

response, distinct from T- and B-cell antigen-specificity. Recent examples of pathogen specificity involve the influenza virus and parainfluenza virus hemagglutinins, now known to be specific ligands of the human NK activating receptor NKp46 (ref. 15). There may be specific interactions between different activating NK receptors and non-self cell-surface-expressed molecules encoded by different groups of viruses.

It is now clear that mouse NK receptor Klra8 is critical to host defense against experimental MCMV infection. Thus, the study by Vidal *et al.* allows one to conclude, unequivocally, that NK cells have an essential role in immunity to infectious

agents. Whether NK cells defend against viruses under natural conditions of infection will be difficult to ascertain in mice.

Most humans are infected by human cytomegalovirus (HCMV); the infection is generally asymptomatic or self-healing. People with inherited T-cell deficiencies, however, are vulnerable to HCMV. In addition, unexplained severe clinical illness occurs in a small number of fetuses and newborns, and, more rarely, in children and adults. Whether human activating NK receptors confer protection from HCMV and whether symptomatic patients lack these receptors remains to be determined. □

1. Biron, C.A. *et al. Annu. Rev. Immunol.* **17**, 189–220 (1999).
2. Lee S.-H. *et al. Nature Genet.* **28**, 42–45 (2001).
3. Raulet, D. *et al. Annu. Rev. Immunol.* **19**, 291–330 (2001).
4. Tomasello, E. *et al. Immunity* **13**, 355–364 (2000).
5. Bakker, A.B.H. *et al. Immunity* **13**, 345–354 (2000).
6. Paloneva, J. *et al. Nature Genet.* **25**, 357–361 (2000).
7. Idris, A.H. *et al. Proc. Natl. Acad. Sci. USA* **96**, 6330–6335 (1999).
8. Scalzo, A.A. *et al. J. Exp. Med.* **171**, 1469–1483 (1990).
9. Scalzo, A.A. *et al. J. Immunol.* **149**, 581–589 (1992).
10. Smith, H.R.C. *et al. J. Exp. Med.* **191**, 1341–1354 (2000).
11. Karre, K. *et al. Nature* **319**, 675–678 (1986).
12. Farrell, H.E. *et al. Nature* **386**, 510–514 (1997).
13. Reyburn, H.T. *et al. Nature* **386**, 514–517 (1997).
14. Delano, M.L. & Brownstein, D.G. *J. Virol.* **69**, 5875–5877 (1995).
15. Mandelboim, O. *et al. Nature* **409**, 1055–1060 (2001).

Linking microarray data to the literature

Daniel R. Masys

University of California, San Diego, School of Medicine, La Jolla, California, USA. e-mail: dmasys@ucsd.edu

The availability in computerized form of the published literature on genes is a potentially rich source of information for the interpretation of microarray data. Automated text processing confronts substantial challenges due to variability in the language used by authors, but even incomplete linking of gene clusters to the literature can reveal functional information that is useful in explaining gene expression patterns.

Microarray technology permits the measurement of the expression of thousands of genes simultaneously, and presents some prodigious challenges to data analysis. First-generation tools for microarray analysis have focused on numerical analysis methods such as unsupervised hierarchical clustering to find groups of genes sharing similarities based solely on expression values. Statistical approaches are useful for both class discovery (for example, identifying previously unknown genes and assigning putative functions to them) and class prediction (for example, identifying gene expression profiles that correlate strongly with and optimally classify phenotypic manifestations). Interpretation of the biological basis for observed similarities in gene expression patterns is generally left to the observer, and presents a formidable challenge due to the sheer volume of different genes available on microarrays and the complexity of possible ways that genes might be related structurally and functionally.



Bob Crimi

The availability in computer-interpretable form of the published literature describing genes and their functions is a potentially rich source of information to assist in the interpretation of gene expression patterns. On page 21, Tor-Kristian Jenssen and colleagues¹ describe an analytical application built by searching the titles and abstracts of over 10 million MEDLINE citations for the presence of gene identifiers. Citations containing gene identifiers were

processed to identify the co-occurrence of more than one gene identifier, and the existence of co-occurrences was used to build a network of relationships among genes. Genes were further characterized by extracting the Medical Subject Headings (MeSH) keywords used to index the citations. Using this data, they created a web-accessible application named PubGene (<http://www.PubGene.org>) that combines similarity determination based on numerical analysis of microarrays with analysis based on literature-derived linkages.

