Reactome: clear view of a starry sky

Morag Robertson, m.robertson@elsevier.com

It is now possible to navigate easily through a variety of complex human biological reactions, thanks to Reactome (http://www.reactome.org), an online curated, free-to-access, database of human biological processes, which has recently been launched as a collaboration among Cold Spring Harbor Laboratory (http://www.cshl.org), the European Molecular Biology Laboratory's European Bioinformatics Institute (http://www.ebi.ac.uk) and The Gene Ontology Consortium (http://www.geneontology.org).

Exponential expansion

Genomic information has increased exponentially and, to manage this expansion, a series of gene-centric 'catalogues', such as RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq), GeneCards (http://bioinformatics.weizmann.ac.il/cards) and KEGG (http://www.genome.ad.jp/kegg), have been created. These databases provide a gene-by-gene view of the genome, however, researchers across medical science rarely think of a gene or a protein at a time but are instead concerned with the complex interactions among proteins, protein complexes, nucleic acids and small molecules that carry out a complex biological process. There has therefore been a mismatch between the 'gene-at-a-time' design of the current genome databases and the 'whole-pathway' approach of most scientific literature. Reactome's predecessor, Genome Knowledgebase, was created to narrow this gap.

'Although we don't know the whole picture now, there is a lot of information currently held in the minds and papers of scientists world wide', says Ewan Birney, who heads the

European Bioinformatics Institute side of the Reactome project. The challenge, therefore, has been to create a system for pulling all of this information together into one easy to use resource. With the launch of Reactome, the hope is that the researcher and student alike will be able to find all the information they need in one place.

Systems biology

The emerging field of 'systems biology' has further resulted in the Reactome database, which incorporates all the information available in its predecessor. It has two target audiences: the bench researcher (including undergraduates, graduate students and postdocs), who needs to quickly locate a gene product and the processes it participates in, and learn about the role the protein product has in the larger context of a biological pathway, and the bioinformaticist, who is trying to draw conclusions from a large dataset and can query the database or download it. 'Just as large databases, such as Genbank, have been essential for cataloguing all of our genes, databases, such as Reactome, will be essential for cataloguing all the processes that our genome encodes', says Trey Ideker, Whitehead/Pfizer Computational Biology Fellow, Whitehead Institute Center for Genome Research at MIT (http://www.wi.mit.edu/far/far_ideker_bio.html).

Reactome covers the topics of cell division and its checkpoints, DNA repair and replication, metabolism of sugars, amino acids, lipids and nucleotides, the tricarboxylic acid cycle, the central dogma of gene transcription and processing of mRNA and protein translation. In future updates, the database will further include programmed cell death, nerve-impulse transmission and blood clotting and blood cell development. 'Reactome is a knowledgebase that is converting that information into a single public resource for people now and in the future to use', says Birney.

A complication that has plagued previous attempts at creating a database of biochemical reactions is the redundancy of biological processes. Common practice is to treat functionally equivalent protein isoforms and splice variants, for example, as one entity, implying that any individual entity from the given set could fulfill the same role in a given situation. Although Reactome allows the same type of generalization, it does so in a way that enables specific functions to be traced back to the individual molecules covered by the generalization. Macromolecular complexes are treated in a similar way, with a complex being treated as a distinct entity from its individual components. For example, a molecular entity in one compartment has a different identity in Reactome from an entity with an identical chemical composition in another compartment.

The database is authored by expert biological researchers, maintained by Reactome editorial staff and cross-referenced with PubMed and the sequence databases Ensembl, LocusLink

1359-6446/04/$ – see front matter PII: S1359-6446(04)03217-9

and SwissProt. The basic information in the database is provided by expert bench researchers, edited by Reactome staff and, following peer review, is published online. Knowledge from experts in the field, cast into this unambiguous but impressive data model by human curators (as opposed to computational text processing), makes Reactome data suitable for a variety of computational modelling approaches.

NIH chemical genomics network takes off

Anthony Li, a.li@elsevier.com

The National Institutes of Health (NIH; http://www.nih.gov/) announced that it has established the first in a series of chemical genomics screening centres that will allow academic and government scientists in the USA a currently unprecedented access to large libraries of organic chemical compounds.

Ten pilot centres

The NIH Chemical Genomics Center will be based at the National Human Genome Research Institute (NHGRI; http://www.genome.gov/) with the plan of funding ten pilot centres across the USA within the 2005 fiscal year. To support the network, NIH will establish a repository to acquire, maintain and distribute a collection of up to one million chemical compounds.

The compounds in question, referred to as 'small molecules' because their size is less than that of proteins, can be used to modulate gene function and so improve understanding of the biological pathways involved in human health and disease. Most marketed drugs are small molecules and, as a result, have been the focus of intense screening efforts by the pharmaceutical industry for many years. In contrast, researchers within academia, government and non-profit research institutions have previously not been able to access large libraries of these molecules.

Redressing the balance

Director of the NHGRI, Francis S. Collins, hopes that the centres will help

to redress the balance without impinging on the pharmaceutical industry; 'It should not be viewed as an effort to turn public sector researchers into drug developers. What we are doing is simply giving academic and government researchers a chance to contribute in a much more vigorous way to the earliest stages of the drug development pipeline.'

The network of chemical genomics centres will explore the vast majority of the human genome for which no small-molecule chemical probes have been identified. Pharmaceutical research tends to concentrate on small molecules that act upon molecular targets with known relevance to human disease, which constitutes a relatively narrow group. To date, of the hundreds of thousands of proteins thought to be encoded by the 25,000 genes within the human genome, less than 500 have been discovered to possess a chemical compound with which they interact. The NIH Chemical Genomics Center plans to substantially build upon this number by screening more than 100,000 small molecules within its first year of operation.

Ultra-high throughput screening

To help achieve its goals, the centre has selected a suite of ultra-high throughput target and pathway screening technologies from Kalypsys of San Diego (http://www.kalypsys.com/). The agreement, potentially worth up to US$30 million over the four-year contract, will deliver a series of technologies, materials and services including an automated robotic system capable of screening in excess of a million compounds per day in a variety of biochemical and cellular assays.

It is hoped that the information generated by the chemical genomics network will identify new targets for therapy and the tools to study them and, as a result, enable these targets to move more rapidly through the drug development pipeline.

1359-6446/04/$ – see front matter PII: S1359-6446(04)03218-0