Microarrays—the 21st century divining rod?

Artificial neural networks were used to decipher gene-expression signatures collected with DNA microarrays and to classify cancers into specific categories. Will this technology lead to better diagnostic tools and new therapeutic targets? (pages 673-679)

YUDONG D. HE & STEPHEN H. FRIEND

Primitive tribes believed that the answers to important questions could be found by examining bags of animal bones, studying mammalian entrails or searching with divining rods. Critical decisions in cancer therapeutics are to some extent still guided by fairly unsophisticated determinants, such as what organ the tumor arose from, how big it is, and what it looks like after it has been imbedded in wax and colored with different dyes. Of course we refer to these techniques as tumor staging, and the dyes have sophisticated names like hematoxylin, eosin and tetrabromfluorescein. But for the most part, physicians have had to make treatment decisions based on correlations between tumor markers and past patient outcomes, without understanding the molecular mechanisms that underlie these associations. New technologies such as expression arrays, which allow us to monitor the expression of thousands of individual genes, could overcome the limitations of old methods of diagnosis and prognosis based on a limited numbers of markers[1–5].

In this issue, Khan *et al.*[6] use gene-expression profiling to tackle a particularly frustrating set of pediatric tumors, small round blue-cell tumors (SRBCTs), which often masquerade as each other. These include several tumor subtypes, including the Ewing family tumors, neuroblastoma, rhabdomyosarcoma and non-Hodgkin lymphoma[1]. Many of these cancers can be treated with similar therapies. For example, childhood rhabdomyosarcomas and Ewing sarcomas both respond to adriamycin. However, drugs of this type cause heart damage, and while Ewing sarcomas are aggressive enough to warrant treatment with these dangerous drugs, rhabdomyosarcomas can be treated with other, safer drugs. This is one reason it is so important to be able to quickly and easily identify tumor subgroups.

In an attempt to find new ways to sub-classify SRBCTs, Khan *et al.*[6] used microarrays to monitor the expression changes of over 6000 genes in 88 samples. The authors used artificial neural networks (ANNs) to identify gene-expression signatures associated with specific subtypes of tumors. These ANN models were trained using 63 SRBCT samples (23 tumors and 40 cell lines) and made satisfactory predictions for 25 independent samples (14 tumors, 2 normal tissues and 9 cell lines).

This work presents a new application of expression profiling to a clinically relevant and practically cogent situation. Recent advances in DNA microarray technologies have made it possible to measure genome-wide gene expression for a variety of biological samples. One of the most important uses of microarrays is to study changes in gene expression that accompany changes in cell physiology, such as during development, cell-cycle progression, drug treatment or disease progression. A number of studies over the last few years have established that related phenotypes are generally reflected in related patterns of cellular transcripts, implying that the cellular state can be characterized and classified by gene-expression pattern. In addition to the classification of disease specimens[1–6], this technology has led to a broad spectrum of applications such as drug target validation[7], pathway dissection[8], discovery of gene functions[9] and annotation of the human genome[10].

Kahn *et al.*[6] report that gene-expression patterns of 96 transcripts representing 80 known genes and 13 anonymous expressed sequence tags characterize 4 distinct subtypes of SRBCTs (Fig. 1). Of the 61 transcripts specifically expressed in only one cancer type, 41 (including 7 expressed sequence tags) have not been previously reported to be associated with SRBCTs. The authors made the interesting observation that Burkitt lymphomas—a subset of non-Hodgkin lymphoma—form a distinct cluster, probably because their tissue of origin (lymphocyte) is different from the two sarcomas and the neuroblastomas.

The authors also made the remarkable finding that genes currently used in tumor diagnosis do not have the tumor-specific expression pattern. For example, Ewing sarcoma is typically diagnosed by positive staining for MIC2 antigen and negative staining for the leukocyte common antigen CD45, muscle-specific actin or myogenin. However,

Pattern				# Genes		SRBCT Discrimination
EWS	NB	RMS	BL	Total	New	
				16	14	EWS vs NB/RMS/BL
				1	0	
				15	12	NB vs EWS/RMS/BL
				0	0	
				20	13	RMS vs EWS/NB/BL
				0	0	
				10	2	BL vs EWS/NB/RMS
				12	0	
				5	0	EWS/NB vs RMS/BL
				2	0	
				7	0	EWS/RMS vs NB/BL
				0	0	
				4	0	EWS/BL vs NB/RMS
				4	0	

Fig. 1 Gene-expression patterns used to differentiate four types of SRBCT. Tumors include Ewing sarcoma (EWS), neuroblastoma (NB), rhabdomyosarcoma (RMS) and Burkitt lymphomas (BL, subset of non-Hodgkin lymphoma). Red represents relatively high expression level, whereas green represents either no expression or relatively low expression levels, in arbitrary units. Columns under '# Genes' indicate the total number of genes that had differences in expression levels between tumor types, and number of new transcripts identified by this analysis. Each group of genes has a different role in discriminating diagnostic SRBCT subtypes, as shown in the last column.

