Methodological Overview

Which craft is best in bioinformatics?

T.K. Attwood *, C.J. Miller

School of Biological Sciences, The University of Manchester, Oxford Road, Manchester M13 9PT, UK

Received 11 December 2000; received in revised form 14 January 2001; accepted 16 January 2001

Abstract

‘Silicon-based’ biology has gathered momentum as the world-wide sequencing projects have made possible the investigation and comparative analysis of complete genomes. Central to the quest to elucidate and characterise the genes and gene products encoded within genomes are pivotal concepts concerning the processes of evolution, the mechanisms of protein folding, and, crucially, the manifestation of protein function. Our use of computers to model such concepts is limited by, and must be placed in the context of, the current limits of our understanding of these biological processes. It is important to recognise that we do not have a common understanding of what constitutes a gene; we cannot invariably say that a particular sequence or fold has arisen via divergence or convergence; we do not fully understand the rules of protein folding, so we cannot predict protein structure; and we cannot invariably diagnose protein function, given knowledge only of its sequence or structure in isolation. Accepting what we cannot do with computers plays an essential role in forming an appreciation of what we can do. Without this understanding, it is easy to be misled, as spurious arguments are often used to promote over-enthusiastic notions of what particular programs can achieve. There are valuable lessons to be learned here from the field of artificial intelligence, principal among which is the realisation that capturing and representing complex knowledge is time consuming, expensive and hard. If bioinformatics is to tackle biological complexity meaningfully, the road ahead must therefore be paved with caution, rigour and pragmatism. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Database; Genome; Biological complexity; Ontology; Knowledge representation; Artificial intelligence

1. Introduction

Nobody can doubt that sequencing the entire genomes of diverse organisms, and acquisition of a rough draft of the human genome, represent major technical achievements. But the technological advances that have allowed these breakthroughs now threaten to overwhelm us with data, and problems, of all kinds. At a time when researchers can look ahead to the imminent publication of the complete human genome (February 2001) and rejoice in the making of scientific history, their exuberance is tempered by ominous portents. The now-familiar metaphors of doom used to describe the pace of data acquisition (explosion, avalanche, deluge, tsunami) betray a deep concern; despite the early flood warnings, we appear to have been caught unprepared. Today, we are confronted not only with an increase in the amount of data, but also in the types of accumulating data and in the data structures required to adequately represent the information and its inter-relationships.

Merely increasing the amounts of information we collect does not in itself bestow an increase in knowledge or miraculously endow us with an understanding

* Corresponding author. Tel.: + 44-161-2755766; fax: + 44-161-2755082.
E-mail address: attwood@bioinf.man.ac.uk (T.K. Attwood).
of genomes (Galperin and Koonin, 2000). To achieve this, we must store the information, analyse it in appropriate ways, and ultimately try to understand what it means in order to inform biologists and guide their experiments. Unfortunately, in the panic to automate the route from raw data to biological and biomedical insight, we are generating and propagating innumerable errors both in our valued data collections and in our literature. Thus, bioinformatics has become a Pandora’s box, tempting us with its promise to reveal hidden treasures, but throwing out unbidden troubles when curiosity gets the better of us and we dare to peep inside. In this article, we discuss some of these troubles and note intriguing parallels with developments in artificial intelligence (AI) made during the last 30 years.

2. The trouble with databases

Biological databases are now an essential part of the research environment. But, in many cases, the use we now wish to make of them is different from the motivations behind their original inception. Some have evolved simply as a by-product of a particular research project, with no thought that they might one day become key international resources; others have emerged as repositories to submit data for peer review, rather than as tools to facilitate in silico analyses. Consequently, many are creaking under the strain of information overload, having neither the internal infrastructure nor support software required to meet the demands placed on them.

