Journal of Cellular Biochemistry Supplement 37:110–119 (2001)

Genomics Approaches to Drug Discovery

John F. Reidhaar-Olson,* Brian K. Rhees, and Juergen Hammer

Department of Genomic and Information Sciences, Hoffmann-La Roche Inc., Nutley, New Jersey

Abstract New approaches to drug discovery have come about in recent years as a result of important advances in genomics and bioinformatics. The availability of genome-scale sequence data, the development of new tools for high-throughput gene expression monitoring, and improvements in the ability to analyze large data sets have revolutionized the field. In this article, we discuss three applications of genomics data in the drug discovery process: target discovery, prodrug strategies, and vaccine development. J. Cell. Biochem. Suppl 37: 110–119, 2001. © 2002 Wiley-Liss, Inc.

Key words: genomics; bioinformatics; gene expression; target identification; epitope

Recent advances in genomics and bioinformatics have transformed biology from a science of small-scale experiments to one of high-throughput processes, large data sets and sophisticated analytical methods. As a consequence, new possibilities have arisen for the identification of therapeutic targets and the development of small molecule drugs. These changes have been driven largely by genome sequencing projects and new technologies for high-throughput measurements of gene expression. The sequencing projects, particularly the Human Genome Project [International Human Genome Sequencing Consortium, 2001; Venter, 2001], have put into the hands of researchers the complete sequences of tens of thousands of genes, along with the challenge of sorting out which represent opportunities for therapeutic intervention. To help meet this challenge, gene expression technologies, such as those based on microarrays [Lockhart et al., 1996; Brown and Botstein, 1999], have provided a powerful means of using genome sequence data for the identification of disease-associated genes.

The power of array-based methods lies in their ability to rapidly dissect the transcriptional differences between normal and diseased cells. This capability has important implications throughout the drug discovery process. At the early stages, expression profiling of normal and diseased tissues contributes to target identification. At later stages, expression data can be used to optimize lead compounds and to evaluate toxic effects [Gore et al., 2000]. In addition, expression profiling can lead to the discovery of diagnostic, prognostic, and surrogate markers [Golub et al., 1999]. In this article, we will provide an overview of the genomic and bioinformatic processes involved in large-scale gene expression studies, and focus on three specific applications of these methods, each with a different approach to the development of novel therapeutic treatments (Table I).

Genomics Process

Successful use of expression profiling in drug discovery depends on the incorporation of a set of wet-lab and analytical steps into an integrated process, such as the one outlined in Figure 1. The process starts with the disease of interest and collection of relevant tissues. The number of tissues required for such a study can range from several dozen to several hundred, and must include a sufficient number of each tissue type to allow for meaningful comparisons. In many cases, the situation will be more complex than a simple division of tissues into normal and diseased categories. For example, in a study of diabetes, tissues may be collected from normal, insulin-resistant, and diabetic individuals, with samples collected from each individual under multiple experimental conditions, such as before and after insulin treatment. Moreover, in addition to these tissues,

*Correspondence to: John F. Reidhaar-Olson, Department of Genomic and Information Sciences, Hoffmann-La Roche Inc., 340 Kingsland St. Nutley, NJ 07110-1199.
E-mail: john.reidhaar-olson@roche.com

Received 28 September 2001; Accepted 4 October 2001

DOI 10.1002/jcb.10072

TABLE I. Selected Applications of Expression Profiling to Drug Discovery

Application	Basic elements of approach	Characteristics of identified gene target
Target identification	Mining of expression data and clinical parameters for identification of tractable drug targets	Causative role in disease (Targetable function)
Prodrug strategies	Analysis of expression data to identify enzymatic activities selectively associated with diseased cells	Selective expression in disease (Exploitable function)
Vaccine development	Use of expression data and epitope scanning algorithms to identify disease-specific T-cell epitopes	Selective expression in disease (No functional requirement)

there also must be access to a wide range of samples from other organs, in order to assess the overall tissue distribution of expression for genes of interest. For all tissue samples collected, clinical parameters must be collected and stored in a tissue database for later mining with the expression data, as discussed below.

In the next step, tissue samples are subjected to expression profiling using microarrays. With current profiling methods using oligonucleotide arrays or spotted cDNA arrays, this step can provide expression data for most of the genes in the human genome. Analysis of this data, using an integrated set of databases and analytical tools (Fig. 2), leads to a preliminary set of disease-associated genes. When sufficiently large numbers of tissues have been profiled, clustering algorithms [Eisen et al., 1998] can be applied to group tissue samples based on their gene expression patterns. In many cases, most of the samples will cluster clearly into normal or disease branches of the dendrogram. Outliers may represent suspect tissues, such as very early-stage tumor tissue or presumed normal tissue that is in fact contaminated with tumor cells. The analysis of expression patterns may be improved by removing such tissues and concentrating on those that are more representative of normal or diseased.

