The industrialisation of BIOINFORMATICS

With the completion of the 'draft' sequence of the human genome predicting an estimate of 40,000-120,000 genes that describe us, the real work begins. The problem now is not lack of data but lack of tools to thoroughly analyse what we already have and what's coming in the future will only make that chasm wider.

By Dr Alex Titomirov

In recent weeks, we've witnessed the penultimate step in one of the greatest accomplishments of human science: the completion of a 'draft' sequence of the human genome as a joint effort of public and private sequencing groups. And while there is work remaining to create a 'finished' sequence, this should be easily completed in the next one to three years. What most people outside of the genomics industry do not recognise is that sequencing is only the first step of many to realise the true potential of the human genome.

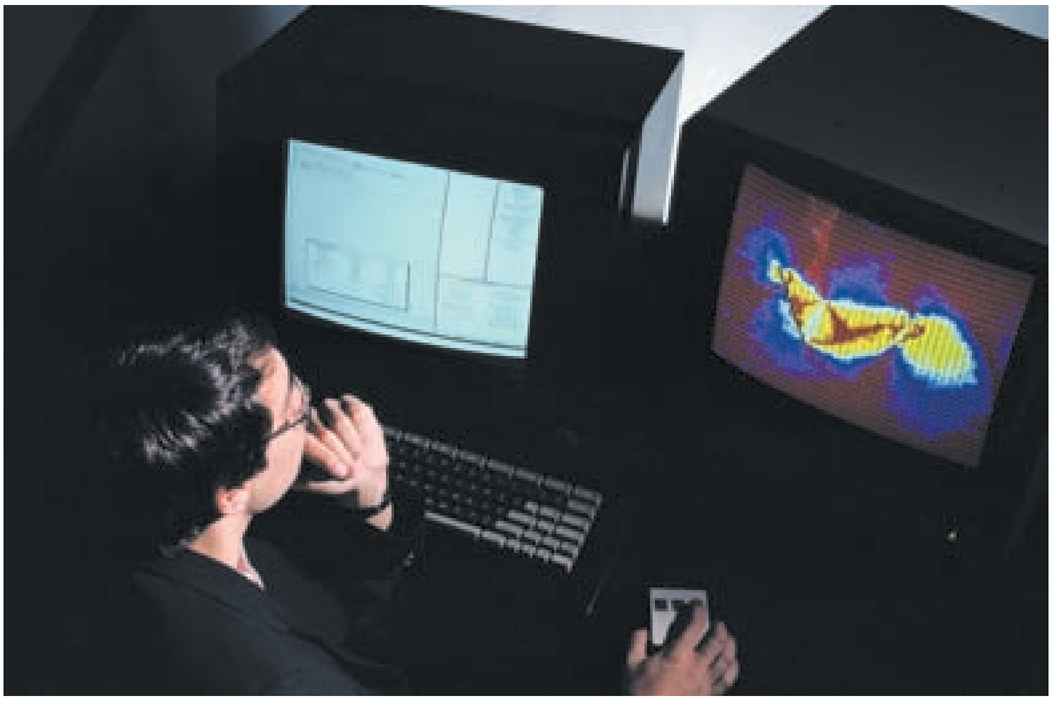
Although the numbers vary, best estimates suggest that a total of 40,000-120,000 genes describe 'us'[1-3]. The drugs currently available today only target about 500 of that total. So one major benefit to human health of sequencing the human genome is that many new targets will be identified as suitable for pharmaceutical intervention. Scientists suggest as many as 5,000-10,000 of the total collection of genes may end up being 'targetable' by pharmaceuticals or biotechnology products. And lest we think that human genomics is the only game in town, there will be more than 60 bacterial genomes, along with drosophila, and perhaps the mouse and rat genomes, completed by the end of 2000[4]. Clearly there's a lot of open ground available to stake a claim for the next blockbuster.

