Overview of Commonly Used Bioinformatics Methods and Their Applications

IZET M. KAPETANOVIC,[a] SIMON ROSENFELD,[b] AND GRANT IZMIRLIAN[b]

[a]Chemopreventive Agent Development Research Group and [b]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, USA

ABSTRACT: Bioinformatics, in its broad sense, involves application of computer processes to solve biological problems. A wide range of computational tools are needed to effectively and efficiently process large amounts of data being generated as a result of recent technological innovations in biology and medicine. A number of computational tools have been developed or adapted to deal with the experimental riches of complex and multivariate data and transition from data collection to information or knowledge. These include a wide variety of clustering and classification algorithms, including *self-organized maps* (SOM), *artificial neural networks* (ANN), *support vector machines* (SVM), *fuzzy logic*, and even hyphenated techniques as *neuro-fuzzy networks*. These bioinformatics tools are being evaluated and applied in various medical areas including early detection, risk assessment, classification, and prognosis of cancer. The goal of these efforts is to develop and identify bioinformatics methods with optimal sensitivity, specificity, and predictive capabilities.

KEYWORDS: bioinformatics; data mining; cancer; early detection; risk assessment; hierarchical clustering; neural networks; support vector machines; fuzzy logic; genomics; proteomics; drug discovery

INTRODUCTION

Technological advances in biological and biomedical areas, especially in genomics and proteomics, are resulting in huge amounts of high-dimensional data with the number of "features" or "predictors" exceeding sample size by orders of magnitude. This new statistical paradigm, termed a "curse of dimensionality", has required the medical researchers to turn to the fields of **artificial intelligence** and **machine learning** in their quest for adequate analysis tools. Analogously as the recognition of the field of medical statistics as a specialized area of applied statistics resulted in the development of numerous new statistical methodologies, the fields of artificial intelligence and machine learning are currently being retooled and refined to suit the practical needs of bioinformatics. Integration of modern computational tools in bioinformatics and innovative high-throughput biotechnologies carry with them a great

Address for correspondence: Izet M. Kapetanovic, Chemopreventive Agent Development Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892-7322. Voice: 301-435-5011; fax: 301-402-0553.
kapetani@mail.nih.gov

Ann. N.Y. Acad. Sci. 1020: 10–21 (2004). © 2004 New York Academy of Sciences.
doi: 10.1196/annals.1310.003

impact potential in health care. The **-omic** (comprehensive analysis of components within a biological grouping, i.e., proteomics for proteins) revolution is stimulating development of new and reinvention of older computational methods. This review describes some of the commonly used and evolving bioinformatics methods and data mining tools and includes examples of their application to early cancer detection or diagnosis, risk identification, risk assessment, and risk reduction. It is not intended to be exhaustive, but only to present a brief overview. Individual topics are discussed in a more comprehensive manner in separate chapters in this volume. Another chapter reviews newer computational methods in cancer-related bioinformatics.

BIOINFORMATICS METHODS

Clustering

An excellent survey of the material on clustering techniques in microarray data analysis is Tibshirani *et al.*[1] (http://www-stat.stanford.edu/~tibs/research.html/). