Genomic analysis methods

James O. McInerney & Kenneth H. Wolfe

When all the contigs have been joined and all the sequence ambiguities have been ironed out, the time comes in any genome sequencing project when it is necessary to annotate the completed sequence. Even scientists who are not directly involved in large-scale sequencing may have an interest in genome annotation and analysis, because bacterial genome sequences are now often released in a preliminary form – finished but not annotated, or even unfinished – many months before they are published. Obviously, annotation is not a trivial undertaking, but rather one that requires the use of an extensive range of bioinformatic skills. The focus of the project moves from the 'wet lab' with its gels, sequencers and PCR machines to the 'dry lab' of hardware, software and algorithms.

The sequencing project could also be said to move from the hard facts (the exact DNA sequence) to the softer inferences: attempting to identify potential open reading frames (ORFs), to assign homologies from sequence similarities, to identify motifs and fingerprints, to study codon usage patterns, to find genes that have been acquired recently through lateral transfer, and so forth. It may sound as though a lot of hand-waving is involved, but with the recent development of exquisitely sophisticated algorithms many of the heuristic elements have been removed. At first glance, a completed microbial genome sequence appears to be an unruly and undisciplined assortment of the four nucleotides. However, a careful analysis, combining a sound knowledge of microbial biochemistry with good computational assistance can provide a surprising insight into the macromolecular architecture of a completed genome.

Finding the genes

It is an easy process to identify potential protein-coding genes, but it is infinitely more difficult to identify those that are *de facto* protein-coding. We can simply look for potential start codons and stop codons. The result of this type of simplistic approach is a collection of ORFs (potential genes) of varying sizes. Rules of thumb can then be applied, for example the observation that real genes almost never overlap by more than a few bases, and that large ORFs are extremely unlikely to be spurious (the threshold chosen might be 100–200 amino acids, depending on base composition). It is advisable, however, to do database searches (see below) with the complete set of all ORFs in the genome, before any other prediction methods or length limits are applied, because the clearest indication that a gene exists is a strong match to a gene from another species.

This approach will find many genes but will leave holes in the genome where there may be ORFs that are real genes without homologues. We need robust methods for eliminating the unlikely candidates while retaining those that might reasonably be expected to encode a functional peptide. The most frequently used gene prediction method is the non-homogenous hidden Markov model (HMM) method, described by Mark Borodovsky. First, a training set of 'known' genes and 'known' non-coding regions must be supplied. These sets are then used to define frequently used and infrequently used oligonucleotides and frequency matrices are constructed. These matrices are used in a moving window analysis of the genome to find regions with oligonucleotide frequencies that correspond favourably with those in the pre-defined matrices. In this way, it is possible to predict the state of each region of DNA as coding or non-coding, independent of ORF content. HMMs are quite efficient at identifying real genes but can have difficulty deciding whether to include or exclude regions between alternative possible start codons at the 5′ ends of genes. Current HMMs are being developed with the goal of improving the reliability of recognizing the correct start codon in a gene.

Database similarity searching

One of our first ports of call when trying to make sense of a new genome is at the sequence collections. These began in the early 1980s and have never failed to double in size each year. Sequence collections come in a variety of guises. There are repositories such as GenBank or EMBL,

BELOW:
Fig. 1. Variation in base composition around the *Chlamydia trachomatis* genome. The radar plot shows the frequency of the four nucleotides at synonymous (third) codon positions, calculated as a moving average from synonymous sites within a window of 40 kb of genomic sequence. In most sequenced eubacteria the leading strand is relatively rich in G+T, and the lagging strand rich in A+C, so in *C. trachomatis* the origin and terminus of replication are probably at about 8 and 2 o'clock, respectively, in the plot.

