From code to mode for orphan genes

Meeting report

Classical genetics is qualitative: Mendel's peas are either green or yellow, either smooth or wrinkly. If one has such an observable phenotype, one can seek the gene, study its meiotic segregation and pin it down by genetic and physical mapping, and then by sequencing. Eventually, one can make the mechanistic connection between genotype and phenotype using biochemical measurements: this is how progress is usually made. But life is rather different in the post-genomic era.

With the genomes of many organisms now completely sequenced and new ones emerging almost every month, it is becoming obvious that the function of fewer than half of their open reading frames (ORFs) is known with any confidence, and science is faced with the challenge of understanding the function of all the newly discovered 'orphan' genes (i.e. those with no known relatives in the databases). In well-studied organisms, these genes have clearly escaped the classical ('function first') genetic approach and may therefore have quantitative, rather than qualitative, phenotypes; the field that is emerging to establish their role is known as functional genomics. It differs from classical genetics both in the comprehensive and integrative nature of its analytical approach, and in the fact that it does not rely on a one-to-one relationship between gene and phenotype.

We therefore need new tools for solving this 'inverse problem' of the post-genomic era, in which we have sequence but no function, and this was the theme of a workshop* held recently by the UK BBSRC.

Big is beautiful: large-scale analysis of molecular composition

The types of approach needed to effect progress in this area (Douglas Kell, University of Wales, Aberystwyth, UK) include: methods for the large-scale analysis of cellular composition in which hundreds of read-outs are obtained simultaneously; advanced genetic techniques for the large-scale production and tagging of mutant lines; methods for establishing functional equivalence between paralogues (evolutionary sisters) in the tractable model systems and in systems with economic importance; and much improved bioinformatics tools to allow us to deal with the flood of different kinds of data that will be generated by these new approaches, while ensuring their integration with the tools we have now.

A major programme in this area is the European Functional Analysis Network (EUROFAN), which is developing methods for functional genomics in yeast (Stephen Oliver, UMIST, Manchester, UK). Of the approximately 5790 protein-encoding genes revealed in the sequence of *Saccharomyces cerevisiae*, the functions of some 2600 (45%) are considered to be known; those of 990 (17%) have been ascribed by homology, and 2200 (38%) are of unknown function. The levels of the transcripts of approximately 20% are below the level of detection under steady-state conditions, although more are transcribed during 'transients' following the change in an environmental parameter. Of the 802 'new' genes studied in EUROFAN to date, 99 (12%) are essential as judged by tetrad analysis, and many others are pleiotropic. One way forward is the simultaneous analysis of the low molecular weight compounds in the cell (the 'metabolome'), which is likely to be much simpler than that of the proteome (as the known metabolic map of yeast contains approximately 564 metabolites, an order of magnitude lower than the number of gene products).

The current flood of genomic data has driven the development of technologies that permit the large-scale and parallel analysis of gene expression at a particular stage of development or in a particular physiological state. The most extreme example of such methods is the use of hybridization-array technologies ('DNA chips') to determine the complete set of genes expressed under a given set of conditions (the 'transcriptome'), usually by making fluorescently labelled cDNA from the transcripts and binding this to the complementary strands in a two-dimensional array. In the 'yeast-chip' variant devised by Affymetrix, each gene is represented by at least 19 different 25mers, with mismatches used as controls and for background subtraction (Elizabeth Winzeler, Stanford University, Stanford, CA, USA); the whole genome is on five chips, with some 65 000 pixels on each. In this system, 85% of genes produce detectable transcripts (the lower limit of 'undetectables' comes from lowering the background), while 7% (425) show cell-cycle regulation (varying in synchronized cultures by a factor of more than two), indicating for the first time the enormously dynamic biochemical complexity of the operation of the living cell. Allelic variation is one of the many other novel questions that become answerable with this wonderful technology. This is seen to occur in some 1 in 150 base pairs, allowing the production of more markers (3808) in one afternoon's experiments than in 40 years of classical genetics!

In addition to the Affymetrix approach, comparable arrays can be produced using PCR-mediated amplification of target sequences and, using commercially available equipment, these can be spotted onto aminated polypropylene at a density of 10 000 spots cm^{-2}; these arrays are even reusable (Jörg Hoheisel, Deutsches Krebsforschungszentrum, Heidelberg, Germany). Peptide nucleic acids, which lack the anionic phosphate in their backbone, may prove more suitable for the production of stable arrays, and may be synthesized directly in peptide synthesizers. At the same time the higher-density formats will make sequencing by hybridization an attractive possibility.

Similar analyses to that of the transcriptome may be carried out at the level of translation to define the proteome, in which two-dimensional gel electrophoresis is used to separate individual members of the proteome and mass spectrometry is employed in identification (Stephen Fey, Odense University, Denmark). Because proteome analysis makes no assumptions about which targets will be important, it is possible to use it to light up the best early markers of diseases of interest, which can then be developed into cheap ELISA-based assays. Thus, in assessing problems of rejection in heart-transplant patients, changes in some markers could be observed (and remedial action taken) several weeks before a clinical manifestation.