In most cases, analysis of expression data must be followed up for interesting genes with additional experimental studies for confirmation. The most sensitive confirmation method uses quantitative RT-PCR [Heid et al., 1996]. This technique has the advantage that it can be rapidly applied not only to the original set of tissues, but also to a much broader range of tissues and tissue types. This process, which we

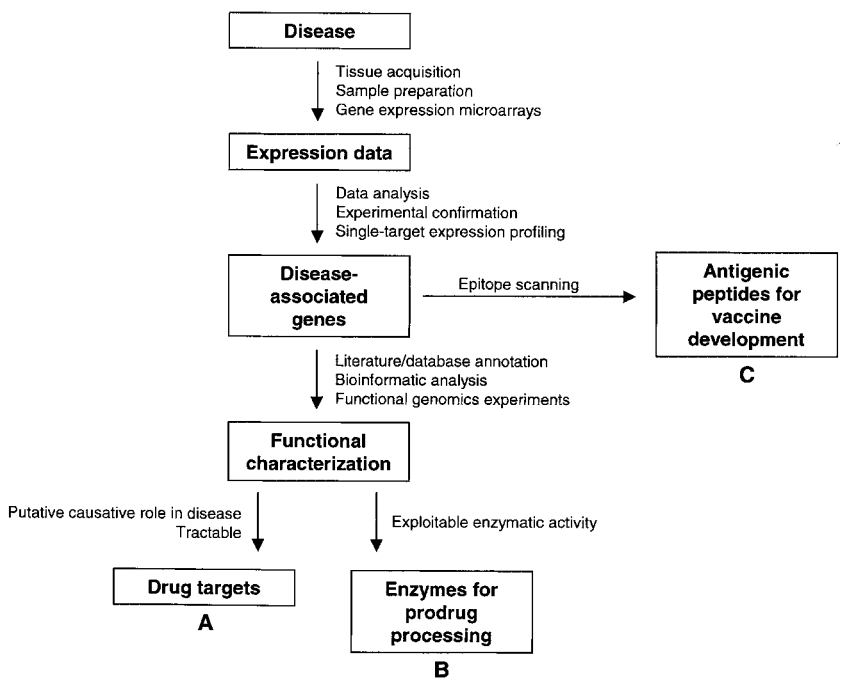

Fig. 1. Workflow for drug discovery projects based on gene expression profiling. Three specific applications of expression data are shown: (**A**) target identification; (**B**) identification of enzymes for prodrug processing; and (**C**) identification of T-cell epitopes for vaccine development. See text for details.

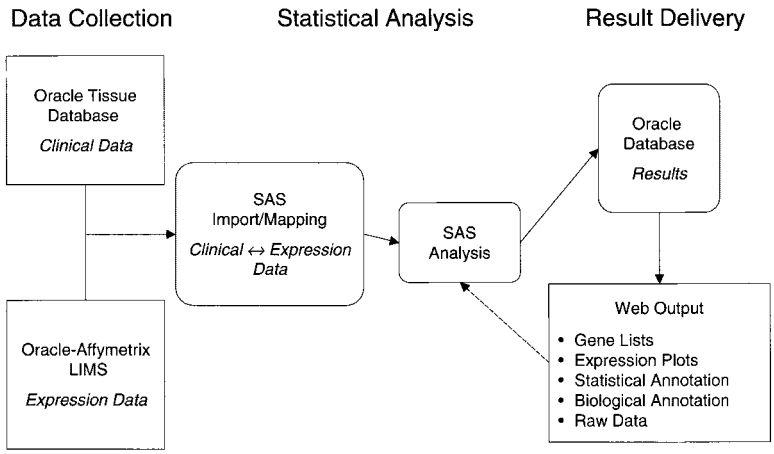

Fig. 2. Process for analysis of gene expression data. Software tools and information (in italics) are indicated for each step. The dotted arrow indicates a potentially ititerative step. The software tools shown are examples; others with the same functionality can be substituted.

refer to as single-target expression profiling (STEP), not only serves as a confirmation step for the array data, but also reveals the broader tissue distribution of expression. This information is crucial for all of the applications discussed here, since detection of high expression in the diseased tissue and low expression elsewhere is the underlying goal.

At this point in the process, different avenues can be pursued depending on the goal. We will focus here on three ways in which a set of disease-associated genes can be used to develop novel therapeutics. In the first application, the goal is identification of new protein targets with causative roles in disease. In the second, the objective is discovery of enzymatic activities