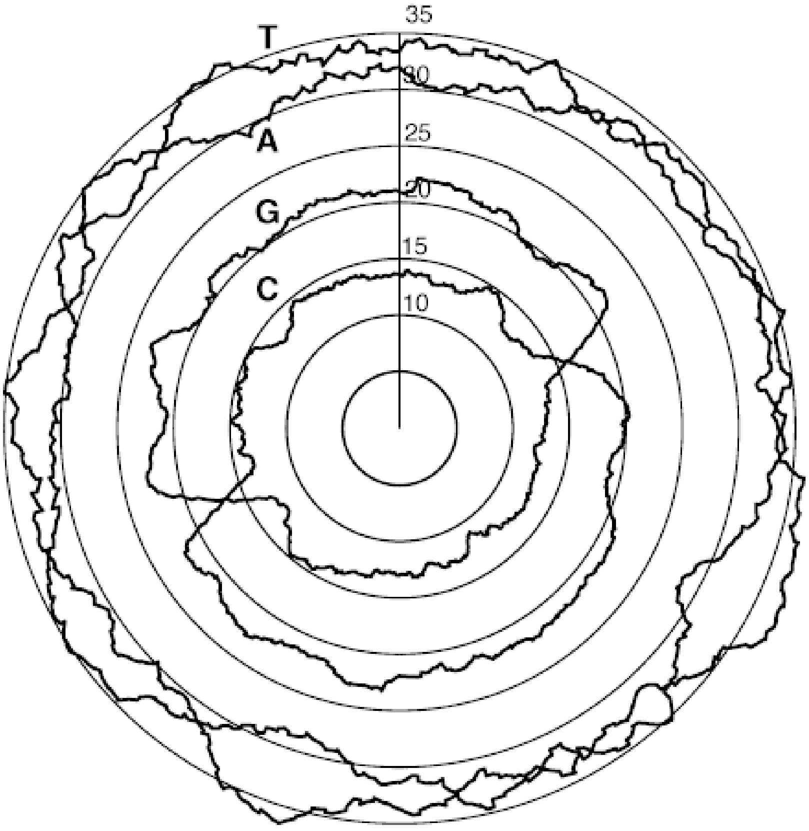

Table 1. Useful websites

GenBank	http://www.ncbi.nlm.nih.gov
BLAST	http://www.ncbi.nlm.nih.gov/blast
EMBL	http://www.ebi.ac.uk/embl
SWISS-PROT and TrEMBL	http://www.ebi.ac.uk/swissprot

which accept almost all sequences submitted to them and make them available to the public in an unaltered form, either immediately or after a short holding time requested by the sequencer. These repositories are the largest and most up-to-date sources of DNA and protein sequences. The downside to using these repositories concerns the deliberately light curatorial process and the large amount of internal redundancy (duplicate sequences or near-duplicates). In addition, sequence annotation can be quite spartan and there is no consistency across repository entries in relation to descriptive terminology and the use of keywords because these are supplied by individual authors.

A sequence database, on the other hand, is a curated, non-redundant collection of sequences, with a certain amount of consistency in nomenclature (both for the taxa and the molecules) and keyword usage. Databases such as SWISS-PROT are annotated to a high level and each entry contains a substantial amount of the available knowledge about that molecule. The downside is that there is a large backlog of sequences that have not yet been annotated and so are not in SWISS-PROT (they can be found in TrEMBL, a sort of limbo for protein sequences).

Then there are the secondary specialist databases which usually contain annotated sequences or alignments or use dedicated software products to look for 'fingerprints' in the supplier's sequence.

All these repositories and databases can be searched via the internet for sequences that show a reasonable amount of similarity to the sequence submitted by the user. These searches can be used both to suggest functions for the genes that have been predicted (e.g. BLASTP searches of proteins predicted by HMMs against SWISS-PROT) and to check for any possible genes or pseudogenes that were overlooked by the prediction algorithm (e.g. BLASTX searches of supposedly non-coding regions against SWISS-PROT).

