Commentary

Pattern Recognition and Arrays

The Times Are A-Changing

Darryl Shibata

From the Department of Pathology, University of Southern California School of Medicine, Los Angeles, California

Not so long ago, very little was known of the estimated three billion base pairs of the human genome. Little bits were painstakingly unraveled. With relatively few genes known, entire laboratories could be devoted to the discovery of a new gene or the characterization of a few genes. Even fewer genes could be considered relevant to cancer and deserving the label of oncogene or tumor suppressor gene. Most of these genes were pounced on quickly, often with rapid publication of the usual spectrum of studies—transfections, knockouts, mutation frequencies in various tumor types, mutational spectrum, molecular epidemiology, biomarker potential, and so on. Some of these genes faded into relative obscurity whereas others, like p53, became “workhorses” or “molecules of the year.”

As long as the pace of oncogene or tumor suppressor discovery was reasonable, sufficient manpower was available to populate the ever-growing number of pages in an ever-growing number of journals. However, the Human Genome Project threatens business as usual. Soon all of the estimated 80,000 human genes will be identified and sequenced. Indeed, one can now make or purchase filters or chips covered with probes representing ~25% (20,000 genes) of all human genes. With a single hybridization, one can potentially obtain more information than a thousand graduate students of yore.

Moch et al present in this issue one of the first studies using such cDNA arrays to probe for differences between renal cell carcinoma (RCC) and normal kidney.¹ From the ~5184 genes on their filters, they noted that vimentin RNA was overexpressed in a RCC cell line. Taking this hint, they examined vimentin protein expression in primary tumors by immunohistochemistry. However, in the spirit of the times, they did not examine just any set of tumors. Continuing their previous work,² they examined a collection of 532 formalin-fixed, paraffin-embedded RCC specimens arranged in microarrays (0.6-mm circular cores with as many as 1000 tumors per block), greatly facilitating the immunohistochemistry. Vimentin was overexpressed in a subset of the tumors, confirming the au-

thors' cell line observations. Furthermore, clinical follow-up was available for 386 of the tumors. Therefore, the authors were able to retrospectively link vimentin overexpression in RCC with a significantly poorer prognosis. From hybridization to immunohistochemistry to a possibly clinically significant biomarker in perhaps weeks. Amazing!

The genetic database is relatively well organized, with definite links between a cDNA and appropriate genomic annotations. The link is usually weaker between a mRNA and its protein, as specific and sensitive antibodies must typically be developed. The authors were fortunate (or, as we shall see, in some ways unfortunate) that vimentin antibodies suitable for fixed tissues have been available for years and that they were able to quickly employ their remarkable tumor block arrays. These arrays, linked to a clinical database, greatly facilitate correlation between expression and clinical significance, and are immensely important in an efficient system geared to high throughput.

Perhaps as a tribute to the current industry of tumor analysis, at the end of this high-technology voyage the authors find that others had already preceded them. They note that one study published in 1989³ also found vimentin overexpression in RCC to be associated with a less favorable prognosis, whereas another study from 1991⁴ found no association between vimentin expression and outcome.

Is vimentin overexpression a clinically useful biomarker for RCC? Only time will tell. However, the focus on a single biomarker is in many ways a relic of a past when only a limited number of markers was available. Given the ~80,000 genes soon to be available and their potential transcriptional and posttranslational modifications, it may not be rational to focus on a single marker. It may be possible and more powerful to consider the expression of all genes. Vimentin, for example, may be neither an oncogene nor a tumor suppressor gene, but rather overexpressed along with a number of other genes when certain critical pathways are altered. Because gene products

Supported by National Institutes of Health grant CA70858.

Accepted for publication February 1, 1999.

Address reprint requests to Darryl Shibata, 1200 N. State Street, #736, Los Angeles, CA 90033. E-mail: dshibata@hsc.usc.edu.

interact, a number of genes are potentially equivalent biomarkers.

One key to future knowledge will be the identification of the pathways linking genes. Perhaps one day there will be wall charts with 80,000 genes connected by arrows, similar to glycolysis and Krebs cycle flow charts. More likely, the charts will be on the Internet and not on a wall. Another likely improvement over static box-and-arrow charts will be interactive computer models. Virtually anything can be simulated if a problem is understood. The physical sciences have taken the lead with simulations, creating complex models of complex phenomena. Descriptions of biological phenomena are still largely qualitative, although the modeling of a promoter after an analogue circuit⁵ illustrates some of the possible ways of describing dynamic biological responses with software.

