily concerned with data management; however, during data production, the shotgun assembly of entire genomes in the range of megabases required refined algorithmic approaches[17].

Hand in hand with the publication of the data came the first efforts at large-scale annotation of genomes, identifying genes (again by straightforward homology searches) and other features of the genome that became apparent by pattern matching using existing tools. A continuing theme has been the surprising degree of diversity seen in each new organism, with typically around 30% of open reading frames showing no obvious similarity to other database entries[18]. This lays great importance on one of the abiding problems of structural biology and bioinformatics, that of elucidating function and/or structure from the primary sequence of novel gene products without a clearly significant BLAST hit.

Again, subsequent efforts added more layers of value to the data in ways that were sometimes unique to the nature of the data and that often only became apparent once a 'critical mass' was available. Pharmaceutical companies were particularly interested in the phylogenetic distribution of genes: the most desirable target is one that maintains a high degree of similarity across bacterial clades (thus promising broad-spectrum antibiotics) while being highly divergent from their homologs in humans, the pathogen's unwilling host. Furthermore, it has been suggested that profiling the phylogenetic distribution of genes can be used to cluster genes of similar function, offering clues to function when this would otherwise be obscure[19].

Putative targets must also be essential to the survival of the pathogen and, on occasion, genes can be ruled out

when there are obvious close homologues or other indications of redundant function. It is also possible to examine the complement of genes within a certain pathway or discrete cellular function to help judge the suitability of a target[20]. In many instances, this depends on the completeness of the genome, to give more confidence that there are no unknown components. On occasion, an apparently missing component in a pathway might suggest a misclassification or the need for a more sensitive search[21].

A recent example demonstrating the advantages of whole genomes is a study of the citric acid cycle in 19 complete genomes[22]. Having complete genomes made it possible to reason confidently about the presence or absence of the different parts of the cycle and various shunts and branches, and even about the overall metabolic schemes of the organisms. With only fragmentary evidence, such inferences are much more risky. When gaps are observed, they can direct the search towards possibly overlooked open reading frames or phenomena like gene displacements (different genes coding for proteins that perform the same function in different organisms). Extending our understanding to entire pathways in this way has obvious benefits for pharmaceutical target selection.

A unique aspect of microbial genomes is the ability to observe operon structure and thus likely coexpressions and/or co-regulations, and, more generally, gene clusters that might imply common function[23]. Observations have been made on chromosomal organization on a broader scale that could also eventually give insights of pharmaceutical relevance: for example, conservations of gene order as well as compositional measures could suggest lateral gene transfer[24], which is vitally important to pharmaceutical companies because of concerns about antibiotic resistance[25].

Genomic sequence

Although significant human genomic sequencing has been under way in one form or another for over a decade, with concomitant bioinformatics support for physical mapping and sequence-data management, rates of sequence production have increased markedly in just the past year. This is a result of improving technology and accelerating schedules, owing in part to new competition in the private sector[26]. This has generated considerable interest in this genomic sequence as a source of new or refined targets for the pharmaceutical industry, for several reasons.

By various estimates[9], it seems likely that 10–20% of genes are missing from EST collections, because they have either a very low abundance when expressed or a highly specific pattern of expression, confined to a narrow cell or tissue type, or to a specific time of expression (perhaps alternatively spliced in those contexts). These 'missing' genes might make the best drug targets precisely because they are likely to be highly directed to tissues or disease processes of interest, and to offer high specificity and thus few side effects. Moreover, even though a large proportion of genes have been 'touched' by ESTs, the coverage of coding regions is far from complete and is, in any case, very error prone, based as it is on single-pass sequences.

With the genomic sequence, there is at least the potential to predict more full-length genes on a 'dry' basis (i.e. by computational methods alone). When this can be done, not only high-quality sequences but also the structure of the genes (the alternating arrangement of exons and introns) are immediately available; this is useful, for example, in screening for mutations, which is usually done by extending primers from the introns (not seen in ESTs) at either end of an exon. The flanking genomic sequence may also be available, which could enable the analysis of promoters and regulatory regions, to predict the expression contexts of genes, including nuclear-hormone receptor and other transcriptional-control sites of immediate interest to pharmaceutical companies. The genomic context also provides useful information on clusters of related genes, syntenic (that is, common gene ordering within broad chromosomal regions) relationships in model-organism genomes and candidate genes from mapping studies. Finally, the genomic framework offers the best possible scaffold for clustering, assembling and resolving errors in ESTs; conversely, one of the greatest assets of using ESTs might prove to be their contribution to the elucidation of gene structures in the genomic sequence[27–29].

However, the effective use of genome-sequence data is not without very serious bioinformatics challenges. The data are more numerous and complex than the EST data; the sources of these data are highly varied and dynamic, including FTP and WWW sites in addition to GenBank. Accurately reconstructing gene structures is a difficult computational task, especially given the emerging prevalence of alternative splicing, pseudogenes and other artifacts. The technology of gene finding has evolved considerably, having been one of the major areas of bioinformatics research, but it must still be considered to be an open problem, in particular for long genomic sequences that might contain multiple genes[30,31]. In this case, the problem becomes one of not only accurately assembling exons, which are often numerous and widely-separated, but also properly segmenting the genes and, in particular, determining the 5′ ends (generally under-represented in EST databases). Despite the difficulties, bioinformatics approaches to gene finding in genomic sequences might yet be preferable to ongoing bench efforts to extend

cDNAs to full length (which is laborious and expensive) as a means of determining complete and accurate gene sequences.

Model organisms

With the effective completion of the genomes of the yeast *Saccharomyces cerevisiae* in 1997 (Ref. 32) and the nematode worm *Caenorhabditis elegans* in 1998 (Ref. 33), new approaches were opened up to functional genomics in the human genome (Fig. 2). Not only did many more informative homologies become apparent with human genes but access was also gained through them to biotechnology that could rapidly elucidate the pathways, interactions and so on in which they participated. Thus, the genomic connection to model organisms has brought the notion of the 'wet–dry cycle' to the forefront in requiring close interactions between bioinformatics groups and bench activities.

As always, the initial phase of simple data management was the first concern but the multiplicity of model organisms quickly brought into focus the upcoming problem of intelligent integration of data. With the need to identify links between corresponding genes and gene families in various organisms, a whole series of challenges became apparent. First, extensive, cross-referencing, genome-wide searches of different organisms were required, identifying orthologs as far as possible. Much of this repeated work that was first done with the microbial genomes[34]. To make best use of the resulting links, it also became necessary to take advantage of the annotations to be found in numerous specialized databases that had long been established around individual organisms by their research communities [e.g. FlyBase[35], YPD (Ref. 36)]. (In fact, these formerly purely academic databases have recently taken on such value that they are now successfully charging licence fees for their commercial use, which is particularly interesting to pharmaceutical companies.) This raised the technical issue of simultaneously querying multiple heterogeneous databases to new prominence[37], as this will be necessary to ensure that users are not limited to simple browsing as an exploration paradigm. This problem is being addressed in several ways[38].

The availability of whole-model-organism genomes for comparative purposes should provide advantages for bioinformatics in many novel ways[39,40]. For example, a recent study suggests that protein–protein interactions and functions can be predicted on a large scale by cross-referencing genes that are separate in one organism but fused in another[41]. Complete sets of gene families in model organisms might also reveal new branches to search for in the human sequence. For example, recent work indicates that the nuclear receptor (NR) superfamily, which includes some important drug targets, is surprisingly diverse in *C. elegans*, with substantially more than 200 members detected thus far[42]. Although this explosion of diversity might be limited to nematodes, it does at least suggest that there are many new examples yet to be found in humans and also promises to help guide the search.