Unfortunately, the problem facing us now is that it is not at all obvious what each gene does, let alone which gene products will make good therapeutic targets. Researchers have already begun to investigate what all these genes actually do but this represents a huge collection of essentially unconnected research projects, even within a single organisation. The problem now is not the lack of data but the lack of tools to analyse the huge amounts of diverse data associated with each gene. And, as new genes are investigated using new techniques, the problem will only increase in magnitude and scope.

Data diversity, not just quantity, is also causing a bottleneck. Scientists are not content with having a gene's sequence, they also want data on the expression of that gene, the levels of translation, the post-translational modifications of that gene product, the enzymatic function (if any) or the other proteins with which it interacts, and the functional pathways involved. In addition, how do these same data vary across different tissues, disease states or drug treatments? As you can see, this creates an extremely complex, multidimensional web of information.

There are many new technologies available to investigate all of these questions. New chip and array-based technologies can generate tens of thousands of data points and their potential relationships in an afternoon. While it is good that such data can be gathered quickly, the sheer magnitude of data generated further complicates both data acquisition and analysis problems. The more data analysis required, the more difficult it is to see the forest for the trees.

Of course, it is not actually the gene directly but rather the gene product (protein) encoded by that gene that carries out a biological function, and proteins have conformations in three dimensions, not just linear sequence information. The task of determining that 3D protein structure is also extremely computationally intensive. It is clear that even though we have the genome sequence, it is merely the starting point.



Bioinformatics is a critical part of the drug discovery process for today's bench scientist

Drug discovery and the post-genomic era

Although many researchers will benefit from the knowledge obtained from the Human Genome Project, it is the pharmaceutical and biotechnology industries that have the most to gain. The traditional process of drug discovery that pharmaceutical companies have been using for years results in, on average, one to two New Drug Entities (NDEs) per year, with the entire process frequently taking 10 or more years from discovery to product rollout[5]. For obvious social, health and economic reasons, it is desirable to speed up the drug discovery process as much as possible.

The human genome data holds great promise and will transform drug discovery in a number of ways. Knowing our complete genetic blueprint opens up many new diseases for treatment, as well as new strategies for existing treatments. Since many diseases are 'multigenic', we will be able to target multiple points in disease pathways, leading to options for patients who are now refractory to current treatments, as well as reducing potential side effects. No longer will we be limited to treating the symptoms of diseases but rather, we can attack the underlying causes and eventually provide individualised and optimised treatments and, ultimately, cures. But to fully realise this future of personalised medicine and consumer genomics, we need to start from the ground up with new tools and approaches to target discovery.

The Industrialisation of bioinformatics

There is a parallel between the 'California gold rush' and the 'genomics race' today. In both cases, pioneers were entering essentially virgin territory looking for gold. Being the first on the spot, some found gold nuggets just lying around. The early pioneers of the biotechnology era, such as Amgen, Biogen and Genetech among others, have turned some of these nuggets, like erythropoietin, interferons and tPA, into some of the biggest selling drugs today. Others have seen how the early pioneers were richly rewarded for their efforts and have followed suit to get the gold. But by now the easy lodes have largely been mined and tested and either discarded or taken into development behind someone's corporate intellectual property firewall. Those following the early pioneers need to work harder and mine deeper below the surface to find untapped opportunities.

To find these undiscovered blockbusters, new approaches must be adopted. There is an extreme shortage of trained bioinformatics specialists in the biotech and pharmaceutical industries and too much data to sift through by hand hoping to find a blockbuster. Much like Henry Ford introduced the assembly line and transformed automobiles from a hand-crafted specialty product for the rich to a consumer product, so too must bioinformatics transform itself from the realm of the specialist to a common tool of the average biologist as another instrument in their research toolbox.

To accomplish this transformation, several critical pieces must either be developed or implemented. One obvious requirement is faster computers. Massive amounts of data requires massive amounts of storage and computational speed. Those familiar with computers recognise the widely cited 'Moore's Law' which postulates that CPU transistor density, which approximately equates with CPU speed, doubles about every 18 months. That principle has helped make PCs the commodities they are today.