Today, there are hundreds of databases world-wide, housing information at the levels of the genome, the transcriptome, the proteome, the metabolome, etc. (Fig. 1). The endeavour to cope with, and to rationalise, these vast quantities of data has involved unprecedented levels of global cooperation and ever-increasing levels of automation in data handling and analysis. But automation carries a price. For example, in building genomic databases, software robots are used in the process of functional annotation of newly determined sequences. These pose a threat to information quality, because the annotations they generate are error prone—the information is neither identified as having been computer-generated, nor does it have levels of confidence assigned to it (Brenner, 1999). As a result, it can be used to generate further annotations, allowing misannotations to propagate. So great has been the concern over the resulting error rates that an ‘error

![Fig. 1. In the genome era, bioinformatics is strategically placed to exploit information directly from sequencing projects, from proteomics and metabolomics, from functional and structural genomics, from the scientific literature, from clinical texts, and so on. A more holistic view of biology requires seamless integration of all the data-types emerging from these fields so that bioinformatics can continue to make a contribution to biomedical science. Thus, bioinformaticians will need to work hand-in-hand with, and draw inspiration from, computational biologists, computer scientists, biochemists, medics, and a host of other professionals in order to turn this aspiration into a reality.](image-url)
catastrophe’ has been predicted (Karp, 1998); some have proposed that the problem could be solved simply by eliminating all archived annotation (Wheelan and Boguski, 1998). Although curators are aware of the problems and strive to improve the quality of their resources, we are fast reaching a point where databases are so large that revisions can no longer be made within acceptable time-frames (Karp, 1998). If we are to continue to use these resources, we must accept that they are imperfect historical products (Bork and Bairoch, 1996). Consequently, to get the most from them, and from the tools used to create and interrogate them, it is important to have an understanding both of their powers and of their pitfalls.

3. The task ahead

Work to unravel the meaning of the mass of accumulated genome data is only just beginning. At a superficial level, the task looks straightforward: first we must locate the genes and translate the coding regions to yield their protein products; then perform similarity searches to establish relationships with previously characterised sequences, and hence assign function by evolutionary inference; we must then rationalise the function in structural terms if the structure is known, otherwise use prediction or modeling techniques to derive a feasible model; and finally, given the quantity of data, we must automate these processes as far as possible.

This, of course, is fiction, and significant problems arise at all levels. First, methods for predicting genes in uncharacterised DNA are unreliable, and it is often unclear what we mean by a ‘gene’. Second, it is not safe to propagate functional annotation from one sequence to another merely on the basis of some degree of shared similarity, and it is often unclear what we mean by ‘function’. Third, by comparison with the number of available sequences, there are very few known structures, and methods to predict structure are unreliable. And finally, as already mentioned, the degree of automation that has necessarily taken place, with imperfect tools and protocols, has created unknown quantities of database misinformation. Likewise, the language we use to describe the ‘bread and butter’ approaches of bioinformatics is often misleading, muddying perceptions of what can actually be achieved. So, given the problems, how good are current bioinformatics tools at gene, function and structure prediction?

4. The trouble with counting genes

The types of information used to predict gene structures include signals in the sequence (such as splice sites), content statistics (e.g. codon bias), and similarity to known genes. Each of these approaches has varying levels of success. In a recent test of gene-detection software on part of the fly genome, although the programs could identify 95% of coding nucleotides, intron/exon structures were correctly predicted for only ~40% of genes; overall, the methods failed to find 4.6–95.3% of genes and incorrectly identified up to 55% (Reese et al., 2000). Possibly the most sobering evidence of the frailty of current gene prediction methods is the uncertainty in the number of genes in the human genome. Estimates vary so wildly (currently from ~27 500 to ~312 000) that human gene prediction has been reduced to a sport, and a book opened (Genesweep, 2000) to reward the luckiest punter. Clearly, the methods used to arrive at these numbers differ, each using different approximations and extrapolations. Nevertheless, it is disturbing that the mismatch between the knowledge required to identify genes correctly and that which is encoded in the algorithms is so great that the different analytical approaches should yield such disparate results.

Perhaps the biggest hurdle for accurate gene counting is that the definition of a gene is unclear. Opinions differ as to whether it should be considered as a heritable unit corresponding to an observable phenotype, or as a packet of genetic information that encodes a protein, or proteins (genes can encode two or more proteins), or even RNA. A further confusion is that genes can exist in multiple copies, and genes do not have to be expressed to exist. So how should we count them? For the purposes of establishing rules for Genesweep, a gene is considered to be a set of connected transcripts, and a set of guidelines is offered as to what will and will not count as a gene within the contest. Even so, estimates of the total gene complement of the human genome still vary.

5. The trouble with counting sequences and structures

There are currently >590 000 protein sequences in NCBI’s non-redundant database (NRDB), and many millions of ESTs stored in public and proprietary repositories. These numbers will increase inexorably with the fruition of further genome projects. By contrast, the number of available structures is <14 000. But these figures are highly redundant and are skewed by a handful of biologists’ favourites (e.g. globins, lysozymes, immunoglobulins, etc.). In fact, the number of unique protein structures is probably still less than ~2000. Of course, we do not know the number of unique sequences—in spite of its name, NRDB is not non-redundant, but non-identical, and is thus massively redundant. Nevertheless, by comparison with the number of available sequences, there is clearly a dearth of
In isolation, protein sequences and structures do not inherently tell us function. However, when we look at alignments, patterns of conservation emerge (motifs) that begin to provide functional clues. Thus, without other information, we cannot easily predict the function of the sequence in (a) or of the structure in (b); but the glycine/serine/threonine-rich region captured in the motif in (c) provides evidence that allows us to infer a role in ATP-binding.

Fundamental to the process of sequence annotation. Unfortunately, the meaning of this simple concept has been widely abused, polluting the literature with nonsense statements that equate homology with similarity. Sequences are homologous if they are related by divergence from a common ancestor (Fitch, 1970). Analogy, on the other hand, relates to the acquisition of common features (folds or functions) via convergent evolution from unrelated ancestors: e.g. β-barrels occur in soluble serine proteases and integral membrane porins, which, despite their common architecture, share no sequence or functional similarity; similarly, the enzymes chymotrypsin and subtilisin share groups of catalytic residues, with near-identical spatial geometries, but no other sequence or structural similarities. Homology is an absolute statement that sequences have a divergent rather than a convergent relation-
ship—it is not a measure of similarity. This is not just a semantic issue. Woolly use of the term obscures our understanding of the evolutionary relationships we are trying to determine. Database search tools identify similarity between sequences—it is for the researcher to ascertain, on the basis of other evidence, whether the similarity is biologically significant and hence whether the sequences are likely to have derived from a common ancestor. The same arguments apply when comparing structures. Structures may in fact be similar, but their common evolutionary origin is a hypothesis that may be correct or mistaken (Rheeck et al., 1987). For a more complete discussion of the confusions relating to the use of the term homology, see Fitch (2000).

Among homologous sequences, it is useful to distinguish proteins that usually perform the same function in different species (orthologues) from those that perform different but related functions within one organism (paralogues). The study of orthologues allows us to chart changes in a given protein across species. By contrast, paralogues arise via gene duplication events—the duplicated genes follow separate evolutionary pathways, allowing new specificities to evolve through variation and adaptation. Such complexity presents significant challenges for bioinformatics. For example, following a database search, how much functional annotation can be legitimately inherited by a query? Many automatic analysis systems have operated on the assumption that it is sufficient to transfer the annotation from the best-matching hit. But how can we be sure that the best match is the true orthologue and not a parologue? Such naïve approaches do not consider the functional plasticity that may arise through divergent evolution, and have spawned numerous annotation errors in sequence databases (Gerlt and Babbitt, 2000).

8. The trouble with modules and domains

Additional problems have arisen from the failure properly to consider the domain and/or modular nature of some proteins. Modules are autonomous folding units, used like protein building blocks to confer a variety of functions on a parent protein, either by multiple combinations of the same module, or via different modules to form mosaics. Following a database search, the best hit could be a match to a single domain or module within a multidomain protein. In this case, inheriting the annotation from the parent protein itself, and not from the domain, is likely to be inappropriate.

As elegantly described by Jacob (1977), nature behaves in the manner of a tinker, fashioning old material to create new systems rather than starting from scratch. Mosaic proteins are excellent examples of nature's tinkering, where individual modules have been recruited for context-dependent roles in a range of different proteins. Such complexity poses important problems for computational systems: knowledge of the particular types of building block used to make a protein, without knowledge of their biological context, is often not sufficient to infer its function (Jacob, 1977; Gold et al., 1997). Thus, identifying the presence of a module tells little of the function of the complete system; knowing most components of a mosaic does not allow us easily to predict a missing one; and modules in different proteins do not always perform exactly the same function. Appreciating the complexity of biological systems and gaining a greater understanding of the evolutionary processes that generated biochemical, cellular and developmental innovations (Jacob, 1977; Gogarten and Olendzenski, 1999) will be paramount if bioinformatics approaches are to improve.

Comparative genomics has shed light on still more complicating factors: gene functions may be redundant; non-orthologous displacement can replace genes with unrelated but functionally analogous genes (Galperin and Koonin, 1998); horizontal gene transfer can introduce genes from different phylogenetic lineages (Fitch, 2000); and lineage-specific gene loss can eliminate ancestral genes. Genomes thus harbour many obstacles that defy an intuitive correlation between sequence, structure and function, and render reliable, fully automatic computational assignment of protein function almost impossible.

9. The trouble with defining function

Understanding what we mean by protein function is not straightforward; it is multifaceted and context dependent (Bork, 2000). Unfortunately, researchers have been vague in using the term, with the result that many database annotations are confused, and confusing. To understand the function of a protein, we need to know what action and what role it performs in the cell (Karp, 2000). In terms of its action (or local function), we want to know with what other cellular components it interacts and of what sort are the interactions (e.g. protein–protein, protein–carbohydrate, protein–lipid, protein–DNA, etc.). To understand its role (or integrated function) is to appreciate how the cell behaves if the protein is absent. But proteins do not function in isolation. They generally have important roles within complex systems and perhaps within several systems. Thus, function can be understood in different contexts and at different levels of complexity, from molecular and cellular, to tissue and organismal. How then can we deconstruct this information to provide a meaningful and accurate one-line database description that adequately encompasses the function of a protein?
Until recently, function has been used to refer somewhat arbitrarily to biochemical activities, biological goals and cellular structures. Thus, for example, we may find the function of actin being described in different places as ‘ATPase’ or ‘constituent of the cytoskeleton’. However, in an attempt to better reflect biological reality, and to introduce rigour into the field, several ontologies have now been created. Ontologies are formal representations of key concepts and the relationships between them, captured in a way that allows computer programs to manipulate the knowledge they encapsulate. In the context of bioinformatics, ontologies are being developed to represent descriptions of function and functional relationships, and will become increasingly important as more genome sequences are generated and require automatic annotation. Notable examples are those aiming (i) to use a set of controlled vocabularies to define more explicitly the relationships between gene products and biological processes, molecular functions and cellular components (Ashburner et al., 2000); (ii) to encapsulate the notions of local and integrated function (Karp, 2000); and (iii) to define relationships between enzymes, their locations in metabolic pathways, their reaction products and substrates, and the superfamilies to which they belong (Goto et al., 1997).

10. The trouble with structure prediction and fold recognition

We have seen that it is hard to define what we mean by a ‘gene’, making it difficult to count genes accurately, and that it is hard to define ‘function’ rigorously, making function assignment tricky. However, at least we know what structures are. They are tangible, measurable things, so should we not be able to predict them reliably?

Structure prediction methods include computationally intensive strategies that simulate the physical and chemical forces involved in protein folding, and knowledge-based approaches that use information from the structure databases to build models. Yet, despite more than 30 years of research, the problem of predicting protein structure remains unsolved: homology modeling and threading techniques typically produce low-resolution models, and no current method yields reliable predictions for remote homologues (Rost and O’Donoghue, 1997). For small proteins, ab initio methods generate models with segments that resemble the correct fold, but results deteriorate beyond ~100 residues. Today, knowledge-based methods, especially those that combine information from different approaches, give the best results (Panchenko et al., 2000). For modeling and fold recognition, advances have accrued in particular from balancing better algorithms with appropriate levels of manual intervention (Sternberg et al., 1999).

Prediction methods do not work well because we do not fully understand how the primary structure of a protein determines its tertiary structure (the protein folding problem)—we cannot yet read the language that sequences use to create their folds. However, structural genomics projects will gradually lessen our reliance on prediction, as they aim to provide experimental structures or models for every protein in all completed genomes. This lofty goal is tempered by an acceptance that the structures we can determine will be limited to those that respond best to the available biophysical techniques (e.g. those proteins that will crystallise, or those that are small enough for NMR); membrane protein structures will remain difficult to obtain.

