The Pharmacogenomics Journal (2002) 2, 207–208

www.nature.com/tpj

PERSPECTIVE

Bioinformatics in a post-genomics world—the need for an inclusive approach

JO McInerney

Bioinformatics and Pharmacogenomics Laboratory, National University of Ireland, Maynooth, Ireland

The Pharmacogenomics Journal (2002) **2**, 207–208. doi:10.1038/sj.tpj.650111

We need to understand what happens in a cell. If we understand this, we have a better chance of understanding what is happening in the organism and if we understand what is happening in the organism, we stand a better chance of understanding what happens when the organism interacts with its environment. This argument, above all others has been driving molecular and cellular biological research for more than sixty years. The answers to what is happening in the cell, will greatly improve our chances of correcting defects, designing drugs to interfere with cellular processes, dealing with disease susceptibility and why there is variation in drug response.

A decade ago when the genomic sciences began to move from the 'cottage industry' of single gene sequencing to genuine high throughput, the tasks facing bioinformatics were clear. What were needed were detailed and accurate comparative genomic analyses. Excellent bioinformatic tools for database searching,[1] multiple sequence alignment[2] and phylogeny reconstruction[3] were already in existence and when the data deluge arrived, bioinformatic methods were sufficiently well-developed to deal with the situation.

Today, much of the focus has shifted to the post-genomic sciences. The development of new technologies such as microarray devices and advances in high-throughput polymorphism detection methods have forced the development of novel algorithms and software tools with which to analyse these data. Bioinformatics is now primarily concerned with two issues—analysis of interactions and analysis of sequence variation.

To fully appreciate the problem, consider that if a proteome is only made up of four proteins there might be 11 interactions (AB, AC, AD, BC, BD, CD, ABC, ABD, ACD, BCD, ABCD). For the entire human genome, the number of possible interactions is enormous. To illustrate the point we can take a recent report of two siblings with the same genotype that were observed to have different phenotypes.[4] The report centered on a condition known as congenital chloride diarrhoea (OMIM 214700), which is a recessively-inherited defect of intestinal electrolyte absorption. It was observed that the two siblings, both having the same mutation in the effecting gene (SLC26A3) had very different clinical outcomes.[4] To make computer-based predictions concerning this disease is obviously much more difficult than simply identifying mutations and supposing that the genotype–phenotype relationship is easily explained. Unfortunately, however, computational modeling of the interactions between all proteins in a cell is unlikely to be a realizable objective in the near future, simply because of the high dimensionality of the interaction space.

The problem I have just outlined would not end even if we had a reasonable idea of how most of the interactions came about. Once the basic interactions between proteins and other molecules have been worked out, it will also become necessary to account for variation in DNA and protein sequence. Even in a small system such as HIV, drug resistance mutations have compensatory substitutions that restore sensitivity to antiviral drugs and *vice versa*.[5] It would be difficult enough to account for all the things that HIV is capable of doing, without having to worry about all the variants of HIV and how these variants interact with the host immune system and how they interact with anti-retroviral therapies. However, this is the task with which we are presented. In humans—a much more complex organism than HIV—similar situations are also likely to present themselves. Variation in sequences results in variation in interactions and this results in variation in phenotype.

Complicating issues even more, we have the situations where environmental conditions affect disease penetration and protein expression levels, gene frequencies vary across populations and the methods of collecting data vary between laboratories and are not documented with any degree of rigor in our current databases. These issues must be addressed going forward.

Fifteen years ago, it was supposed that most phenotypes (even complex phenotypes) were due to small numbers of genes having large effects. Although it was thought that there might be some interactions between these individual components, protein–protein interactions were not thought to be nearly as important as the expression levels of important individual genes. Today, this picture has changed dramatically. The current feeling is that most complex phenotypes are due to interactions between large suites of genes, each of which has a small overall effect.[6] In yeast, for instance a total of 234 genes were

identified whose expression profile changed in response to compounds that inhibit ergosterol biosynthesis.[7] A number of years ago, nobody would have predicted that such a profound change in expression profile would have been seen and that so many genes would be involved. Even more intriguingly, a study of human fibroblasts showed that a large number of genes appear to be transcriptionally-active even though transcription of their mRNAs would normally be associated with other tissue types.[8]

How do we predict cellular interactions computationally? How do we integrate DNA and protein sequence variation into our models of biochemical pathways? In fact at the moment, it is proving difficult to even manage the results of the experiments that give us these data, not to mention predicting protein–protein interactions and gene expression patterns *in silico*. The solution will require a combination of bench experiments, computational analysis and most importantly, integration.