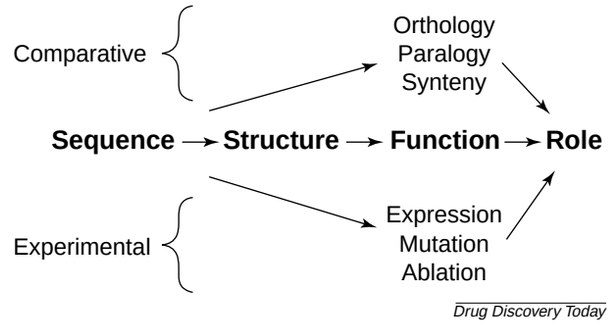

Figure 2. *An important point at which knowledge must be integrated is during the functional assignment of a new putative target, in order to validate it. The traditional sequence–structure–function approach can be supplemented by, for example, the results of comparative genomics and new scaled-up experimental techniques to greater appreciate the role of the target in a particular context.*

As the mouse genome has been completed, there is every reason to expect even greater rewards from comparative genomics. The mouse has useful syntenic relationships with humans more so than with more distantly related organisms[43]. Homologies with yeast and nematodes have been useful for identifying the broad biochemical function of human genes and their position in pathways (e.g. signal-transduction schemes). However, with the mouse, there is a much greater chance of identifying phenotypes that are also directly relevant in humans[44]. Already, the extensive collection of mouse ESTs has proved to be a rich source of useful human homologs[45].

Polymorphisms

Among the latest waves of data beginning to wash over the drug-discovery enterprise is that deriving from genetic markers in the human genome. Genetic maps based on microsatellite markers have been available for several years and have been useful in genetic mapping and helping to localize disease genes through family studies, in many cases leading to candidate genes. However, a new wave of much more densely spaced SNPs is now becoming available, which promises to allow association studies that will be effective even without carefully defined family studies for

disease-related traits. So important is this thought to be that a consortium of pharmaceutical companies has been established that is financing the identification of up to 300,000 SNPs and the mapping of half of them (http://snp.cshl.org/).

SNPs are expected to be the mainstay of efforts not only in the mapping of disease associations but also in pharmacogenetics, which is attempting to identify the genetic determinants of the different reactions to drugs across a population (owing, for example, to differences in drug metabolism or perhaps even variations in target molecules themselves). The so-called coding SNPs (cSNPs) are the most interesting to pharmaceutical companies. These occur in coding regions and thus can themselves be responsible for either disease or a different drug response when they cause non-synonymous changes to proteins. However, these will represent a tiny minority of all SNPs and there is evidence that their allele frequencies are lower, probably because of selection[46].

Several systematic attempts have been made to identify putative SNPs electronically from EST data[47,48], in which cSNPs might be expected to be found most readily. However, this is susceptible to the high sequencing error rate of ESTs and other difficulties[49]. The major contribution that bioinformatics will make to SNP studies begins, as always, with data management. The long-standing discipline of population genetics will provide the primary data-analysis techniques and will also help to establish the parameters of large-scale efforts to accumulate SNPs (Ref. 50), but bioinformatics will provide a key link to the underlying genes and the interpretation of associations.

Gene expression

Tissue localization of gene expression is an increasingly important aspect of target validation. With genomic initiatives providing putative targets in profusion, some notion of function beyond what can be discerned from homology is crucial to the decision to continue with a target. Knowing when and where a gene is expressed can be an important input to this process.

The first contribution bioinformatics made to this question was, as noted above, through counting the ESTs contributing to given transcripts from different libraries. This method has proved to be useful in detecting abundantly expressed genes that are restricted to certain tissues. For example, cathepsin-K expression was highly localized to osteoclasts, the cell type responsible for bone resorption, because approximately 4% of ESTs from a human osteoclastoma cDNA library came from this one transcript[51]. This was apparent after sequencing only a few thousand ESTs but this is an inefficient and probably inaccurate means of routinely assessing transcription levels, especially given that not every pattern can be expected to be so pronounced. Many biotechnologies have become available for this purpose[52], with those that appear to be the most scaleable and adaptable being oligonucleotide-array DNA chips[53,54] and microarray grids[55].