It is important to put the value of protein structure in perspective. Protein folds provide different scaffolds, which can be decorated in different ways by different sequences to confer different functions. Knowing both the fold and the function allows us to rationalise the mechanistic process by which the structure affects its function at the molecular level. Structures provide one part of a complex biological puzzle. Without supporting information, knowing structure alone will not inherently tell us function (Fig. 2). For example, determining the structure of a hypothetical protein and discovering that it binds ATP (Zarembinski et al., 1998) may shed light on possible aspects of its functionality, but it does not reveal its specific biochemical and cellular role.

Structures, and structural alignments, are often described as the ‘gold standard’ or ‘standard of truth’. They are not. Structure alignments, like sequence alignments, are simply models from which we can make biological extrapolations. It is important to understand the strengths and weaknesses of these models, and pointless to assert superiority of structure alignments over sequence alignments—both are equally valid models for the different dimensions of information they are intended to represent and for the different questions they are intended to answer.

11. The trouble with defining structure

Just as we can view function at different levels of complexity, so we can think of structure at different levels: from the high-resolution view of its atomic coordinates, to increasingly lower-resolution views of its topology (the connectivity of its secondary structures), its architecture (the gross arrangement of its secondary structures), or its structural class (mainly α, mainly β, etc.). In the context of fold recognition and prediction, therefore, it is helpful to be precise about the level of structure we are meaning. For example, does a ‘good’
prediction correctly reproduce atomic positions, topology, architecture, or merely the structural class? Where does a ‘reasonably good’ prediction fall in the structural hierarchy? And what level of structural detail does a ‘tough near miss’ (Olszewski et al., 2000) reveal?

In predicting genes, protein functions and structures, it is helpful to define our terms precisely and to be honest about our achievements, otherwise bioinformatics will descend into Babeldom (Attwood, 2000) and we will continue to be baffled by paradoxical new ‘prediction’ methods that yield >80% error rates. Getting computers to solve biological problems is hard, but with the relentless accumulation of sequence data, improvements continue to be made in all areas.

12. The challenges ahead

With the first fruits of the human genome now available, and the genomes of many organisms now complete, bioinformatics is strategically placed to develop the necessary tools to help the research community both to interpret and to exploit the accumulating molecular and cellular data. However, the true challenge for bioinformatics lies not with the quantity, but with the complexity of post-genome data, in part because the proliferation of data is allied with a similar growth in data-types that echo the biological concepts we wish to consider (Fig. 3). Thus, for example, in order to improve existing techniques for gene or function prediction, proteins must be described in terms of their constituent modules or domains, genes in terms of their internal structure, and genomes in terms of their gene products.

If bioinformatics is to make use of new data sources, proteins will have to be considered in the context of their molecular and cellular roles. It is the totality of molecular and cellular networks that defines the phenotype of an organism: changes in the networks resulting from mutation or altered gene expression lead to disease phenotypes. The full value of informatics to medical science will only be realised by assimilating this complex information and offering more holistic analyses of biological and pathological processes than have been possible before. For example, it will be necessary to integrate gene sequences, polymorphisms, splice variants, protein structures, phenotypes and functions, biological texts, clinical data, etc. (Fig. 1). The sum total of this knowledge, captured for tens of thousands of genes, will result in millions of facts that need to be characterised, classified and manipulated. This is in stark contrast, say, to the 400 numbers used by an alignment algorithm to capture its knowledge in a scoring matrix.

It has been suggested that databases of gene structure and expression patterns may prove insufficient to elucidate gene function and, even when integrated, they will be of limited value, because they do not contain information on non-linear responses and thresholds (Miklos and Rubin, 2000). Thus, our ability to unravel, for example, the complex networks that control gene expression and differentiation, cell cycle regulation, signal transduction, and so on, will depend on a deeper understanding of the time- and place-dependency (context), complexity and degeneracy of biological systems (Miklos and Rubin, 2000).

13. The trouble with integrating existing tools and data

Simply representing knowledge is pointless without a set of appropriate mechanisms for navigating the
knowledge-base in order to extract relevant information. Moreover, to allow databases to interoperate, both with the algorithms that are going to make use of their data and with other databases, requires well-defined Application Programming Interfaces (APIs), and standard formats for data exchange. The volume and complexity of the data mean that this is not straightforward (Lewis et al., 2000).