*The BBSRC *Technologies for Functional Genomics* workshop was held in Warwick, UK, 25–26 June 1998.

0167-7799/98/$ – see front matter PII: S0167-7799(98)01254-2

KO is OK

The use of antisense RNA to produce specific gene knockouts is attractive but has proved to be haphazard; only recently have the mechanistic and structural reasons for this become apparent from binding studies using oligonucleotide arrays up to 27 bases in length (Edwin Southern, Oxford University, Oxford, UK). A cocktail of antisense reagents may prove to be advantageous, owing to the cooperativity of binding.

Signature-tagged mutagenesis is a powerful technique that allows the construction of many mutant strains and their simultaneous analysis by negative selection (e.g. for virulence) *in vivo* (David Holden, Hammersmith Hospital, London, UK). Virulence determinants so identified often occur grouped as 'pathogenicity islands' with a GC content atypical for the organism, suggesting that they were acquired by horizontal gene transfer – a feature increasingly recognized with the large-scale genomic analyses now becoming possible.

With a time from lead discovery to drug launch that still exceeds ten years, the race is on to exploit the genomic data in genome-based pharmacology or 'pharmacogenomics' (David Bailey, Incyte Pharmaceuticals, Cambridge, UK), and functional-genomics-based drug discovery has already enjoyed success against the HIV proteases. Proprietary databases now contain several hundred thousand expressed sequence tags (ESTs), and the number of human genes thereby uncovered is probably 50 000–100 000. The human-targeted oligonucleotide arrays using cDNA of at least 1 kilobase require only 200 ng of mRNA and have a detection limit of <1 transcript cell^{-1}, while automated proteomics using two-dimensional gels followed by micro-high-performance liquid chromatography with tandem mass spectrometry is identifying 30 new proteins per day. With these tools, we see that at least 265 genes in tissue culture show a more-than-tenfold increase in their expression when their serum is replenished. In addition, the area of molecular diagnostics, especially for the analysis of polymorphisms, will be a major beneficiary of these technologies.

Bioinformatics: turning data into information

With the flood of data comes the requirement to turn this data into information, and for the development of more sensitive and powerful tools for doing so (Lee Beeley, SmithKline Beecham Pharmaceuticals, Harlow, UK). Standard BLAST analysis would not have uncovered the leptin gene, but more sophisticated methods showed it to be a member of the four-helix-bundle family of cytokines, whch includes prolactin. 'In database mining, the hits are, at best, as good as the query', and there is a major need for closer integration between the databases and the burgeoning literature.

The 100 Mb genome of the first multicellular eukaryote, *Caenorhabditis elegans*, is due to be completed by the end of this year and is available as it comes out through the ACeDB database, which contains sequence, physical and genetic maps, and gene- and literature-based information (Richard Durbin, Sanger Centre, Hinxton, UK). Some 18 600 genes are predicted, 30% of the sequence codes for them and the density is about 1 gene in each 5 kilobases. The automated generation of functional information is being carried out using software tools such as Pfam (http://www.sanger.ac.uk/Software/Pfam/).

The availability of increasing numbers of genome sequences allows large-scale comparisons of their organization and the solution of important and long-standing evolutionary questions (Eugene Koonin, National Center for Biotechnology Information, Bethesda, MD, USA). Modern sequence-comparison tools allow clusters of orthologous genes (COGs) to be described, which are the engine for functional annotation of new genomes. The COGs (whose members are, by definition, in at least three genomes) typically account for some 50% of microbial genes; these genes are thus of ancient origin. Generally, and particularly for DNA-repair enzymes, a considerably greater number of domain rearrangements and horizontal gene-transfer events than had previously been suspected has occurred during evolution.

A major area in which quantitative analyses of the contribution of individual genes to complex phenotypes has been performed is in the mapping of quantitative trait loci (QTLs) (Michael Kearsey, University of Birmingham, Birmingham, UK). However, in higher organisms, the resolution is often quite poor and the addition of many more markers (which might be obtained using ESTs) probably will not help, because the limitation arises from the low level of recombination observed in these organisms. Only fine-structure mapping using near-isogenic lines provides an obvious way forward.

Organisms

Developing the tools of functional genomics will require both generic and organism-specific approaches. Among the generic approaches, the analysis of 'neighbourhoods' is likely to prove rewarding (Antoine Danchin, Institut Pasteur, Paris, France). These proximities might include chromosomal position, operon structure, position in metabolism, isoelectric point, the formation of complexes, codon usage, evolutionary relatedness, coregulation and even proximity in the literature. In *Bacillus subtilis*, 87% of the genome is used for coding, but this organism diverged from *Escherichia coli* more than a million years ago, so there is little conservation of operon structure and so on. Indeed, *rpsA* in *E. coli* codes for ribosomal protein S1 but, although there is a homologous gene in *B. subtilis*, this organism has no ribosomal S1 protein at all! Like Grecian columns, genomes have their own styles, which can mask functional similarities (or differences). Two major strands will be the recognition of complex formation between proteins thought not to be functionally related (and often observed in the earlier literature as 'contaminants') and the need for metabolic biochemists in the post-genomic era.

