Genome sequencing and analysis

Biology was the keynote for Dr Craig Venter (Director, The Institute for Genome Research or TIGR, Rockville, MD, USA) when he first organized the annual *Genome Sequencing and Analysis Conference*. It continued to be so for the ninth event at Hilton Head, SC, USA in September. If biology is the study of 'normal' function and pathology the study of 'abnormal' function then the aim in drug discovery is to compare the two in order to identify possible pharmacological solutions. Genomics guides those efforts by highlighting target genes for which altered expression correlates with disease.

Human sequence generation, analysis, polymorphism designation (distinguishing 'normal' and 'mutant' variants) and interpretation, which started with expressed sequence tags (ESTs) a decade ago, is projected to cover the whole genome by 2005. In the process, the technology has improved significantly, and methods are faster and cheaper (current cost is approximately 50 cents per base). Megabase projects are now being completed by TIGR, and others, in 3–4 months.

Prospects for drug discovery

What are the implications of these advances for drug discovery? The protein functions to be eliminated or enhanced by the next generation of antibiotics and antifungals are being identified. Complex human disease is being dissected into convenient, treatable entities. Drugs are being matched more accurately with appropriate patient populations (individuals) to make them more effective and to reduce side effects. The meeting could cast only sidelong glances at such issues, but it did indicate a technological and biological state-of-the-art that promises much for drug discovery.

The complete DNA sequences of more than ten bacterial strains are now established. Not only does this indicate their evolutionary relationships, but it also exposes a number of new attributes and unknown biochemical pathways for old attributes that the next generation of drugs can address. Accelerated sequencing abilities permit rapid comparison of 'new', virulent and traditional strains. Dr Fred Blattner (University of Wisconsin, Madison, WI, USA), for instance, has not only completed the sequence of the innocuous *E. coli* K12, but has compared it with its enteric pathogen cousin *E. coli* O157:H7. His team showed that the virulence determinants were clustered in 50 kb 'pathogenicity islands', which probably are horizontally transmissible.

Orphan genes

In fact, many of the 'orphan' genes (those without homology to known sequence) in both auxotrophs and the heterotrophs seem to be responsible for horizontal transfer of information between bacteria (e.g. transferring the phage or factors/vectors that confer drug resistance). If the sequence of these genes is known, the process of information transfer can be understood and a way to benefit the patient deduced. In the case of *Mycobacterium tuberculosis*, Dr Stuart Cole (Institut Pasteur, Paris, France) pointed out the importance and vulnerability of another previously identified mechanism, that of protein splicing.

Prospects for new antimicrobial agents

Bacterial and yeast sequencing emphasizes the importance of ion and organic molecule transport systems. Present antibiotics and antifungals concentrate on one family of transport molecules. Twelve major groups exist in yeast (400 proteins; Professor André Goffeau, Université de Louvain, Belgium) and bacteria (Professor Milton H. Saier, Jr, UCSD, CA, USA). The yeast study identified several transporters and transcription factors implicated in multidrug resistance (MDR). Given the coding capacity dedicated to transport systems, these studies expose what must be a vital set of functions for the microorganisms, and, thus, putative targets for antifungal drugs. Interestingly, evolutionary trees indicate something about the generation of these families by distinct gene duplication and/or acquisition events (Dr Karen Ketchum, TIGR).

The sequences also provide information about the function of unknown but related human sequences. Thus, the yeast homologue of the human gene associated with adrenoleucodystrophy was shown, by knock-out studies, to be involved in ion transport (Professor André Goffeau). A human protein known to be involved in genetic imprinting is homologous to a *Bacillus subtilis* protease (Dr Frank Kunst, Institut Pasteur). This provides a strong suggestion that a protease is involved in imprinting; the release from imprinting is involved in carcinogenesis.

Transgenic models

Once a membrane or secreted protein is identified, its murine homologue can be evaluated in a transgenic model. Dr Sid Suggs (AMGEN, Thousand Oaks, CA, USA) demonstrated how just such a candidate, osteoprotegerin (OPG), is being studied as a means to counter bone degradation, as seen in osteoporosis. Professor Lee Hood (University of Washington, Seattle, WA, USA) took this one step further with a detailed comparison of the mouse and human T-cell receptor (TCR) loci. The analysis is going to affect our vision of the relationship between autoimmunity and the TCR repertoire.

'Big' dye terminators

Dr Hood and Dr Mark Adams (TIGR) pointed out the importance of recent advances in cloning and sequencing, such as the large dye terminators, and computer technology in making rapid comparisons possible. The use of a large dye terminator – a large fluorescent group attached to the dideoxy terminator (in the sequencing reaction) – results in cleaner reactions and longer readlengths; when combined with the heat-stable

  1359-6446/97/$17.00. PII: S1359-6446(97)01119-7

topoisomerase it avoids compression artefacts and increases sequence fidelity. Dr Richard Gibbs (Baylor University, TX, USA) has been using 'big' dye terminators that he designed to triple output from 2 Mb of human DNA at the end of 1996 to almost 7 Mb in six months; he expects 15 Mb of finished sequence by April 1998. At 5,000 sequencing reactions per week, he is heading for 100 Mb of raw sequence per year.