The obvious solution to the problem of predicting protein interactions is to employ biology as a collaborator. In the same way that solutions for database searching were greatly helped by our knowledge of evolution (descent with modification from a common ancestor), solutions to the new bioinformatic problems will be greatly helped by our knowledge of biology. Many of the central biochemical pathways are known. The current literature contains more than 11 million references (Source: Pubmed). Methods for effectively mining these data[9] are becoming more sophisticated and promise to improve as we move forward. Integration of the results of the millions of experiments that have been carried out over the past 50 years combined with the results of new experiments is a powerful part of the discovery process and a challenge for bioinformatics.

One of the most frequently used methods of finding genes responsible for a particular trait is to measure changes in their expression.[10] Similarly, the newest of the laboratory technologies—proteome arrays[11]—is promising to be one of the most exciting. If the results of these and other experiments are integrated together, then the prospects for future knowledge generation will be greatly enhanced. The role of bioinformatics in this effort will be as much about data integration as algorithm development.

There are still, however, issues that need to be tackled at the algorithmic level. The methods for data analysis are—as usual—the subject of some debate. In the area of microarray research, clustering methods,[12] multivariate analysis methods,[13] and others are all being used with a great deal of effectiveness. However, there are outstanding concerns relating to the appropriateness of each method and their accuracy. Evolutionary analyses of DNA and protein sequences are also conducted in a variety of different ways with no single method gaining universal approval from all concerned. Methods that seek to identify 'important' amino acid residues (those that are under either strong negative or positive selection pressure)[14] are central to the understanding of protein function, however again there is lack of agreement concerning the choice of method.[15]

The requirements for bioinformatics have changed. Ten years ago, it was possible to develop algorithms knowing the datatype with which you were working and the desired outcome (cf database searching algorithms, alignment algorithms etc). Today, with the focus centered on how the organism works and how to integrate large amounts of data, it will be necessary for a more intimate working relationship between bioinformaticist and bench scientist. It will be no more reasonable for the bioinformaticist to work independently on problems relating to cell biology than it will be sufficient for bench scientists to work without bioinformatics. To build a picture of the interactions in the cell, strategies will have to be devised where bench science and bioinformatics work hand-in-glove. Then perhaps in the not too-distant future we will understand what is happening in a cell.

DUALITY OF INTEREST

None declared.

Correspondence should be sent to
JO McInerney, Bioinformatics & Pharmacogenomics Laboratory, National University of Ireland, Maynooth, Co. Kildare, Ireland.
Tel: +353 (0)1 708 3860
Fax: +353 (0)1 708 3845
Email: james.o.mcinerney@may.ie

1 Pearson WR, Lipman DJ. *Proc Nat Acad Sci USA* 1988; **85**: 2444–2448.
2 Higgins DG, Sharp PM. *Gene* 1988; **73**: 237–244.
3 Felsenstein J. *PHYLIP: Phylogenetic inference package*. In. 3.5c ed. University of Washington: Distributed by Author; 1993.
4 Hoglund P *et al. Gut* 2001; **48**: 724–727.
5 Nijhuis M *et al. Aids* 1999; **13**: 2349–2359.
6 Phillips RL *et al. Science* 2000; **288**: 1635–1640.
7 Bammert GF, Fostel JM. *Antimicrob Agents Chemother* 2000; **44**: 1255–1265.
8 Iyer VR *et al. Science* 1999; **283**: 83–87.
9 Tanabe L *et al. Biotechniques* 1999; **27**: 1210–1214, 1216–1217.
10 Alizadeh AA *et al. Nature* 2000; **403**: 503–511.
11 Zhu H *et al. Science* 2001; **293**: 2101–2105.
12 Lukashin AV, Fuchs R. *Bioinformatics* 2001; **17**: 405–414.
13 Fellenberg K *et al. Proc Natl Acad Sci USA* 2001; **98**: 10781–10786.
14 Yang Z. *Mol Biol Evol* 1998; **15**: 568–573.
15 Suzuki Y, Nei M. *Mol Biol Evol* 2001; **18**: 2179–2185.