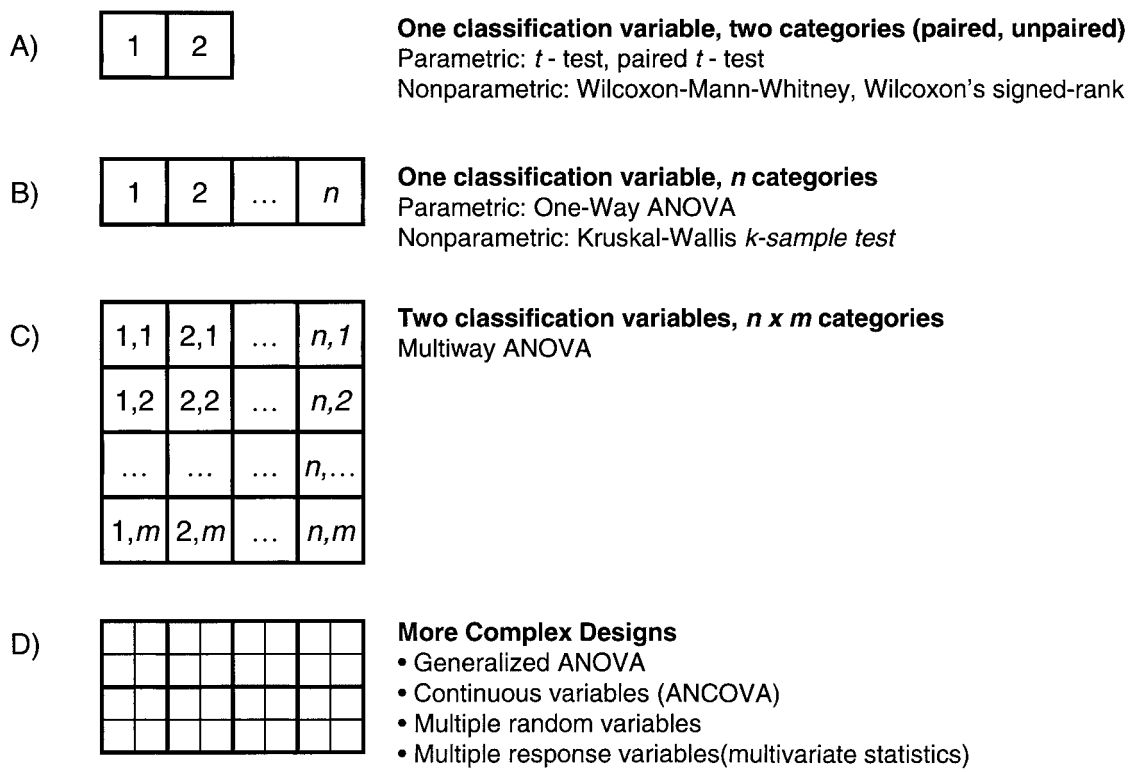

Fig. 3. Statistical models used in the analysis of gene expression data.

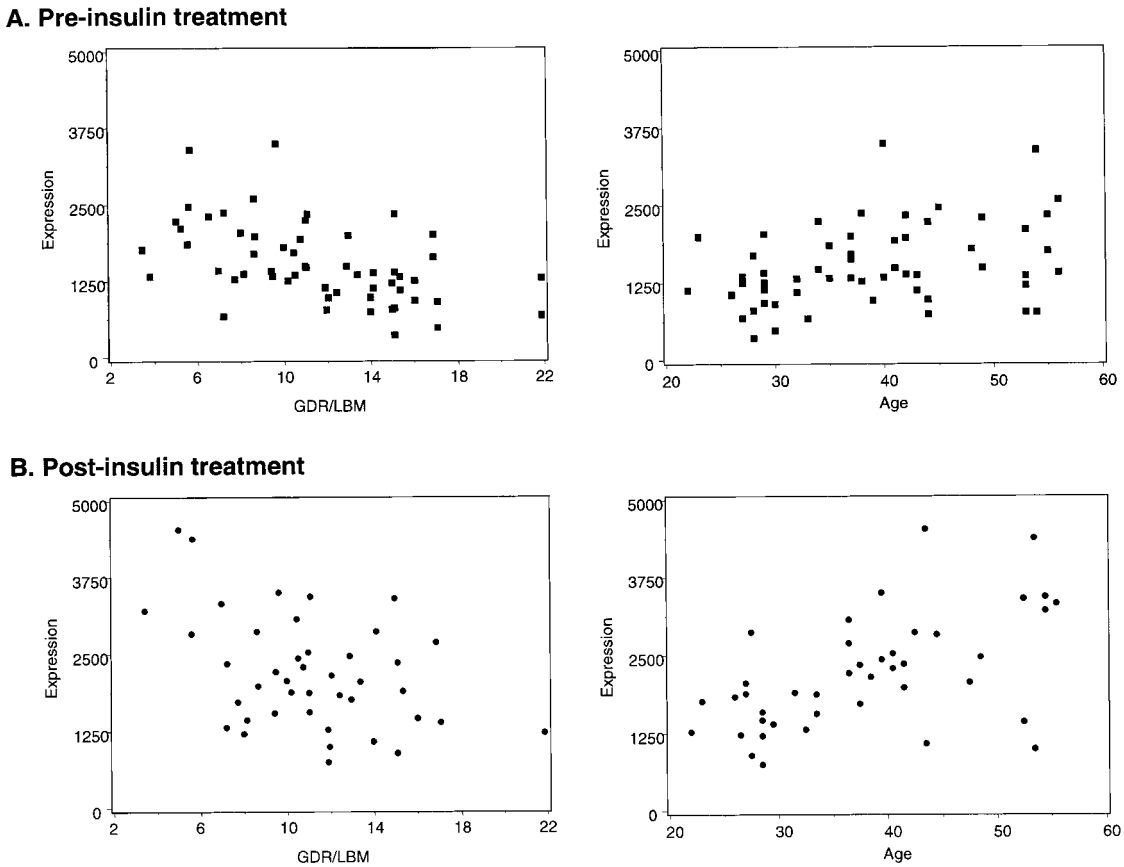

Fig. 4. Correlation of expression data with clinical parameters. The expression levels of metallothionein 1F in skeletal muscle samples (**A**) before and (**B**) after insulin treatment were determined using Affymetrix GeneChip® arrays. A clear correlation ($P < 0.001$) is observed between glucose disposal rate/lean body mass (GDR/LBM) (**left panels**); however, after accounting for variation in clinical and demographic parameters, the GDR/LBM effect is no longer significant ($P < 0.28$). In this case, the loss of significance comes from accounting first for variation in patient age (**right panels**). In both cases, expression data were analyzed by ANCOVA by fitting a model that included GDR/LBM, insulin treatment, and their interactions as fixed effects. The full analysis also included patient age, sex, ethnicity, body mass index, percent body fat, waist-to-hip ratio, and triglyceride levels as fixed effects or covariates, as appropriate. Models were fit using SAS procedure GLM [SAS Institute, 1988].

selectively expressed in diseased tissue that can serve as the basis for prodrug development. The third application identifies proteins expressed specifically in diseased tissue, with or without known function and with or without causative roles in the disorder, as a step toward vaccine development.