In order to make software and databases exchange data efficiently, bioinformatics needs more robust plumbing. Currently, the output of one piece of software is typically piped into the input of the next by using a set of Perl scripts to provide the data transformation. This is inefficient, time-consuming and brittle—a simple change to the output file format can easily break the software, or worse, make it perform inconsistently. A more subtle issue concerns the way the problem scales combinatorially. Connecting \( N \) outputs to \( M \) inputs requires \( N \times M \) different scripts. Two technologies are often proposed as solutions to this problem: XML and CORBA. Whilst these both provide a standard syntax for defining the data to be exchanged (XML using a set of tags similar to those used in HTML, CORBA using a special Interface Definition Language, or IDL), the problem is more one of semantics than of syntax. For example, how does one compare the results of three different alignment algorithms, let alone compare them with the output of a neural network or a Hidden Markov Model? CORBA and/or XML might help make the pipes fit together, but they do not tackle the more onerous task of dealing with what one puts down them in the first place.

14. The transition from numeric to symbolic computation

In a move beyond mere comparison of biological sequences and structures, bioinformatics is becoming increasingly focused on manipulating their annotations rather than the sequences or structures themselves. This is associated with a switch from numeric algorithms, such as those that generate a similarity score from an alignment, to symbolic ones, which manipulate entities (such as ‘promoter’ or ‘helix’). This is similar to the paradigm shift AI underwent in the 1970s, and should set alarm bells ringing.

One of the early goals of AI was to produce computer programs that were able to read and interpret written text so that it could, for example, be translated into a different language. In order to do this effectively, it is necessary to do more than merely apply syntactic rules. Successful translation relies on understanding the meaning of the text, not just the grammatical relationships between the words. This requirement, coupled with similar demands from other arenas, resulted in an enormous amount of effort being focused on the development of ontologies capable of representing the kind of common-sense knowledge we take for granted.

The largest of such systems, Cyc (Guha and Lenat, 1990), arose out of a multi-million dollar research grant, and aimed to produce a system capable of reading and automatically assimilating any encyclopedia article, and thus augmenting its knowledge. Cyc is an ontology and reasoning engine, containing about 100 000 primitive concepts, with many more composite ones generated automatically. Each concept is involved in a few dozen assertions, resulting in a semantic network containing millions of relationships (Cycorp, 2000).

Interestingly, the contribution made by Cyc to the AI community has been in terms of the tools, mechanisms and techniques developed in the course of producing such a large body of formally structured data, rather than its availability as an information resource for use by the wider community. Fully automated machine translation systems, backed by ontologies such as Cyc, have failed to materialise. The state of the art in machine translation now relies on a collaborative relationship between man and machine, in which the computers concentrate on laborious, dumb jobs (such as dictionary-lookup), while a human operator injects the spark of intelligence necessary to make the system work. It is likely that successful bioinformatics analysis will take a similar course—the Holy Grail of fully automated annotation will be abandoned in favour of a more pragmatic, machine-aided, decision-support approach.

15. What hope is there?

With the embarras de richesses bequeathed to bioinformatics by the genome projects, we have entered an encyclopaedic era, driven on by a seemingly insatiable desire to capture ‘omic’ data of all kinds (Fig. 1). To date, however, bioinformatics has only allowed us to pick the low hanging fruit; if we are to synthesise our current flake of knowledge (Eisenberg et al., 2000) into a future corpus of understanding, it is clear that much more demanding tasks lie ahead. The identification of genes and the prediction of protein structure and function are non-trivial computational tasks. But will they remain arcane arts, or will they one day allow us to compute biology effortlessly?

Experience with projects such as Cyc has shown us that any attempt to capture a significant amount of complex knowledge is time consuming, expensive and hard. The majority of the facts encoded in the Cyc Knowledge-Base have been encoded manually, because ‘‘the hardest truth to face, one that AI has been trying to wriggle out of for 34 years, is that there is probably
no elegant, effortless way to obtain this immense knowledge base. Rather, the bulk of the effort must (at least initially) be manual entry of assertion after assertion.” (Guha and Lenat, 1994). Cyc thus represents at least a person-century of work, conducted by highly skilled professionals.