Having found a database hit, we can then decide on the basis of similarity values whether or not the database entry is more similar to the query sequence than would be expected by random chance alone. If the similarity is significantly higher than would be expected by random chance, then we can infer that these two molecules are homologous. In the words of Walter Fitch, '*Homology is like pregnancy, two molecules are either homologous or they are not. There is no such thing as 70% homologous*'. Homology means descended from a common ancestor and the term was coined by Richard Owen in 1843, who defined it (for the purposes of morphological systematics) as "*the same organ under every variety of form and function*". In molecular terms, we define two genes as being homologous if we can infer that they arose from a common ancestor. If they show a higher degree of similarity than would be expected by chance, then it appears to be reasonable in most cases to make the leap of faith and assign both sequences to the same homologous superfamily. Molecular homology can be subdivided into orthology and paralogy: human insulin and mouse insulin are orthologues (they diverged due to species formation), while human α- and β-globin are paralogues (they diverged due to a gene duplication within a species).

The process of trawling through sequence collections looking for sequences of high similarity can be quite rewarding and usually results in many ORFs being assigned to a homologous superfamily. In many cases, it may even be possible to infer a specific function for the ORF, although this must be done with great care because of the problem of 'annotation transfer'. Many genes in GenBank are annotated as having a particular function, whereas in fact this is just an inference based on sequence similarity to some other organism where the wet-lab experiments have been done. When these 'friend of a friend' chains of inferred function become too long, the inference becomes unreliable. A related problem is that any mistake in the annotation of a gene's function in one genome sequence (e.g. a poor judgement call by an author) can become perpetuated by annotation transfer to other genomes and then becomes very difficult to clear up.

● When there are no similar sequences

In every completely sequenced genome there are some probable genes that cannot (even tentatively) be assigned a function on the basis of sequence similarity. Some of these fall into gene families whose sequences are conserved but where a function has not yet been discovered for any member of the family. Other ORFs without homologues must be incorrect predictions and not genes at all. A third group are 'orphans' – genes without families. A combination of evolutionary rate and time since separation from the most recent common ancestor may conspire to make sequences that are homologous appear to be unrelated in sequence. Alternatively, a homologue of the ORF in question may exist somewhere in nature but never have been sequenced before. The question then arises "*What, if anything, do these ORFs encode?*"

Plainly, this question represents something of a holy grail in molecular biological terms. If it were possible to assign functional information to ORFs in the absence of database similarity, then we would be able to make a substantial contribution to the understanding of the biology of every completed genome. Progress in this area

Further reading

Dujon, B. (1996). The yeast genome project: what did we learn? *Trends Genet* **12**, 263–270.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129.

Hayes, W. S. & Borodovsky, M. (1998). How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res* **8**, 1154–1171.

has been made recently by using DNA microarrays to study expression patterns, particularly in yeast. Groups of genes involved in the same biochemical pathway or physiological process tend to be co-regulated. Consequently, if an ORF's expression pattern under a wide variety of conditions groups it with several other genes, all of which are involved in some pathway, the ORF is also likely to be involved in this pathway. This approach can be used to generate hypotheses that can be tested in the lab, but is not yet at the stage where predictions can be made infallibly.

Other genomic patterns

There are other patterns that can be found in completed bacterial genomes. For instance, when the completed sequence of *Mycoplasma genitalium* became available, it was possible to find an unusual wave of base composition heterogeneity around the genome. This wave of base composition variation has a knock-on effect on codon usage within the genome. As yet, the reason for this peculiarity remains unknown. In the spirochaetes *Borrelia burgdorferi* and *Treponema pallidum* there is another unusual phenomenon. There is a significant difference in base composition between the leading and lagging strands of DNA replication and usage of most of the synonymously variable codons is significantly different between the two strands. This time there is a clear explanation for this base composition bias and it is related to replication.

Some extraordinary claims concerning lateral gene transfer have been made since complete and nearly complete genomic sequences have been made available. It has been suggested, on the basis of nucleotide composition analysis, that a very large portion of the *Escherichia coli* genome has been acquired by horizontal transfer since it shared a common ancestor with *Salmonella*. This is a difficult hypothesis to test using an independent dataset, although denser sampling of taxa, combined with phylogenetic inference does have the potential to clarify the issue. Should it prove that these figures are a reasonable estimate of the levels of lateral transfer, there are considerable implications for molecular biologists and evolutionists to reflect upon.