Target Identification

The appropriate analytical approach for identification of potential drug targets depends on the nature of the disease and the study design. In the simplest case, individual tissue samples can be considered as either normal or diseased (given their respective origins), without consideration of additional clinical parameters. In such a case, the analysis can be relatively straightforward, involving a comparison of the set of diseased tissues with the normal samples to identify genes that show significant differences in expression level (using, e.g., a t-test or paired t-test for parametric analysis, or a Wilcoxon's signed-rank for nonparametric analysis; see Fig. 3). In most cases, the emphasis will be on genes that are upregulated in diseased tissues, since drug development efforts can then focus on inhibiting the activity of the corresponding protein.

In other cases, however, the analysis can be considerably more complicated, requiring the use of more sophisticated statistical methods, such as analysis of variance (ANOVA) and analysis of covariance (ANCOVA), to incorporate clinical variables as well as expression data into the analysis. A diabetes study provides an example of this type of situation. Diabetes is

TABLE II. Genes with Significant Correlations to Demographic Variables

Effect	Affymetrix ID	Chromosome	Gene Symbol	Title	Refseq ID
Sex	34477_at	Yq11	UTY	Ubiquitously transcribed tetratricopeptide repeat gene	NM_007125
	37583_at	Yq11	SMCY	SMC (mouse) homolog	NM_004653
	38355_at	Yq11	DBY	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide	NM_004660
	41214_at	Yp11.3	RPS4Y	Ribosomal protein S4, Y-linked	NM_001008
	34842_at	15q12	SNRPN	Small nuclear ribonucleoprotein polypeptide	NM_003097
	38446_at	X	NR1I3	Nuclear receptor subfamily 1, group I, member 3	NM_005122
	45324_at	X	NR1I3	Nuclear receptor subfamily 1, group I, member 3	NM_005122
	47940_at	16q13	MT1E	Metallothionein 1E (functional)	
Age	31622_f_at	16q13	MT1F	Metallothionein 1F (functional)	
	31791_at	3q27-q29	TP63	Tumor protein 63 kDa	NM_003722
	31794_at	10cen-q26.11	NT5B	5′-nucleotidase (purine), cytosolic type	NM_012229
	37592_at	5q13.3	CKMT2	Creatine kinase, mitochondrial 2	NM_001825
	34811_at		ATP5G3	ATP synthase	NM_001689
	37027_at	11q12-q13	AHNAK	AHNAK nucleoprotein (desmoyokin)	
Ethnicity	34499_at	11q13-q14	ACTN3	Actinin, alpha 3	NM_001104
	36736_f_at	7p21-p15	PSPH	Phosphoserine phosphatase	NM_004577
	37208_at	7q11.2	PSPHL	Phosphoserine phosphatase-like	NM_003832
	37209_g_at	7q11.2	PSPHL	Phosphoserine phosphatase-like	NM_003832
	36587_at	19pter-q12	EEF2	Eukaryotic translation elongation factor 2	NM_001961
	36595_s_at 38833_at	15q11.2	GATM	Glycine amidinotransferase	NM_001482
	1728_at	10p13	BMI1	Murine leukemia viral (bmi-1) oncogene homolog	NM_005180
	674_g_at	14q24	MTHFD1	Methylenetetrahydrofolate dehydrogenase	NM_005956
	45546_at		R33729_1	Hypothetical protein R33729_1	
	55022_at		SES2	Sestrin 2	NM_031459

not a condition that can be simplified reasonably to a two-state model; rather, it involves a complex interplay of many continuous variables. The relationship of changes in gene expression to changes in any of those clinical variables is complicated to determine, and misleading results can be obtained if the data is not analyzed properly. For example, Figure 4 shows the results of an expression profiling study we have performed using skeletal muscle from individuals exhibiting clinical measurements, ranging throughout the spectrum from normal to insulin-resistant to diabetic. Samples were obtained from each individual both before and after insulin treatment using a euglycemic/hyperinsulinemic clamp. The graphs on the left show the results of a simple analysis, designed to detect genes showing significant correlation between expression level and glucose disposal rate adjusted for lean body mass (GDR/LBM), a variable that reflects insulin sensitivity. In this example, we show the results for metallothionein 1F, a gene that would likely have been found interesting under the simple model, since it shows a clear, negative correlation between GDR/LBM and expression level. However, the results are no longer statistically significant once variation in clinical and demographic parameters are taken into account. As shown in the right panels, the loss of significance can be attributed to variation in patient age. Similar analysis identifies other genes correlated with demographic variables (Table II). Analyzing the data using all the available patient information leads to a more relevant set of disease-associated genes.

Ultimately, the goal of target identification is a set of genes that are not just associated with the disease but that play a causative role. Consequently, functional information is almost always crucial. In some cases, there may be sufficient literature annotation for a gene of interest that no further functional studies are warranted. However, in many cases, the function of a gene will be unknown, or its role in the disease poorly understood. In these instances, a variety of experimental and analytical tools can be applied. Bioinformatic analysis of the gene

TABLE III. Known Genes Deregulated in Colon Cancer

Pathways	Enzyme classes	Protein functions
17 Differentiation/proliferation	8 Kinases	32 Organellar structure
15 Protein cleavage/degradation	4 Metalloproteases	29 Extracellular matrix
14 Transcriptional regulation	2 Phosphatases	16 Integral membrane proteins
11 Immunity/inflammation	3 Serine proteases	14 Transcription factors
10 Phosphorylation/dephosphorylation	2 Ubiquitin ligases	7 Cytoskeleton
6 Cell cycle	2 Carboxylic ester hydrolases	6 Adhesion
4 DNA replication	13 Other enzymes (1 each)	5 Peripheral membrane
4 Apoptosis		6 Non-receptor kinases
4 Secretion and trafficking		4 Growth factors
3 Lipid modification		4 Metalloproteases
3 Mitosis and meiosis		4 DNA synthesis/modification
3 Wound healing		3 Serine proteases
14 Other pathways (< 3 each)		3 Protease inhibitors
		3 Chemokines/cytokines
		3 G-protein signalling
		17 Other functions (< 3 each)

sequence provides clues to function, membership in a structural class, and cellular localization. Protein-protein interaction screens using two-hybrid methods [Uetz et al., 2000] place unknown genes into functional pathways. Assays based on down-regulating genes of interest reveal phenotypic effects of inhibiting activity. Many of these types of experiments are amenable to automated methods and thus can be simultaneously applied to a number of prioritized candidate target genes.