Describing the expression of 80,000 genes is a daunting task. For reference, there are approximately 4000 stocks on the New York Stock Exchange, and no single easily understood value fully describes its changes. However, because expression of genes is contingent on other genes, networks or patterns are certain to emerge. Comparisons between specimens should fall into three basic patterns—the same, different, and impossible. The same and different patterns are based on normal physiology. For example, studies with a yeast array containing 6400, or virtually all, open reading frames illustrate that expression of some genes remains the same, but others exhibit coordinated up-regulation or repression as a yeast culture switches from anaerobic to aerobic growth.⁶ Similar studies of human cells under a variety of conditions should reveal the multitude of potentially acrobatic expression patterns reflecting normal physiology. For example, genes associated with proliferation and wound repair are coordinately expressed by quiescent human fibroblasts on exposure to serum.⁷ The patterns can be accumulated in a large database of normal physiology, allowing comparisons with individual tumors.

A cancer cell has much in common with a normal mitotic cell; both require expression of genes that allow survival, growth, and division. However, critical genes (and hence the software of a cancer cell) are altered, which could lead to abnormal expression—patterns which are impossible in that they are never observed in normal cells under a variety of conditions. Identification of the impossible combinations may yield insight into which genes are critical for carcinogenesis and provide hints of underlying genetic alterations.

New technology brings the need for new talents. It used to be that molecular genetic data were accurately portrayed with black and white photographs, whereas histology findings required color photographs to convey their complexities. However, color pictures are necessary to portray the complexities of large cDNA arrays; words are inadequate. For now the colorful array patterns are impressive but largely incomprehensible. Although computers are up to the task of comparing all 80,000 spots (3,199,960,00 pairwise combinations), the software remains to be written and current technology may be insufficient to distinguish accurately between the dynamic ranges critical for this type of analysis.

Only a small fraction (~1.5%) of genes was found to be significantly overexpressed or underexpressed between normal and colon cancer tissues by serial analysis of gene expression.⁸ Smaller changes may be physiologically relevant, but technically harder to distinguish. In addition, rational analysis of tumors might require extensive microdissection, as tumor expression heterogeneity is likely. Fortunately, well characterized microdissected tumor specimens are being acquired and analyzed for expression via the Cancer Genome Anatomy Project.⁹

Although genomic approaches should lead to a better understanding of the fundamental biology of cancer, experience has shown that a lack of understanding does not preclude clinical utility. Pattern recognition underlies virtually all current clinical tissue diagnosis. An image that can be crudely projected on my computer screen or digitized as 480,000 dots or pixels at 800-by-600 resolution is analyzed by eye and compared to my private database of experience (have I seen this before?) to yield a diagnosis comprehensible for a clinical plan of action. In the future, one can envision hybridization of tumor RNA to an expression array of 80,000 or more dots, analysis by a scanner, and comparison to an ever-expanding international electronic database to yield a molecular-based plan of attack. This is pattern recognition again, but with a molecular-electronic interface. The digital aspects of genomics ensure that even patterns without matches will remain available for recall instead of fading with organic entropy. The high throughput approach of Moch et al¹ represents a step toward harnessing the practical promise of genomics.

References

1. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP, Sauter G: High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol* 1999, 154:981–986
2. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP: Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998, 4:844–847
3. Donhuijsen K, Schulz S: Prognostic significance of vimentin positivity in formalin-fixed renal cell carcinomas. *Pathol Res Pract* 1989, 184: 287–291
4. Dierick AM, Praet M, Roels H, Verbeeck P, Robyns C, Oosterlinck W: Vimentin expression of renal cell carcinoma in relation to DNA content and histologic grading: a combined light microscopic, immunohistochemical and cytophotometrical analysis. *Histopathology* 1991, 18: 315–322
5. Yuh C, Boloun H, Davidson EH: Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 1998, 279:1896–1902
6. DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278: 680–686
7. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson Jr J, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO: The transcriptional program in the response of human fibroblasts to serum. *Science* 1999, 283:83–87
8. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein M, Kinzler KW: Gene expression profiles in normal and cancer cells. *Science* 1997, 276:1268–1272
9. www.ncbi.nlm.nih.gov/ncicgap/