But biologists working with the tremendous amount of sequence information know the real truth. The amount of biological data is increasing at a faster rate than CPU speed is increasing. So even doubling your computational power every year won't keep pace with the speed with which data is being acquired. Also, biologists working with the data know the quantity is only going to increase at a dizzying pace as we look past human genomes and start trying to understand many of the important plant and animal species around us. So processing and understanding the genomic data is going to become more and more expensive as the computational requirements continue to increase.

Another requirement is more sophisticated software analysis tools. The real use of all that sequence information is to predict in silico the function of a protein. We predict the function of an unknown protein by comparison to databases of genes and proteins whose functions are already known. The process of Functional Genomics is the actual laboratory confirmation of a proposed gene function. Only by giving highly educated 'guesses' and highly accurate functional predictions to the bench can bioinformatics find its true worth.

One problem today is that given a list of 20 potential drug targets, perhaps as few as 3-5 actually have the predicted function confirmed in the laboratory. If there were ways to increase those odds from 1 in 5 to 4 in 5, much less time would be lost in expensive wet laboratory testing, and more targets per year would advance to the next stages of the process.

Research is ongoing in many academic and commercial institutions to develop more sophisticated and sensitive algorithms to predict protein function where there is very low similarity to known sequences. Categories of tools such as Hidden Markov Models (HMMs)[6], Phylogenetic Analysis, and Protein Threading[7] are the cutting age computer reagents needed to identify proteins with very remote sequence homology. Popularly referred to as 'The Twilight Zone', correct identifications of protein similarity in these nether regions will find truly novel drug targets, and convey proprietary advantage to the company or researchers who find them first.

As important as genomics and sequence analysis are for the future of the drug discovery process, they are not the only new technologies that are critical. Others include Combinatorial Chemistry[8], the process of quickly and systematically producing thousands of new chemical entities for testing, and High Capacity or High Throughput Screening, where advances in robotics and miniaturisation allows major pharmaceutical companies to test those several hundred thousand new chemical compounds a month for possible activity in disease models. These technologies generate not only huge quantities of data, but information of entirely different sorts than biological sequence information.

The importance of data integration

An obvious side-effect of all these new technologies is the wealth of data that scientists must analyse to decide whether their projects are a success. These changes include such seemingly mundane areas as data collection from experiments. Scientific research has always generated large amounts of data, but commonly much of that information has been recorded by hand in conventional paper notebooks. Automated data collection is now making its presence felt in pharmaceutical laboratories worldwide, and with it, computer systems designed to analyse and store all that data. But frequently each type of data, be it DNA sequence data, chemical informatics, results from high capacity screening, or data from clinical trials, is handled by unique, specialised, and separate software and computer systems.

More than ever, drug discovery is a multi-disciplinary science and researchers must work with many disparate data types and sources. And while much of the data has moved from the laboratory notebook to individual PCs, there is more data than ever and it is still separated and often hidden from any but the individual researcher. In some ways, this has exacerbated the problem because, in the past, you could always go back to the researcher's notebook. But now, where is that file? What is it called? Is it on the data collection computer or the researcher's computer or is it somewhere on a network drive? Which program do you need to read that file? The present system is clearly inefficient and contributes significantly to the cumbersome and protracted drug discovery process.

What is needed is an integrated, easy to use environment where all the necessary data and results are available to the researcher. Without more integrated and automated analysis systems, potentially valuable data and relationships will be overlooked because they cannot be found, cannot be correlated between different experiments or for lack of time for manual conversion and examination. Automated 'first pass' analysis systems are needed to sift through these mountains of data and present the most favourable results to the researcher for final scrutiny. By developing automated software analysis solutions to weed out poor candidates early in the process, ideally before they enter the laboratory for functional confirmation, more effort is focused on the most likely successes.