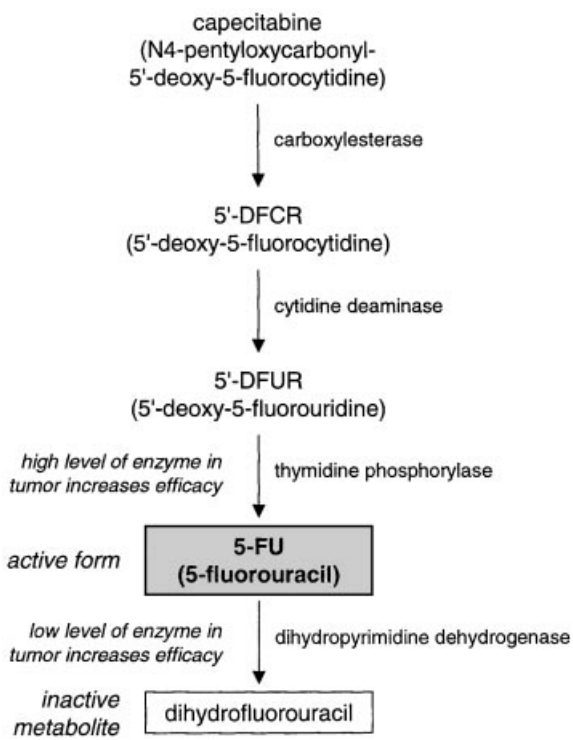

Fig. 5. Enzymatic conversion of Xeloda to active (cytotoxic) and inactive (non-toxic) metabolites. Many tumors show high levels of thymidine phosphorylase activity and low levels of dihydropyrimidine dehydrogenase activity; this combination leads to higher concentrations of 5-FU in tumor tissue than in normal tissue.

Prodrug Development

Another application of expression data is identification of enzymatic activities that are up-regulated in diseased tissues and that can be recruited for activation of prodrug molecules. A typical expression profiling experiment will identify a large number of known genes whose expression is altered in the disease. Table 3 lists pathways, enzyme classes, and protein functions of 145 known genes identified in a colon cancer tissue profiling study at Roche. As part of

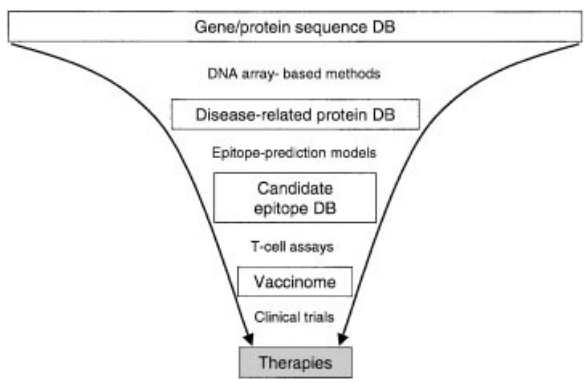

Fig. 6. Epitope prediction models can serve as a filtering tool to create disease-specific candidate epitope databases. This scheme shows how such a model can be applied to reduce significantly the size of the peptide repertoire to be tested for T-cell recognition, thus having an important impact on the amount of laboratory work leading to vaccine design and drug discovery.

a prodrug development strategy, such data can be used to identify enzymes with appropriate activities. An example of such a prodrug is capecitabine (Xeloda®), currently used in treatment of metastatic breast and colorectal cancer [Blum, 2001]. Capecitabine itself is inactive; however, enzymatic conversion results in the formation of 5-fluorouracil, an active cytotoxic agent (Fig. 5). The enzyme that catalyzes the final step in the conversion to the active form, thymidine phosphorylase, is significantly more active in tumor cells than in normal cells. In addition, dihydropyrimidine dehydrogenase, which converts 5-fluorouracil to an inactive metabolite, is less active in many tumors [Ishikawa et al., 1998]. Thus the prodrug is preferentially activated in tumor tissue, leading to greater anti-tumor efficacy and lower general