With its tightly woven web of interactions, locations, entities and parts, changing over time and experiment, bioinformatics confronts us with data that are of similar complexity to the common-sense knowledge in Cyc, and of similar scale (assuming we wish to represent the 30,000–100,000 genes in the human genome). If bioinformatics is going to meet the challenge of fully automated functional assignment successfully, a monumental effort on the scale of Cyc is required, and we must hope that the lessons learnt, and the experience gained, by our colleagues is sufficient to guide us to a successful conclusion. Alternatively, like AI, we can seek compromises. We are fortunate that bioinformatics is an application-driven discipline—it exists to help life scientists answer questions about their specific domain, and occasionally to generate new hypotheses for exploration. This offers two potential routes forward: one limited by the domain in which software is expected to operate, the other by the amount of work the software is expected to undertake without human intervention.

A fully automated system is plausible, but the kind of ab initio knowledge generation promised by the most extreme forms of data mining can realistically only be achieved within limited domains, and then only after an expert has spent a considerable amount of time in the company of a knowledge engineer. Whether their time could be better spent investigating a specific aspect of the data, rather than attempting to describe everything they know, in the hope that the computer will generate new knowledge, is a moot point. Another approach is to build more comprehensive systems, using controlled vocabularies or taxonomies to help guide a human being through the data, but, like the commercially available translation systems, the spark of intelligence will come from the user, not the machine.

We should not underestimate the body of knowledge and well-honed tools that AI has developed over the last half-century in tackling many of the issues that now face bioinformatics. This not only offers us a significant head start in terms of available software, but also provides a set of techniques that have been described in papers and books, and, perhaps most importantly, a set of trained individuals who know how to use them. Associated with the technical and intellectual challenges of adequately capturing bioinformatics knowledge is the social one of establishing meaningful dialogues between biochemists, medics, computer scientists, computational biologists and all of the other professionals involved in the enterprise. AI provides us with a ready-made language for communicating many of the issues associated with knowledge representation, and valuable lessons on how best to move forward.

16. Conclusion

The head start that bioinformatics has gained from the foundations of AI is of limited value if future experiments and computational tools are not designed in concert. High-throughput approaches compel us to think in a disciplined way about the experiments we should perform (Murray, 2000). It is pointless merely to collect ever-increasing volumes of data without a full understanding of the questions we are trying to answer; of whether the technology is in fact suitable for answering those questions; and of whether the models available to represent the data are compatible. In short, the more data we have to handle, the more rigorous we must be if we are to make sense of the complexities.

Appreciating the complexity of biological systems and the difficulties of making predictions about them is not new: Jacob (1977) noted that the processes responsible for blood coagulation, inflammatory reactions against foreign agents, and complement-mediated immunological defences exhibit unexpected complexity. He observed: “For the biologist, it is thus generally impossible to make a prediction, or even an inspired guess, about the nature of such molecules and their structural relations with other constituents”.

In reviewing some of the troubles arising from the last decade of bioinformatics research, we have tried to paint a realistic picture of the significant difficulties that must be overcome if bioinformatics is to continue to make a contribution to the advancement of biomedical science. We have seen that none of the current tools of the trade is best, none is able to address biological complexity in an effective manner. AI has highlighted the need to define concrete, specific concepts before attempting to tackle abstract ones, and has shown us the wisdom of trying to establish the right relationship between man and machine. There is thus clearly a trade-off to be made between the amount of knowledge we are able to represent, the level of ambiguity in that representation, and the time we choose to spend on the task. As Fig. 4 shows, a reductionist approach is required to direct our efforts appropriately—i.e. while we cannot adequately describe all the relationships depicted in the figure, it is possible systematically to model deep, but thin, slices through it. This satisfies our need to acquire holistic views, but limits the domain in which we operate.

The challenges facing bioinformatics are labyrinthine, each begging a harder question than the one before. The convoluted, almost fractal nature of the problems cautions that what we see depends on the perspective
Fig. 4. A true representation of cellular function involves considering the context in which genes or gene-products occur. Whilst many databases hold particular types of data, such as expression, sequence or interaction (depicted by the white bands), a full functional description requires all these levels to be considered in concert. A complete description of everything would require many decades of work; one restricted to a narrow domain (bold) is more likely to be achieved in a realistic timeframe.

from which we look. Bioinformatics must take care to find the right vantage point.

References