Kahn *et al.*[6] discovered that although MIC2 is a highly expressed in Ewing sarcoma samples, it cannot be used exclusively to identify this cancer because it was also expressed in several rhabdomyosarcoma samples.

Kahn *et al.*[6] demonstrate that the ANN—an information-processing algorithm modeled on the structure and behavior of neurons in the human brain—can be trained to recognize and categorize complex patterns. But how different would the results be if other supervised classification techniques were used instead, such as Bayesian classifiers and support vector machine classifiers? Despite the choice of classification algorithms, the performance of a classifier depends critically on the classifier complexity and the size of training set. By no means should the set of 96 transcripts be considered a complete list of SRBCT-related genes, given the limited number of genes represented on their microarrays and the stringent criteria used in gene selection. Nevertheless, the authors found that for the given training set and with dimension reduction to 10 principal components, the optimal classifier complexity is reached without 'overtraining' when the gene set size is on the order of 100.

We are still a long way from translating the clues provided by DNA microarrays into diagnostic tools. Although microarrays might eventually be the diagnostic method of choice, they are currently too expensive for routine clinical use. However, microarrays might initially be used to identify a subset of marker genes whose expression could then be analyzed in the commonly used antibody-based diagnostic assays or by other types of diagnostic kits. Time will tell whether microarrays are best used to confirm pathological diagnoses, or whether they will have diagnostic and prognostic value independent of current pathological classifications.

The large number of new tumor-cell markers is likely to allow physicians to eventually individually tailor cancer therapies for their patients. The benefits of microarray analyses are reciprocal in that the molecular information they provide about cancer pathogenesis should eventually lead to the design of more specific drugs. As gene-expression profiling becomes more widely used in medicine, gene sets may be used in diagnosis and prognosis of different cancer types. Patient responses to specific therapies may also be predicted by the expression pattern of a set of specific marker genes. If microarrays can eventually be manufactured at low cost, and reproducibility issues relating to sample purity and signal amplification can be resolved, expression profiling will provide high levels of specificity and sensitivity in situations where classical histo- or immunopathological approaches are unsatisfactory. Tools such as DNA microarrays and ANNs or other sophisticated classifiers are likely to go a long way to improve the divining rods currently used to determine cancer type and treatment.

1. Pizzo, P.A. Principles and practice of pediatric oncology (Lippincott-Raven, Philadelphia, 1997).
2. Golub, T.R. *et al*. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
3. Alizadeh, A.A. *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
4. Bittner, M. *et al*. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
5. Hedenfalk, I. *et al*. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med*. **344**, 539–548 (2001).
6. Khan, J. *et al*. Classification and diagnostic prediction of cancers using gene expression profiling and artificial networks. *Nature Med.* **7**, 673-679 (2001).
7. Marton, M.J. *et al*. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med*. **4**, 1293–1301 (1998).
8. Roberts, C.J. *et al*. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
9. Hughes, T.R. *et al*. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
10. Shoemaker, D.D. *et al*. Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).

Rosetta Impharmatics
Kirkland, Washington, USA
Email: yhe@rii.com or sfriend@rii.com

A dominant-negative therapy for anthrax

The oligomeric structure of anthrax toxin might be exploited as a preventative or therapeutic vaccine.

STEPHEN H. LEPPLA

Anthrax, a disease caused by the Gram-positive bacterium *Bacillus anthracis*, is rare in animals and even more rare in humans under normal circumstances. However, *B. anthracis* is high on everyone's list of potential biological warfare and bioterrorism agents because it forms stable spores that can be sprayed into the air. The infective lethal aerosol dose for humans is estimated to be 5,000 spores. The accidental release of spores into the air in Sverdlovsk, Russia in 1989 led to many deaths, and vividly demonstrated the infectivity and potential threat of the bacterium[1]. Existing anthrax vaccines, including the one currently administered to United States military personnel, induce protective immunity at least in primate models of aerosol exposure. Infections in non-immunized people are effectively treated by the timely administration of antibiotics, provided that the infecting strain has not been rendered antibiotic resistant. However, a human exposure to spores might go unrecognized for some time, allowing the spores to be taken up by macrophages and carried to local lymph nodes. There, the spores can germinate and initiate growth in the blood stream. Once the bacteria grow dense enough to secrete anthrax toxin (after 1–3 days), antibiotics would cease to be effective. For these reasons, additional therapeutic measures are needed to protect against anthrax. In a recent issue of *Science*, Sellman *et al.*[2] describe a novel therapeutic approach involving a modified toxin protein that blocks anthrax toxin action both *in vitro* and *in vivo*.

Anthrax toxin is the major virulence factor produced by *B. anthracis*. As in other toxin-dependent diseases like diphtheria and tetanus, the pathogenesis of anthrax is entirely attributable to the action of the secreted protein exotoxin. Thus, *B. anthracis* strains unable to produce toxin are avirulent and antibodies to the toxin protect against infection. Anthrax toxin consists of three proteins, with protective antigen (PA) being the central player. PA binds to cells and forms a channel through which the two alternative catalytic moieties, edema factor (EF) and lethal factor (LF) are delivered to the cytsosol[3]. EF is an adenylate cyclase that produces