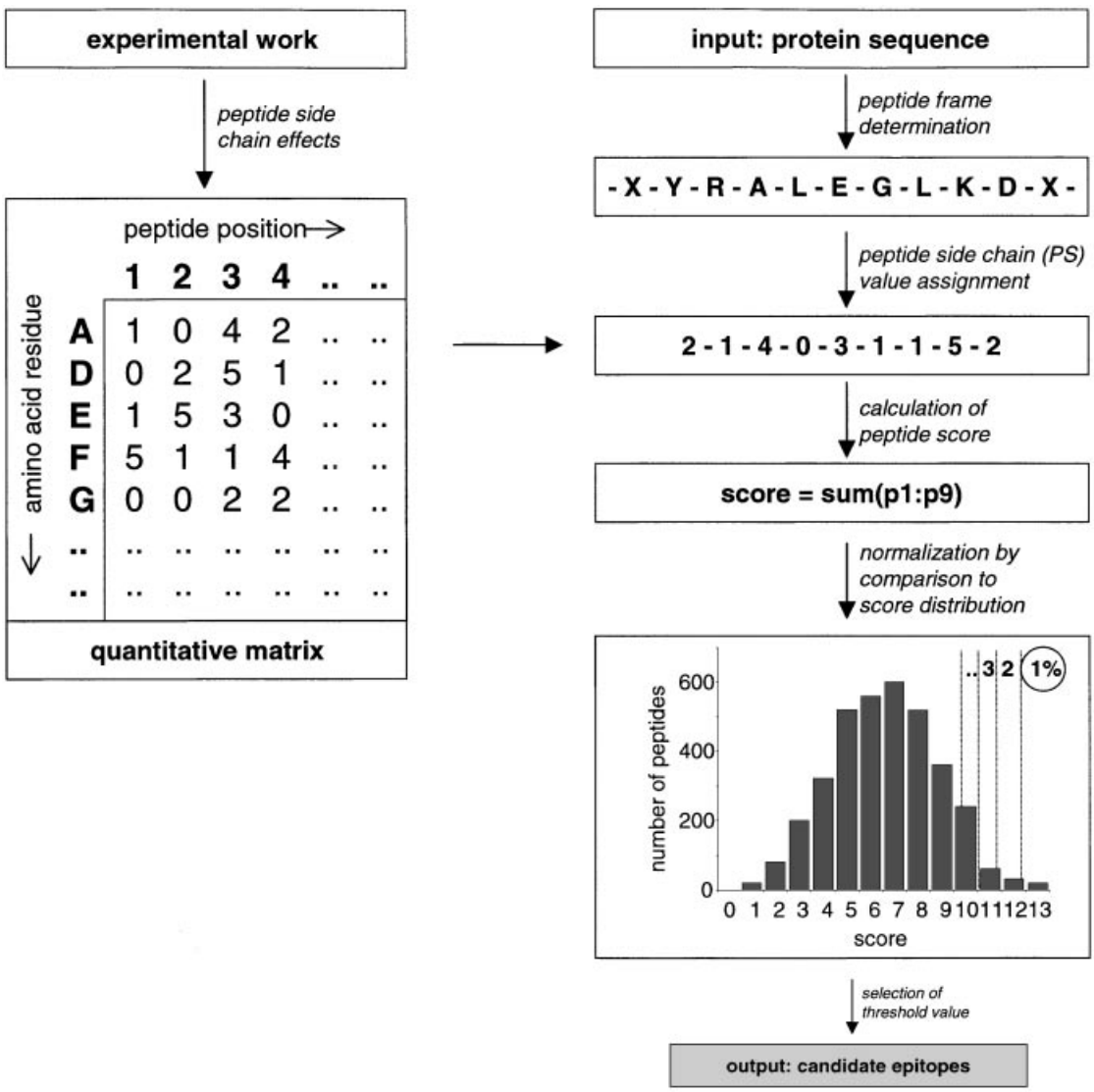

Fig. 7. The quantitative matrix-based epitope prediction process. The flow chart illustrates the basic steps required to create (left) and utilize (right) quantitative matrices for linear epitope prediction, starting from the experimental measurement of all amino acid side chain effects at all peptide positions of HLA-II binding peptides, and resulting in a score value assigned to each peptide frame (PF) contained in the analyzed protein sequence. This score is then compared to the score distribution, calculated for each matrix, of all the PF contained in a representative natural peptide sequence database. If the score of the analyzed PF is equal to or higher than the score values belonging to the percentage threshold value initially selected by the user (1% in this diagram), the output of the algorithm is the prediction of peptides containing such PF as candidate HLA-II epitopes.

toxicity. By analyzing expression profiling data with such applications in mind, other prodrug approaches will be revealed.

Vaccine Development

The final application of expression data involves identification of potential peptide sequences that can be used in the development of epitope-based vaccines. The availability of genomic-scale sequence information, coupled with genome-wide expression monitoring tools, has dramatically increased the number of possible disease-specific antigens. Because of the scale of sequence data involved, experimental approaches to identifying T-cell epitopes within these antigens, such as the synthesis and assay of overlapping peptides from proteins of interest, is not feasible. However, computer models capable of simulating and predicting the biological process of antigen presentation can be used to minimize the number sequences of interest. With a greatly reduced number of experiments, a systematic scanning for candidate T-cell epitopes is possible (Fig. 6).

Advances in understanding antigen presentation at the molecular level have accelerated the development of computer models capable of predicting T-cell epitopes [Hammer et al., 1997]. Although some of the molecular aspects of antigen presentation are insufficiently defined to be of value for epitope prediction models, others, such as the interaction of peptide fragments with HLA models, have been characterized thoroughly. As a consequence, most of the