As before, bioinformatics is most immediately concerned with data-management issues[56], although, in this case, the line between bioinformatics and conventional laboratory-information-management systems (LIMS) can become blurred because of the novelty of the technology and the need for close integration. The volume of data expected from microarray experiments will be enormous, given that these systems have now been scaled up to accommodate tens of thousands of targets, in many cases with several conditions and replicates (or even extensive time series[57,58]) for each experiment. In addition, there is a backlog of interesting experiments to attempt. The interest in this technology within pharmaceutical companies is by no means limited to the characterization of putative drug targets; microarrays are also being actively investigated as an aid to studying molecular toxicology, for example, in what has been termed toxicogenomics[59]. In fact, the uses are not limited to the analysis of gene expression. For example, oligonucleotide arrays can be used for the rapid screening of SNPs and thus for exploring variation[60,61].

With the advent of techniques for assessing the expression of unheard-of numbers of genes (and potentially the entire genome) at once, it has become possible to detect clusters of genes exhibiting co-expression under varying conditions or between normal and abnormal tissues[62,63]. Given the signal-to-noise ratios historically available with these techniques and the ambiguity of interpretation of some patterns, it is clear that one of the challenges to bioinformatics in this area is the establishment of consistent analytical and statistical techniques for these experiments[64,65]. Novel techniques for the visualization and data-mining of results are now being actively explored[66,67] and a proposal has been made for a central repository of array results that will enable the meta-analyses of thousands of experiments[68].

When co-regulation schemes have been suggested and the genomic sequence is available (as is already the case for yeast), it will be possible to examine the upstream regions of putatively co-regulated genes for indications of common sequence elements[69,70]. This should add impetus to bioinformatics approaches to understanding the genetic code of gene regulation. Although methods for characterizing and recognizing transcription factor binding sites have not been

impressively useful to date, recent approaches combining more sophisticated statistical techniques and phylogenetic footprinting (comparing the genomic sequence from model organisms to detect conserved regions) have recently shown promise[71]. The ultimate hope of all these approaches, of course, is to discern genetic networks or circuits that control the overall expressed genome[72].

A recent example of such an approach involved the use of data on gene expression during the cell cycle in yeast[73]. Genes were first clustered based on common periodic patterns of expression in the cell cycle. Upstream regions of the clustered genes were then aligned using a method called Gibbs sampling[74], which is well suited to finding common short motifs in sets of sequences. The resulting motifs were then tested against all the clusters and, in many instances, a high degree of specificity was observed for the original clusters, suggesting biological relevance. When such techniques can be applied to human genomic sequences and expression data, it could be possible to understand regimes of gene regulation at unprecedented levels of detail.