Sequencing technology has evolved rapidly since the last meeting. Professor Bruce A. Roe (University of Oklahoma, Norman, OK, USA) gave an energetic and amusing exposition of his team's sequencing of the Cat Eye and DiGeorge syndrome critical regions on chromosome 22, and homologous murine chromosome 16, plus two bacterial genomes and an *A. nidulans* (fungus) EST library (a big hand for the large dye terminator–heat stable topoisomerase combination to accelerate sequencing across the infamous repeats). His analysis suggests that a long repeat in the DiGeorge region exists on 22q. The mouse–man comparison emphasizes the importance of this region and may help in choosing between candidate protein targets.

IT and automation

Dr Christopher Martin (Lawrence Berkeley National Laboratory, Berkeley, CA, USA) demonstrated the need to integrate computing into the project management as well as the analysis stage. The system that they incorporated during the last year first indicated a bottleneck in contig closure across their P1 phage (resolved by increasing the number of subclones randomly chosen for end sequencing). Recently their data indicated problems and cost overruns at acrylamide (gel)-related steps. This should spur their interest in a 96-capillary system.

In fact, the above mentioned 96-capillary automatic sequencing machine (and its 'clones') is likely to extend daily sequencing capacity 3–4 fold (Molecular Dynamics-Amersham have begun distributing the first machines, one of which was extensively tested and described by Dr Michael Zaro of Incyte, Palo Alto, CA, USA). Runs take only 2 h, instead of the current 8 to 9 h, and no problems with interference between lanes (which are actually separate capillaries) has been observed. It is likely that upgrades of the present machine will add permanently reusable capillaries (100 independent sequencing runs between changing at present), capacity for continuous use (especially minimizing movable parts, like the scanning system) and plate-loading possibilities in order to permit week-long gambits of 3,000 sequencing reactions between reloads.

Gene chip technology

Professor David Cox (Stanford University, CA, USA) and collaborators demonstrated the effectiveness of chip technology to measure the frequency of mRNAs corresponding with a known sequence. Dr Eric Lander (Whitehead Institute Genome Center, Cambridge, MA, USA) and his team verified the ability of genome chips to distinguish the polymorphisms essential to the study of inherited disease. There are, in fact, three ways to attach the nucleic acid to a chip. Affymetrix (Santa Clara, CA, USA) uses photolithography, Incyte's subsidiary Combion and Mosaic (Seattle, WA, USA) uses the 'inkjet' piezoelectric principle that has proven so successful in oligonucleotide synthesis, and everyone (TIGR, Whitehead and Stanford University included) uses mechanical deposition, especially of EST fragment collections. Scientists at Professor Ron Davis' laboratory at Stanford are comparing the various chips on the basis of issues such as cost, versatility and application.

Software and hardware

The recently opened floodgates of sequence are encouraging new and faster methods of data verification, homology searching and coding–noncoding sequence discrimination. Avoidance of data duplication and discrepancies is the touchstone of GDB (Genome Database). Many groups are employing the PHRED software devised by Dr Phil Green (Washington University, St Louis, MO, USA). There are hardware as well as software solutions to perform Smith–Waterman sequence homology searches, cutting time up to 15-fold (R. Mark Adams, Alpha Gene, Woburn, MA, USA). Dr Rajeev Aurora (Johns Hopkins University, Baltimore, MD, USA) exquisitely demonstrated the power of protein structure homology searches (α-helices, β-sheets, coiled coils, etc.). His software detected sequences that had 9% homology and shared three-dimensional structure, and ones with 30% homology that shared none – talk about a sieve for 'good' gene candidates if you know the functional domain that you are after.

In fact, the cutting edge analysis emphasizes the need for full length cDNA sequences. There are a number of new techniques for generating these, including some from Dr Levy Ulanovsky (Weizmann Institute, Rehovot, Israel), who is known for his innovative oligo-primed DNA-sequencing protocols. Also, there is a new set of algorithms to check complete 5′ (Dr Wyeth Wasserman, SmithKline Beecham, King of Prussia, PA, USA) and select 5′ (Dr Jean-Baptiste Dumas, Genset, Paris, France) ends of clones, including promoters. Splicing alternatives can be detected, in part using more accurate three-dimensional protein gel technology (Professor P.C. Andrews, University of Michigan, Ann Arbor, MI, USA) based on direct detection of mass (to 5 kDa; about 100 peptides) on an improved (smaller and cheaper) mass spectrophotometer in the second dimension. Finally, the US Department of Energy is organizing and supporting these efforts in all scientific domains (academic, public and private).

Full programme

There was much more discussed at Hilton Head, such as new enzyme combinations for longer PCR and sequencing reads, pyrosequencing, DENS sequencing using oligos and boronated nucleotide sequencing. Hilton Head provided a very encouraging overview of technical and intellectual innovation on the genome.

J. David Grausz
fax: +33 1 44 05 1970
e-mail: jd.grausz@chu-stlouis.fr