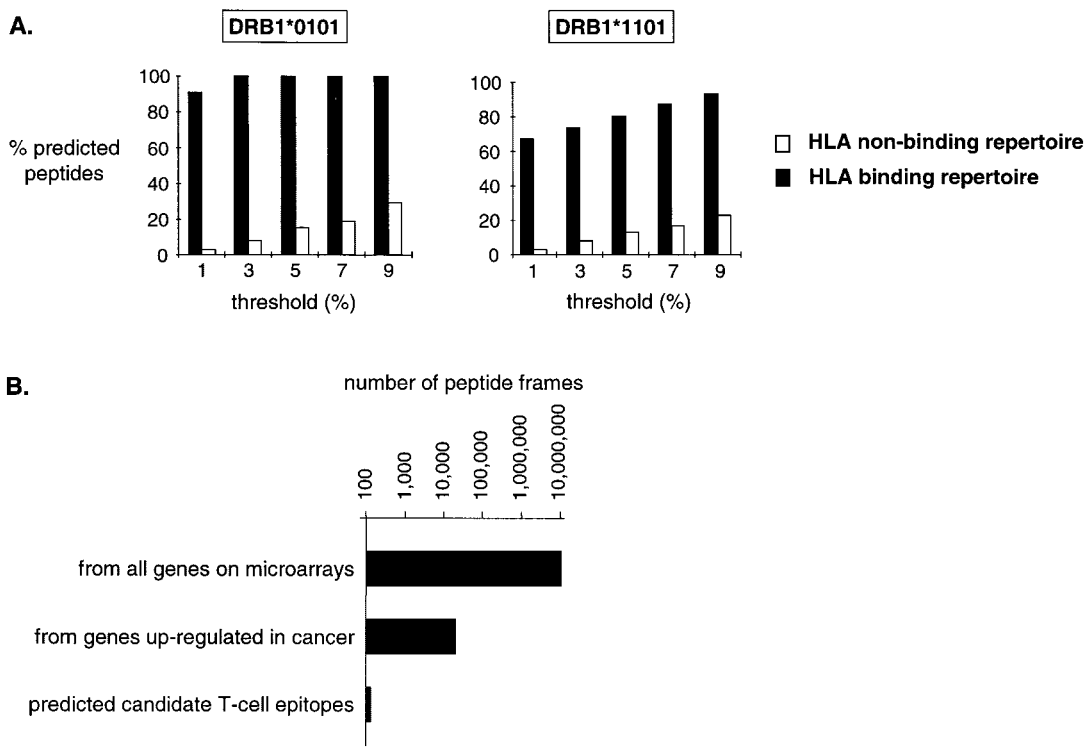

Fig. 8. Predictive power of HLA-DR virtual matrices. **A**: The binding of hundreds of randomly selected natural peptide sequences was experimentally tested, to generate a repertoire of HLA-DR binding and non-binding peptides. Analysis by the matrix-based prediction model of this repertoire showed that, at stringent threshold levels, most of the binding peptides were indeed predicted as HLA-DR ligands (black bars), while only a low percentage of the non-binders (white bars) was predicted. The data for two HLA-DR alleles are shown. **B**: In this example, microarrays were used to identify gene transcripts up-regulated in tissue samples derived from colon cancer patients. The initial number of protein sequences to be screened for epitope identification was thus reduced from ~19,000 (number of gene sequences and contigs represented on the DNA microarrays used for this analysis) to 34 gene products found up-regulated in at least 50% of the primary colon cancer tissues analyzed (middle bar). Analysis of this set of protein sequences using the HLA-II virtual matrix-based algorithm TEPITOPE identified 130 promiscuous candidate T cell epitope sequences (bottom bar), which represent a manageable amount of data for subsequent laboratory testing. The top and middle bars of the histogram show an estimate of the number of peptides corresponding to the gene/protein sequences analyzed.

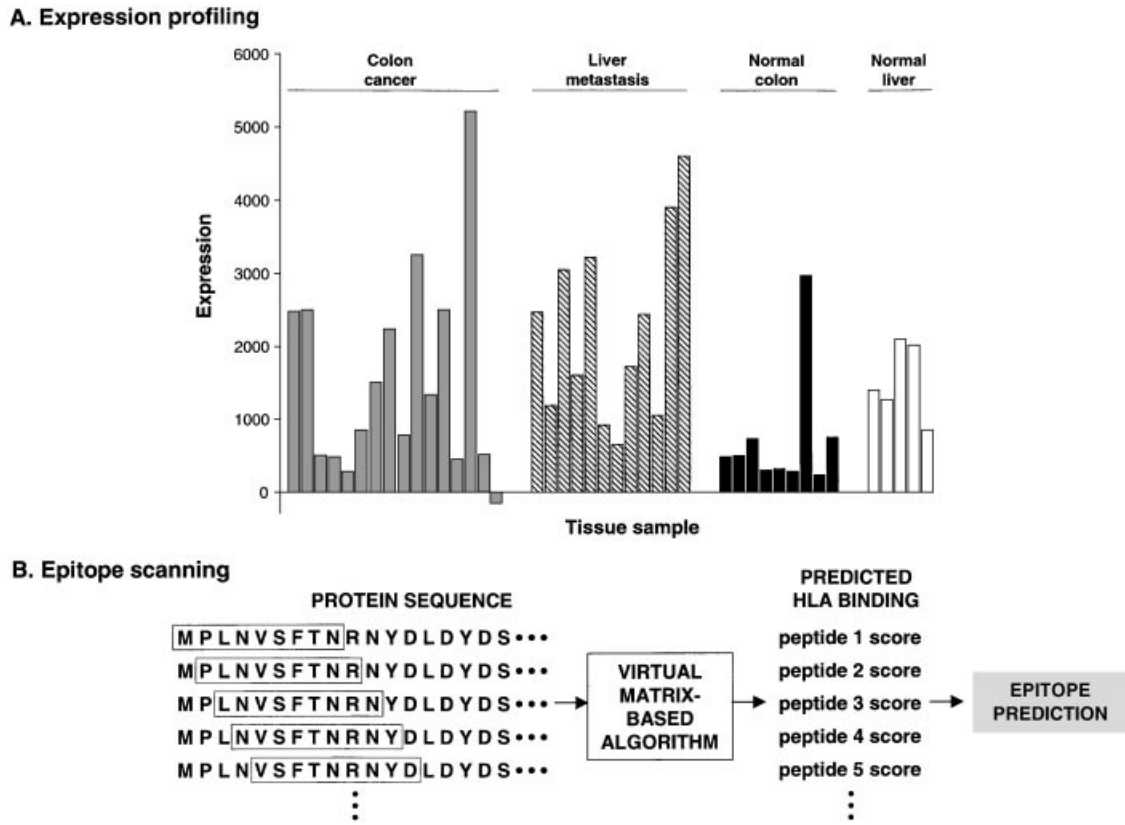

Fig. 9. Identification of T-cell epitopes by expression profiling combined with predictive algorithms. **A**: Expression profiling using Affymetrix GeneChip arrays indicates overexpression of c-myc in colorectal tumors and liver metastases. **B**: The epitope scanning procedure is outlined using c-myc as an example. The HLA binding affinities of peptides in a sliding window along the length of the protein sequence are predicted using a virtual matrix-based algorithm.