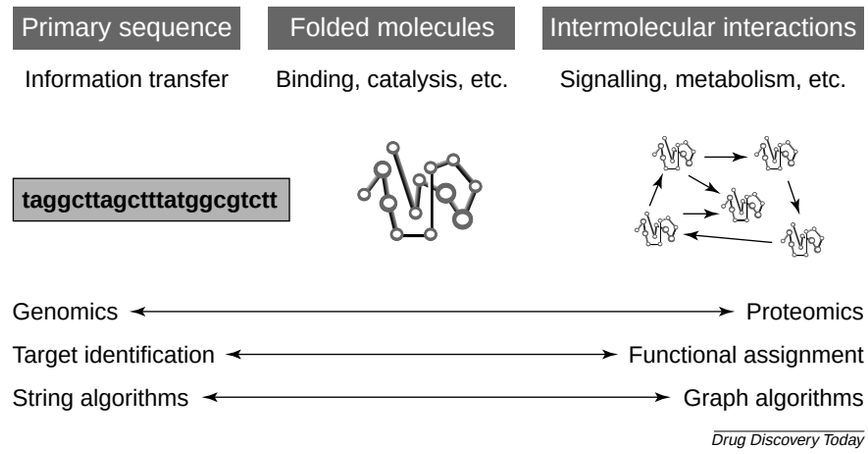

Figure 3. *The concern of bioinformatics in the new, integrative, post-genomic world is shifting from dealing solely with primary sequences to dealing with structural information and, hence, to pathway-oriented knowledge of cellular and organismal physiology. In so doing, the focus of bioinformatics will shift from the genome* per se *to the proteome and even higher-order notions of the physiome. As has been noted, it also entails the transition from target identification to validation or functional genomics and concern for the role of targets in much larger contexts. Finally, the techniques brought to bear by bioinformatics will change; for example, graph algorithms will be of primary importance, together with machine learning, data mining, database integration and advanced representation techniques for the more complex integrative biology of the future.*

Proteomics and beyond

Many believe that the proteome is the next frontier at which bioinformatics will crucially contribute (Fig. 3). Computational methods have already contributed to the large-scale identification of proteins from two-dimensional gel electrophoresis and mass spectrometry[75], and protein microarrays are currently being explored[76]. Structural genomics promises, via high-throughput structure determination, to produce a quantum leap in the number of available protein folds, making fold recognition and comparative protein modelling efforts much more effective[77]. This will undoubtedly be complemented by improved techniques for protein alignment and distant-homologue detection[78,79], creating the strongest connection yet between the sequence-oriented world of bioinformatics and the structure-oriented world of proteins and, hence, with the realm of small molecules and putative drugs. As the 'post-genomic' era ushers in entirely new biotechnologies in increasingly diverse areas, it becomes more crucial than ever to emphasize the integration of these various knowledge sources for functional prediction[80] and, beyond that, for a 'wet–dry cycle' spanning the entire drug discovery process.

Indeed, this must be seen as the greatest challenge facing bioinformatics for the future. The days of target discovery when promising new targets were being generated in profusion have now given way to the urgencies of target validation and a need to reorient bioinformatics efforts to this end. It is not enough to say that, with the easy-to-reach targets gone, bioinformatics must shift its attention to the twilight zone of more distant homologues, because the functional assignment of such targets is often more problematic, and thus they tend to be even further from validation. Rather, it is important for bioinformatics to find ways to extend its useful reach further into the discovery process itself. Some ways to accomplish this are suggested in the introduction. At SmithKline Beecham, these have been manifested in so-called Target Validation Checklists, which are associated with every target in, or about to enter, the pipeline. These checklists consist of a standard template of detailed information on sequences, homologues, variants, mapping information, disease associations, expression data, citations and more, and are updated regularly at each major checkpoint in the progression of a target. In this way, pro-

ject teams are kept up-to-date on the full range of bioinformatics-derived information available about a target in a constantly changing, data-intensive universe.

This is a first step in, but by no means the culmination of, efforts to achieve the sort of intelligent integration of all sources of knowledge useful in decision support, eventually extending to and incorporating clinical evidence that must ultimately validate any target. From the technological perspective, this will also challenge bioinformatics to create database systems capable of integrating knowledge from multiple sources, in fact from multiple heterogeneous databases, in order to span the domains of biological, chemical and clinical data. The acknowledged difficulty of achieving such an integration of diverse database-management systems, schemas, data models and so on, at the level of database technology, is a reflection of the underlying challenge of integrating these world views in the pharmaceutical discovery enterprise itself.

Acknowledgements

I thank James Brown, Pankaj Agarwal, Dominic Bevilacqua and other colleagues at SmithKline Beecham for their helpful comments and suggestions.