Regardless of how effective these new tools are at finding potential new drug targets, their utility is limited if the researchers do not use them regularly. Ease of use must be considered when making software for biologists. Unlike the typical computational chemist or physicist, who are comfortable spending up to eight hours a day in front of their computers planning experiments or analysing data, many biologists would rather be working in the wet lab. Unlike in physics or analytical chemistry, where small differences between data points may be significant, in biology, a 5- or 10-fold difference may result from the variability inherent in biological systems. It is the nature of biological systems to be somewhat unpredictable and frequently difficult to compartmentalise. Given this degree of variability, many biologists view computers and software as occasionally useful, but less important than the results of the electrophoretic gel they use in the laboratory. But while the results of laboratory experiments are indeed the gold standards, the proper computer tools can find those answers more quickly, and with less frustration.

The key to maximising the value of new tools is first to ensure that the results are truly useful for researchers and then to make them easy to use. The tools must be highly integrated, requiring only a few mouse clicks to go from DNA sequence information to highly clustered and sorted gene expression data to predicted 3D structural information and on to a list of relevant literature reporting other researchers work on that gene.

Furthermore, advanced graphical visualisation is required for the outputs of these analysis tools. The human mind is exquisitely evolved to discern patterns in data in visual data but can be easily overwhelmed by reams of numbers. So it is necessary to find ways to present complex numerical data in visual ways that allows the researcher to find the patterns. Finding those relationships also requires highly connected storage systems of databases that allow researchers to pose natural queries about biological problems and have the information automatically sifted and then stored in relational databases so other researchers can build upon previous work, rather than simply repeating parts of it.

The challenge

To meet the growing needs for new drugs to treat disease, a new approach must be undertaken, and soon. Only by rethinking the drug discovery process at many levels, starting with the most upstream part of the process, the Target Discovery phase, can these companies double or triple the number of successful drugs reaching consumers, and do so in 25% less time than the decade or longer it currently takes. Bioinformatics is key to the first step of that process. By validating more and better targets in silico, the entire downstream Drug Discovery process is accelerated. DDW

*Upon his arrival in the United States in 1989 **Alex Titomirov PhD** served as a visiting scientist at the Department of Microbiology at Columbia University and went on to participate in research on the developmentally regulated expression of mammalian cells for gene targeting at the Laboratory of Mammalian Genes and Development at the National Institute of Health. While in the former Soviet Union, Dr Titomirov served as Group Leader of a research team in the field of DNA transfer technology at the Institute of Molecular Biology in Moscow and was Head of Theoretical Seminars at the Laboratory of Functional Morphology of Chromasones. He has also served as a member of the Grant Committee of the Russian Academy of Sciences and an instructor at the Moscow Physical Technical Institute. Dr Titomirov is the founder and the current President, CEO and Chairman of Informax, Inc.*

References

1 Crollius, HR, Jaillon, O, Bernot, A, Dasilva, C, Bouneau, L, Fisc, C. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nature Genetics Volume 25 Number 2 Page 235 – 238 (2000).

2 Liang, F, Holt, I, Pertea, G, Karamycheva, S, Salzberg, SL, Quackenbush, J. Gene Index analysis of the human genome estimates approximately 120,000 genes. Nature Genetics Volume 25 Number 2 Page 239 – 240 (2000).

3 Ewing, B, Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. Nature Genetics Volume 25 Number 2 Page 232 – 234 (2000).

4 Complete genomes available online: http://www.ncbi.nlm.nih.gov/Entrez/Genome/main_genomes.html.

5 Pharma 2005. An Industrial Revolution in R&D. 1998. PricewaterhouseCoopers, page 1.

6 Karplus, K, Sjolander, K, Barrett, C, Cline, M, Haussler, D, Hughey, R, Holm, L, Sander, C. Predicting protein structure using hidden Markov models. Proteins 1997; Suppl 1:134-9.

7 Dunbrack, RL Jr, Dunker, K, Godzik, A. Protein structure prediction in biology and medicine. Pac Symp Biocomput 2000;(12):93-4.

8 Leach, AR, Hann, MM. The in silico world of virtual libraries. Drug Discov Today 2000 Aug;5(8):326-336.