currently available epitope prediction models are based on HLA peptide binding data. This approximation is supported by the observation that HLA peptide binding is a major bottleneck in the selection of epitopes, as indicated by the finding that most peptide sequences lack the capacity to interact with HLA molecules.

Several approaches have been developed for HLA-based epitope prediction, reflecting both the different characteristics of peptide interaction with HLA-I and HLA-II, and the increasing structural and functional information that has become available over the past decade [Raddrizzani and Hammer, 2000]. An effective HLA-II epitope prediction tool is TEPITOPE, which is based on so-called quantitative matrices [Sturniolo et al., 1999] (Fig. 7). Quantitative matrices provide very detailed models in which the contribution to binding of each amino acid at each position within a binding core of a peptide is quantified. The position-specific amino acid values reflect the structural properties of HLA alleles, therefore constituting a "fingerprint" of HLA binding domains. Quantitative matrix-based prediction systems are linear models that are easy to implement and that result in a binding score for each query peptide.

Matrix-based prediction models have been validated for HLA-II in several retrospective studies (Fig. 8A). Furthermore, they have been successfully applied to predict T-cell epitopes in the context of oncology, allergy, and autoimmune diseases [Gross et al., 1998; de Lalla et al., 1999; Manici et al., 1999; Cochlovius et al., 2000; Stassar et al., 2001]. By starting with microarray data rather than whole genome sequence data, the number of possible tumor antigens can be reduced to a manageable number that can, in turn, be experimentally tested as part of a tumor vaccine development program (Figs. 8B and 9).

As these examples illustrate, there is tremendous potential for using large-scale gene expression data in the drug discovery process.

These new approaches to drug discovery have come about because the technical hurdles to genome-scale sequencing and high-throughput expression monitoring have been largely overcome. The next challenge is to increase the speed at which relevant functional information can be obtained for genes of interest, and eventually for every gene in the genome. As expression profiling methods continue to gain in sensitivity, genome coverage, and speed, the pressure for developing high-throughput functional genomics methods will only increase. As functional genomics technologies improve, they will continue to transform the drug discovery process by completing the process of rapidly moving from sequence to expression pattern to protein function.

REFERENCES

Blum JL. 2001. The role of capecitabine, an oral, enzymatically activated fluoropyrimidine, in the treatment of metastatic breast cancer. The Oncologist 6:56–64.

Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nat Genet 21(Suppl 1):33–37.

Cochlovius B, Stassar M, Christ O, Raddrizzani L, Hammer J, Mytilineos I, Zoller M. 2000. In vitro and in vivo induction of a Th cell response toward peptides of the melanoma-associated glycoprotein 100 protein selected by the TEPITOPE program. J Immunol 165:4731–4741.

de Lalla C, Sturniolo T, Abbruzzese L, Hammer J, Sidoli A, Sinigaglia F, Panina-Bordignon P. 1999. Cutting edge: Identification of novel T cell epitopes in Lol p5a by computational prediction. J Immunol 163:1725–1729.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531–537.

Gore MA, Morshedi MM, Reidhaar-Olson JF. 2000. Gene expression changes associated with cytotoxicity identified using cDNA arrays. Funct Integr Genomics 1:114– 126.

Gross DM, Forsthuber T, Tary-Lehmann M, Etling C, Ito K, Nagy ZA, Field JA, Steere AC, Huber BT. 1998. Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. Science 281:703–706.

Hammer J, Sturniolo T, Sinigaglia F. 1997. HLA class II peptide binding specificity and autoimmunity. Adv Immunol 66:67–100.

Heid CA, Stevens J, Livak KJ, Williams PM. 1996. Real time quantitative PCR. Genome Res 6:986–994.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Ishikawa T, Sekiguchi F, Fukase Y, Sawada N, Ishitsuka H. 1998. Positive correlation between the efficacy of capecitabine and doxifluridine and the ratio of thymidine phosphorylase to dihydropyrimidine dehydrogenase activities in tumors in human cancer xenografts. Cancer Res 58:685–690.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680.

Manici S, Sturniolo T, Imro MA, Hammer J, Sinigaglia F, Noppen C, Spagnoli G, Mazzi B, Bellone M, Dellabona P, Protti MP. 1999. Melanoma cells present a MAGE-3 epitope to CD4(+) cytotoxic T cells in association with histocompatibility leukocyte antigen DR11. J Exp Med 189:871–876.

Raddrizzani L, Hammer J. 2000. Epitope scanning using virtual matrix-based algorithms. Brief Bioinformatics 1:179–189.

SAS Institute. 1988. SAS/STAT User's Guide, Version 6, Ed. 4. Cary, NC: SAS Institute.

Stassar MJ, Raddrizzani L, Hammer J, Zoller M. 2001. T-helper cell-response to MHC class II-binding peptides of the renal cell carcinoma-associated antigen RAGE-1. Immunobiology 203:743–755.

Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J. 1999. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotech 17:555–561.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627.

Venter JC, et al. 2001. The sequence of the human genome. Science 291:1.