EDITORIAL

PREDICTING FUNCTION

Microbial genomes are being sequenced at an impressive rate, but how can we take the next step and use this information to understand how bacterial cells work?

It's an exciting time to work with a bacterial model system — as bacteriologists are awash in a sea of genomic information that is promising to revolutionize microbiology. Using all the different 'omics' technologies — genomics, transcriptomics, proteomics and metabolomics — should allow researchers to understand how bacterial cells function and, moreover, how these functions are integrated. However, once the genome of your favourite bacterium has been sequenced, the fun has only just begun. How are functions assigned to the genes (or proteins) that seem to have interesting roles in different bacterial processes?

Last year, a group of experts in bioinformatics, microbiology and biochemistry met to brainstorm ideas on how to advance not only microbiology, but biology as a whole, by unlocking the knowledge that is contained in the vast reams of bacterial genome sequence information that are deposited in public databases. Proceedings were keenly observed by representatives from major funding bodies and a report ('*An Experimental Approach to Annotation*') of the findings of this microbiology think tank was published in January.

First, the problems that face microbiologists after the genome has been sequenced were discussed. When a genome has been completely annotated, there remain almost 40% of genes — many of which are conserved among several different species — for which no function can be predicted. Even for characterized proteins, the corresponding gene sequence has, in some cases, not been found. Plus, experimental validation of predicted functions has lagged far behind the speed of annotation. In fact, an inverse pyramid of information is present, in which annotations of huge numbers of sequenced genes are based on a relatively tiny number of functionally characterized genes.

Most microbiologists rely on annotators to assign functions to genes. However, it is wise to remember that annotations aren't gospel truth, but instead rely on the interpretation of the annotator, which in turn depends on available predictive bioinformatics programs combined with analyses of published literature.

"the really interesting thing about bacterial genomes is that we know what we don't know"

Where predicted functions of genes have been tested, the results (especially if negative) might not have been published. Unless specific funding has been allocated, annotators do not update predictions, so new data on the validation of predicted gene functions might not be incorporated into existing annotations. When incorrect annotations — and as many as 5–10% of predicted gene functions may be incorrect — are found during subsequent experimentation, it's rarely the sequence that is wrong but more often the annotation.

The group discussed what actions could be taken to improve both bioinformatics prediction programs and the information available to annotators. Starting with prokaryotic genomes is an ideal way to tackle these problems, which are even worse in sequenced eukaryotes. Sequenced prokaryotic genomes are small and plentiful, prokaryotes are often genetically and biochemically tractable, and, as Peter Karp pointed out, "the really interesting thing about bacterial genomes is that we know what we don't know".

The solution it seems is to tackle the problems of low levels of functionally characterized genes and high levels of 'conserved hypotheticals' by setting up a central database to function as a repository for bioinformatics information, including predicted gene functions and lists of enzymes for which no genes have been assigned. Bench scientists could use the database to devise simple experiments to test predictions, and apply for funding to carry out these biochemical analyses. A central database would allow all the results to be collated and used to feed back into future and ongoing annotation projects, and to improve the predictive capabilities of bioinformatics programs.

The aim would be to assign functions (rather than just predicted functions) to genes, with all the information housed in a central clearing house that is accessible to all. It would benefit all biologists by increasing the power of gene function prediction. This solution might just allow genomics to progress and open the door to understanding basic prokaryotic biology.