REFERENCES

1 Waterman, M.S. (1995) *Introduction to Computational Biology*, Chapman & Hall
2 Gusfield, D. (1997) *Algorithms on Strings, Trees and Sequences*, Cambridge University Press
3 Baldi, P. and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach*, MIT Press
4 Baxevanis, A. and Ouellette, B.F.F. (1998) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, John Wiley & Sons
5 Searls, D.B. (1998) Grand challenges in computational biology. In *Computational Methods in Molecular Biology* (Salzberg, S.L. *et al.*, eds), pp. 3–10, Elsevier
6 Brenner, S. and Lewitter, F., eds (1998) *Trends Guide to Bioinformatics*, Elsevier Science
7 Adams, M.D. *et al.* (1993) 3400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* 4, 256–267
8 Adams, M.D. *et al.* (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* 4, 373–380
9 Williamson, A.R. (1999) The Merck Gene Index project. *Drug Discovery Today* 4, 115–122
10 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
11 Schuler, G.D. (1997) Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med.* 75, 694–698
12 Vasmatzis, G. *et al.* (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl. Acad. Sci. U. S. A.* 95, 300–304
13 Hawkins, V. *et al.* (1999) PEDB: The prostate expression database. *Nucleic Acids Res.* 27, 204–208
14 Burke, J. *et al.* (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* 8, 276–290
15 Chou, A. and Burke, J. (1999) CRAWview: For viewing splicing variation, gene families and polymorphism in clusters of ESTs and full-length sequences. *Bioinformatics* 15, 376–381
16 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512
17 Sutton, G. *et al.* (1995) TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* 1, 9–19
18 Brown, J.R. and Warren, P.V. (1998) Antibiotic discovery: Is it all in the genes? *Drug Discovery Today* 3, 564–566
19 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
20 Ogata, H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34
21 Selkov, E. *et al.* (1997) A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* 197, GC11–GC26
22 Huynen, M.A. *et al.* (1999) Variation and evolution of the citric-acid cycle: A genomic perspective. *Trends Microbiol.* 7, 281–291
23 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
24 Nelson, K.E. *et al.* (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329
25 Brown, J.R. *et al.* (1998) A bacterial antibiotic resistance gene with eukaryotic origins. *Curr. Biol.* 8, R365–R367
26 Venter, J.C. *et al.* (1998) Shotgun sequencing of the human genome. *Science* 280, 1540–1542
27 Xu, Y. and Uberbacher, E. (1997) Automated gene identification in large-scale genomic sequences. *J. Comput. Biol.* 4, 325–338
28 Jiang, J. and Jacob, H.J. (1998) EbEST: An automated tool using expressed sequence tags to delineate gene structure. *Genome Res.* 8, 268–275
29 Bailey, L.C., Jr *et al.* (1998) Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* 8, 362–376
30 Guigo, R. (1997) Computational gene identification: An open problem. *Comput. Chem.* 21, 215–222
31 Claverie, J.M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 6, 1735–1744
32 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 546
33 The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282, 2012–2018
34 Tatusov, R.L. (1997) A genomic perspective on protein families. *Science* 278, 631–637
35 Ashburner, M. and Drysdale, R.A. (1994) FlyBase – the *Drosophila* genetic database. *Development* 120, 2077–2079
36 Hodges, P.E. *et al.* (1999) The yeast proteome database (YPD): A model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* 27, 69–73
37 Markowitz, V.M. *et al.* (1995) Characterizing heterogeneous molecular biology database systems. *J. Comput. Biol.* 2, 547–556
38 Letovsky, S.I. ed. (1999) *Bioinformatics: Databases and Systems*, Kluwer Academic Publishers
39 Koonin, E.V. (1999) The emerging paradigm and open problems in comparative genomics. *Bioinformatics* 15, 265–266
40 Clark, M.S. (1999) Comparative genomics: The key to understanding the

