WestminsterResearch
http://www.wmin.ac.uk/westminsterresearch

Bioinformatics: a useful tool for the molecular microbiologist?

Pamela Greenwell
Sanjiv Rughooputh
School of Biosciences

This is an electronic version of an article published in The Biomedical Scientist, 48(5), pp. 494-495, May 2004.

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners. Users are permitted to download and/or print one copy for non-commercial private study or research. Further distribution and any use of material from within this archive for profit-making enterprises or for commercial gain is strictly forbidden.

Whilst further distribution of specific materials from within this archive is forbidden, you may freely distribute the URL of WestminsterResearch. (http://www.wmin.ac.uk/westminsterresearch).

In case of abuse or copyright appearing without permission e-mail wattsn@wmin.ac.uk.

A R T I C L E

Vast amounts of data are now available on individual microbial species through the application of molecular techniques such as genomics, proteomics and metabolomics. Here, Pamela Greenwell and Sanjiv Rughooputh consider the value of the information obtained and how it can be understood.

Bioinformatics:

a useful tool for the molecular microbiologist?

Following hard on the heels of the human genome project, microbial genome versions have now begun to produce vast amounts of information on the nucleotide sequences of specific microbes. How useful is this information and how can researchers wade through the millions of base pairs of sequence data to find genes or sequences of interest for either diagnostic or therapeutic strategies? In theory, the answer lies with the new specialty of bioinformatics, which covers genomics, proteomics and metabolomics – terms that are more recognisable to many as molecular genetics and biochemistry.

Many scientists have the impression that once the genome of an organism is sequenced then everything about it is known. But is this really true? At this stage, it might be useful to look at an organism whose genome has been sequenced completely. Take *Chlamydia trachomatis* serovar D, for example.

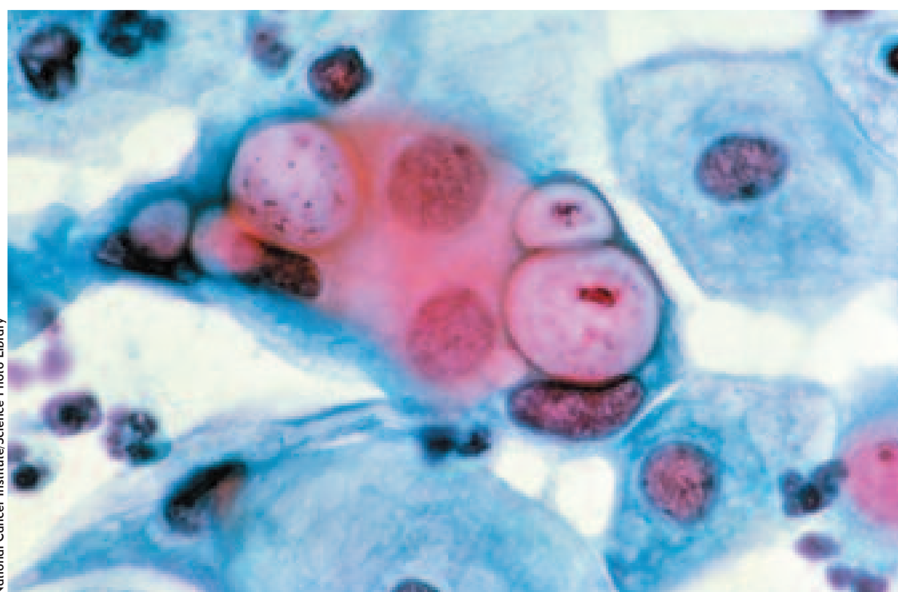
Chlamydia trachomatis

The *C. trachomatis* serovar D genome was sequenced in 1998 and found to be 1.045 Mb in length. Full details of the genes present, the proteins they encode and any references can be obtained by logging on to the TIGR website (www.TIGR.org) and then clicking on the 'Comprehensive microbial resource' and then the 'Visit a CMR page for an individual genome' icons. This will reveal the names of all the organisms cloned. In this example, *C. trachomatis* serovar D would be the destination, where everything that is known about the genome of the organism can be explored.