● *Dr James McInerney is based at the Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland and at The Natural History Museum, Cromwell Road, London SW7 5BD*
e-mail james.o.mcinerney@may.ie
● *Dr Ken Wolfe is based at the Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland*
e-mail khwolfe@tcd.ie

MicroShorts

Funding

MRC/Royal Colleges of Physicians & Pathologists Training Fellowships in Clinical Infection and Medical Microbiology

The aim of these Fellowships is to encourage promising young clinicians who are pursuing clinical training under RCP or RCPath schemes to undertake a period of research training in clinical infection and medical microbiology which involves work in the clinic as well as in the laboratory. Applications are particularly encouraged from those intending to obtain joint accreditation under the new RCP/RCPath training scheme in medical microbiology and virology and infectious diseases. The Fellowships provide the opportunity for specialized or further research training in relevant fields within the UK, leading to the submission of a doctoral thesis. There will be two awards available under the scheme in the 1999/2000 session and the closing date for the competition is **26 January**. Further details can be obtained by e-mail from **fellows@headoffice.mrc.ac.uk**

Pathological Society Fellowships

The Pathological Society sponsors fellowships to enable members of the medical and scientific professions working in the UK or Ireland in experimental and/or pathologically or microbiologically related research to travel to other institutions for periods of up to 12 months to learn new techniques of value in their research. Deadlines: **1 October** and **1 March**. Application forms and details are available from 2 Carlton House Terrace, London SW1Y 5AF (**www.pathsoc.org.uk**).

Research

Microbes in Norwich (MICRON)

A new website has been created to unite microbiological research in Norwich, UK. This brings together information on over 25 group leaders working within the University of East Anglia, the John Innes Centre, the Institute of Food Research and the Sainsbury Laboratory who have research interests in numerous aspects of microbiology. The site features an organism-based structure containing introductory information on all the microbes under investigation. The site also contains listings of microbiology seminars and meetings in Norwich and the surrounding area. The URL for MICRON is **http://www.jic.bbsrc.ac.uk/hosting/microbes/index.html.** For more information contact either Mike Merrick (**mike.merrick@bbsrc.ac.uk**) or Gavin Thomas (**gav.thomas@bbsrc.ac.uk**).

Education

American Society for Microbiology Resources

Our colleagues in the States have been busy expanding the range of resources that they produce to promote microbiology as a subject and as a career. A particularly exciting project, as part of the Microbial Literacy Collaborative, is the production of *Unseen Life on Earth*, a comprehensive new video series and television course. A four-part documentary is to be screened on national television in America this autumn and the video series will be launched in January, covering Microbial cell biology, Microbial genetics, Integrating themes, Micro-organisms and the environment and Micro-organisms and human life. Books and other resources to complement the series will also be available from the ASM next year. Different aspects of the project are aimed at specific audiences such as schools and colleges, distance learners, the general public and libraries. Check out **www.microbeworld.org** or contact the SGM External Relations Office for a leaflet (**info@socgenmicrobiol.org.uk**). Details of other ASM educational resources may be found at **www.amusa.org/edusrc/edu4c.htm**

Association for Science Education

The 2000 annual meeting of the ASE will take place 6–8 January at the University of Leeds. Thousands of science teachers, technicians and advisers from many other countries as well as the UK will attend. There is a very full programme of lectures, workshops, INSET courses and visits. Over 200 manufacturers, publishers, suppliers and organizations providing services to education will be exhibiting at the meeting. SGM and MISAC will be there on a joint stand, promoting the theme 'Building up . . . Breaking down' (microbial growth versus decomposition) and distributing posters, factsheets and other materials. SGM and the NCBE will also be launching the pack of fermentation activities for 16+ that has been developed by John Schollar and Bene Watmore with sponsorship from the Society. If you require further information or a copy of the ASE Advance Programme, please contact **Janet Hurst** or **Dariel Burdass** at Marlborough House.