The market for cereals dwarfs that for pharmaceuticals, so it is not surprising that the major seed companies are investing heavily in the tools for functional genomics (Keith Edwards, IACR, Long Ashton, UK). These include huge EST databases, physical and genetic maps, and transformation systems, all of which can be used to produce defined knockout lines. In crops such as maize, 90% of the genome is not transcribed, and even for that which is, only 1% of knockouts produce a readily observable phenotype. Given the enormous synteny (conservation of gene ordering) between the cereal grasses, this is likely to be true generally. The analysis of phenotype needs to become as automated as that for genotype.

Most of the first plant genome to be sequenced, that of *Arabidopsis*

thaliana, with 110 Mb (cf. rice with 450 Mb and wheat with 16 000 Mb), should have been completed by the end of 1999 (Michael Bevan, John Innes Centre, Norwich, UK). Of the 30 Mb done to date, homology searching reveals that some 50% of the genes are of unknown function; of those with identified functions, 30% are involved in primary and secondary metabolism, 14% in defence against disease, 15% in transcription, and 8% in signalling. Knockouts of some 1 in 200 genes give an obviously unusual phenotype. Armed as we already are with the excellent transposon tags largely available, the humble thale cress should become a major test bed for functional genomics tools in plants.

There are major conserved segments, too, between man and mouse (Stephen Brown, MRC Mouse Genome Centre, Harwell, UK). The large-scale production of mutant sperm and the screening of appropriate mutant progeny has allowed the phenotypic analysis of some 5000 lines to date, of which 153 (3%) have phenotypes observable in a complex battery of semiquantitative but sophisticated physiological tests. This resource should give much-improved annotation over that available from the human genome.

Conclusion

Overall, the major themes coming from this workshop were that the large-scale technologies now being developed will allow us to ask and answer the truly major questions, both long-standing and novel, that will revolutionize our understanding of the workings of living organisms; that large-scale expression analyses show that most genes are functional (and thus contribute to fitness) most of the time; that horizontal gene transfer has been widespread; that our traditionally single-gene-based phylogenetic analyses are consequently no longer sustainable; and that the recent emphasis on the study of macromolecules at the expense of metabolism will need modifying. The post-genomic era will not be dull.

Douglas B. Kell

Institute of Biological Sciences, Cledwyn Building, University of Wales, Aberystwyth, UK SY23 3DD.
(E-mail: dbk@aber.ac.uk; WWW: http://gepasi.dbs.aber.ac.uk/home.htm)

Solvent-tolerant bacteria in biocatalysis

Jan A. M. de Bont

The toxicity of fine chemicals to the producer organism is a problem in several biotechnological production processes. In several instances, an organic phase can be used to extract the toxic product from the aqueous phase during a fermentation. With the discovery of solvent-tolerant bacteria, more solvents can now be used in such two-liquid water–solvent systems. We are gaining new insights into the mechanisms of bacterial solvent tolerance, such as the active efflux of solvents from the cytoplasmic membrane and solvent-impermeable outer membranes.

In 1989, Inoue and Horikoshi[1] reported the isolation of a toluene-tolerant *Pseudomonas putida* strain that grew in a two-phase toluene–water system. This finding came as a surprise because it had been known for a long time that toluene and other solvents such as benzene or octanol were very toxic to microorganisms[2]. The initial observation was confirmed for some other strains of *P. putida*[3–5] and for other representatives of the genus *Pseudomonas*[6–8]. Furthermore, solvent tolerance has been found in strains of the Gram-positive bacteria *Bacillus*[9,10] and *Rhodococcus*[11].

It is interesting to note that it has been possible to improve the solvent tolerance of bacteria such as *Escherichia coli*. Such mutants are not able to withstand toluene but can usually grow in the presence of less-toxic solvents such as xylenes, and provide a means of studying various aspects of solvent tolerance[12]. The solvent-tolerant *Pseudomonas* species can be used in environmental biotechnology[13] as well as in biotechnological production processes in two-liquid water–solvent systems.

J. A. M. de Bont (Jan.deBont@imb.ftns.wau.nl) is at the Division of Industrial Microbiology, Department of Food Technology and Nutritional Sciences, Wageningen Agricultural University, PO Box 8129, 6700 EV Wageningen, The Netherlands.

Solvents in membranes

A relationship has been established between the toxicity of a solvent for an organism and the partitioning of a solvent to octanol from the water phase[14]. The logarithm of the octanol–water partition coefficient is termed log $P_{O/W}$, and this parameter has been taken to be an indicator of the solvent's partitioning from the aqueous medium to the membrane of organism[15].

It is generally accepted that it is the solvent's effects on the cytoplasmic membrane, where it will preferentially accumulate, that result in the destruction of the organism[2,14,16]. As a consequence of the high solvent concentrations in this compartment, the cell is no longer able to perform essential biochemical reactions and eventually loses its integrity.

Various analytical techniques[15,17] can be employed to determine the dose–response relationship of a solvent in a membrane. The different techniques reveal similar patterns, supporting the view that the actual membrane concentration of a solvent is an important parameter.

 0167-7799/98/$ – see front matter PII: S0167-7799(98)01234-7