So, if everything is known about *C. trachomatis* serovar D, why is work still being undertaken on it? The truth is that everything is known about the organism's nucleotide sequence and researchers can predict that it codes for 877 proteins, 103 of which are secreted as judged by the presence of signal peptides and 241 have transmembrane domains and therefore appear to be membrane bound. However, the function of more than 20% of the proteins remains a mystery.

One way to identify genes is by their homology with other cloned genes, on the assumption that bacteria must have evolved from each other and therefore will share genes, albeit with some changes that have resulted from evolution. To observe such homology, an alignment search facility such as BLAST (basic local alignment search tool) can be used, which permits comparison of nucleotides or derived proteins from *C. trachomatis* serovar D with all other genes or proteins lodged in databases worldwide or with subsets (eg prokaryotic genes and proteins) of that data.

'One way to identify genes is by their homology with other cloned genes'



National Cancer Institute/Science Photo Library

The appearance of *Chlamydia trachomatis* infection in a cervical smear. The *C. trachomatis* serovar D genome was sequenced in 1998.

The results show areas of homology as alignments and give the statistical probability that the homology is significant and not due to chance. However, work tends to be done with the translated amino acids that are derived from the nucleotide sequences because different organisms do have different codon preferences that can make nucleotide comparisons difficult.

Of course, as the databases contain billions of sequences it might be expected that most proteins could be identified by their homology to other known proteins. This, however, is not the case, as has been seen with *C. trachomatis* serovar D. Is it an unusual microbe? No, *Trichomonas vaginalis* has also been studied and this has shown that most of its cloned genes encode proteins with little homology to other cloned genes.

Widening the focus

So, does this mean that bioinformatics is not useful? Although it is not possible to identify all genes by this method, some proteins can be identified and then other useful data can be derived from them. For example, when proteins encoded by *C. trachomatis* serovar D genes were compared to those of *Escherichia coli* and *Bacillus subtilis*, 195 showed better homology with *E. coli* proteins, whereas 259 showed greater homology with those of *B. subtilis*.

This implies that *C. trachomatis* serovar D has characteristics of both Gram-positive and Gram-negative organisms. When *C. trachomatis* serovar D proteins were compared to those of other sexually transmitted organisms (eg *Treponema pallidum* and *Mycoplasma genitalium*), 68 of the encoded *C. trachomatis* serovar D proteins showed homology with *M. genitalium* proteins and 286 showed homology with *T. pallidum* proteins.

However, the majority of *C. trachomatis* serovar D proteins show no similarity with those produced by either of the organisms. Nonetheless, the proteins that are conserved between these organisms may be of great interest as targets for therapeutics strategies against more than one organism simultaneously.

Genomic comparison

One of the most useful tools for the molecular microbiologist is whole genome comparison, using traditional tools such as BLAST. To achieve the best visualisation, however, Artemis (www.sanger.ac.uk/Software/Artemis/) is the tool of choice, with ACT (www.sanger.ac.uk/Software/ACT/) required for whole genome comparison.

> Comparisons of pathogenic and non-pathogenic bacteria have highlighted the presence of pathogenicity islands’

ACT permits direct comparison of two genomes and, in the case of mycobacteria, has shown that *Mycobacterium leprae* and *M. tuberculosis* originally had similar genomes, but that *M. leprae* went on to lose genes, which resulted in its current smaller size. However, small fragments of these lost genes are still visible using ACT.

In other organisms, comparisons of pathogenic and non-pathogenic bacteria have highlighted the presence of pathogenicity islands, which can be investigated as potential therapeutics targets. Similarly, the evolution of antibiotic resistance can be studied and areas of the microbial genome suitable as diagnostic targets can be determined.

In addition, nucleotide information permits the identification of specific areas of the genome and tools are available (www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) for designing primers for polymerase chain reaction (PCR)-based detection.

Leading question

So, is bioinformatics useful? Yes, but it cannot tell the whole story. It is simply a tool, albeit a very powerful one, to aid the understanding of microbial genomes.

In conclusion, it is interesting to ponder the value of the human genome project in light of the limitations found with microbial analysis. Clearly, the take-home message must be, ‘you can know everything about the nucleotides in the genome, but that does not mean that you know everything about the organism’. ■

Dr Pamela Greenwell and Dr Sanjiv Rughooputh are members of the Molecular and Medical Microbiology Research Group at the School of Biosciences, University of Westminster, London.