In the construction of this new resource, the authors encountered what Samuel Johnson identified in 1775 as “the boundless chaos of a living speech”². The detection of gene names and symbols present in titles and abstracts was hampered by two characteristics of language well known in the field of text processing: synonymy and polysemy. Synonymy refers to the fact that there are many ways to refer to the same object, and polysemy means that a given word may have multiple meanings. In the searching of large citation collections



such as MEDLINE, synonymy causes reduced recall (that is, inability to find records that are truly related to the gene being sought) and polysemy causes reduced precision (that is, retrieving records that are not related to the gene, due to word sense ambiguities such as insulin being a gene, a protein, a hormone and a therapeutic agent). A more fundamental limitation is that the full text of the published biomedical literature contains far more information than is contained in the corresponding titles and abstracts of articles.

To overcome the problems of unpredictable word usage of authors, libraries and information scientists developed the notion of controlled vocabularies. Terms selected from 'preferred term lists' by professionally trained indexers (who, in the case of MEDLINE, are taught to read and apply keywords to the full text of the article *before* reading the abstract) have the potential to improve both recall and precision relative to searching of author-supplied text words. But even trained biomedical indexers reading the same article choose the same preferred keywords only 40–60% of the time³, so a central issue in language-based systems is making them sufficiently robust and tolerant of the variety of ways of representing the biological ideas in the literature.

Associating experimental microarray results with the published literature is an

example of a 'data mining' tool that uses explicit linkages between two independently constructed sources of information that contain conceptually related records. Successful data mining depends critically upon reliable sets of what computer scientists call "foreign key references"—that is, the existence of the same unique words or record identifiers in records of each of the sources to be linked. The PubGene application uses HUGO gene symbols associated with specific loci on microarrays as the common currency for linking to the literature, and displays characterizations of genes using the MeSH keywords from the literature written about those genes. At the University of California, San Diego, we have developed a similar application (<http://www.array.ucsd.edu>) for interpreting gene clusters that uses GenBank accession numbers as the common currency for linking to the literature and extends the notion of characterizing groups of genes through literature-derived keywords by placing those keywords in concept hierarchies.

Interpretive tools that are currently available for analyzing microarray results provide only a partial view of the relevant literature, as if gazing through a picket fence. As Jenssen *et al.* have pointed out, PubGene gene pairs identified by co-occurrences within titles and abstracts accounted for only 45% of gene pairs

described in the text of the records of Online Mendelian Inheritance in Man, and 51% of the pairs of proteins described in the Database of Interacting Proteins. Although this is far above a chance level of performance, it means that such approaches are useful primarily as tools of intellectual exploration and browsing, and not as comprehensive or definitive tools for global characterization of expression results. Additional limitations include the inescapable fact that expressed sequence tags and genes without associated publications do not participate in the analysis, and the more subtle bias that well-known, better-characterized genes are over-represented in the literature relative to newly discovered genes.

The usefulness of automated linkages to the literature in assisting in the interpretation of array data will improve as the literature expands and becomes increasingly available as electronic full text, and as computational tools for processing language become more powerful and robust. In the meantime, the view through a picket fence is clearly better than no view at all. □

1. Jenssen T.-K., Laegrid, A., Komoroski, J. & Hovig, E. *Nature Genet.* **28**, 21–28 (2001).
2. Johnson, S. Preface to *Dictionary of the English Language* (London, 1775).
3. Funk, M.E., Reid, C.A. & McGoogar, L.S. *Bull. Med. Library Assoc.* **71**, 176–183 (1983).
4. Masys D.R. *et al. Bioinformatics* **7**, 319–326 (2001).

Packaging paternal chromosomes with protamine

Robert E. Braun

Department of Genetics, University of Washington, Seattle, Washington 98195-7360, USA. e-mail: braun@u.washington.edu

The chromosomes of sperm cells are tightly packaged into a complex of DNA and protamines. Converting the chromatin from a nucleohistone to a nucleoprotamine structure may serve both biophysical and developmental functions. Several recent genetic studies have shown unexpected findings of the dosage requirements for the genes involved in sperm chromatin remodeling.

As spermatids undergo the terminal stages of spermatogenesis, compacting the DNA requires the replacement of the histones with a class of arginine- and cysteine-rich proteins called protamines. Although the reason for replacement of the histones with protamines is unknown, one possibility is the generation of a more hydrodynamic sperm head that speeds the transit through the female reproductive tract and

across the zona pellucida surrounding the egg. It may also be that the nucleoprotamine structure protects the genetic material in the sperm head from physical and chemical damage. Alternatively, packing of sperm chromatin may serve to reprogram the paternal genome so that the appropriate genes from the father's chromosomes are expressed in the early embryo. Whereas most mammals contain a single prota-

mine, mice and men have two. A study by Chunghee Cho, William Willis and colleagues¹ (see page 82) indicates that each is vital to paternal procreation.

Using gene targeting in mouse embryonic stem cells to investigate the function of the mouse protamine genes (*Prm1* and *Prm2*), Cho *et al.* show that mutation of either haploid-expressed gene leads to defective sperm. Unexpectedly, removal