Human Genome Project. *Bioessays* 21, 121–130
41 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
42 Sluder, A.E. *et al.* (1999) The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. *Genome Res.* 9, 103–120
43 Gilley, J. and Fried, M. (1999) Extensive gene order differences within regions of conserved synteny between the *Fugu* and human genomes: Implications for chromosomal evolution and the cloning of disease genes. *Hum. Mol. Genet.* 8, 1313–1320
44 Brown, S.D. and Nolan, P.M. (1998) Mouse mutagenesis – systematic studies of mammalian gene function. *Hum. Mol. Genet.* 7, 1627–1633
45 Marra, M. *et al.* (1999) An encyclopedia of mouse genes. *Nat. Genet.* 21, 191–194
46 Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238
47 Gu, Z. *et al.* (1998) Single nucleotide polymorphism hunting in cyberspace. *Hum. Mutat.* 12, 221–225
48 Buetow, K.H. *et al.* (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21, 323–325
49 Forsberg, L. *et al.* (1999) Low yield of polymorphisms from EST blast searching: Analysis of genes related to oxidative stress and verification of the P197L polymorphism in GPX1. *Hum. Mutat.* 13, 294–300
50 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144
51 Drake, F.H. *et al.* (1996) Cathepsin K, but not cathepsins B, L or S, is abundantly expressed in human osteoclasts. *J. Biol. Chem.* 271, 12511–12516
52 Carulli, J.P. *et al.* (1998) High-throughput analysis of differential gene expression. *J. Cell Biochem.* 30/31 (Suppl.), 286–296
53 Ramsay, G. (1998) DNA chips: State-of-the art. *Nat. Biotechnol.* 16, 40–44
54 Marshall, A. and Hodgson, J. (1998) DNA chips: An array of possibilities. *Nat. Biotechnol.* 16, 27–31
55 Debouck, C. and Goodfellow, P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.* 21, 48–50
56 Ermolaeva, O. *et al.* (1998) Data management and analysis for gene expression arrays. *Nat. Genet.* 20, 19–23
57 Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297
58 Iyer, V.R. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87
59 Nuwaysir, E.F. *et al.* (1999) Microarrays and toxicology: The advent of toxicogenomics. *Mol. Carcinog.* 24, 153–159
60 Hacia, J.G. (1999) Resequencing and mutational analysis using oligonucleotide microarrays. *Nat. Genet.* 21, 42–47
61 Hacia, J.G. *et al.* (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* 22, 164–167
62 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
63 Perou, C.M. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9212–9217
64 Wittes, J. and Friedman, J.P. (1999) Searching for evidence of altered gene expression: A comment on statistical analysis of microarray data. *J. Natl. Cancer Inst.* 91, 400–401
65 Greller, L.D. and Tobin, F. (1999) Detecting selective expression of genes and proteins. *Genome Res.* 9, 282–296
66 Cole, K.A. *et al.* (1999) The genetics of cancer – a 3D model. *Nat. Genet.* 21, 38–41
67 Toronen, P. *et al.* (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.* 451, 142–146
68 Khan, J. *et al.* (1999) DNA microarray technology: The anticipated impact on the study of human disease. *Biochim. Biophys. Acta* 1423, M17–M28
69 Bucher, P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* 9, 400–407
70 Zhang, M.Q. (1999) Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.* 23, 233–250
71 Wasserman, W.W. and Fickett, J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* 278, 167–181
72 Sharp, D.H. and Reinitz, J. (1998) Prediction of mutant expression patterns using gene circuits. *Biosystems* 47, 79–90
73 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285
74 Neuwald, A.F. *et al.* (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Prot. Sci.* 4, 1618–1632
75 Jensen, O.N. *et al.* (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: Strategies and applications. *Proteins* 2 (Suppl.), 74–89
76 Lueking, A. *et al.* (1999) Protein microarrays for gene expression and antibody screening. *Anal. Biochem.* 270, 103–111
77 Sali, A. (1998) 100,000 protein structures for the biologist. *Nat. Struct. Biol.* 5, 1029–1032
78 Altshul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
79 Park, J. *et al.* (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284, 1201–1210
80 Bork, P. (1998) Predicting function: From genes to genomes and back. *J. Mol. Biol.* 283, 707–725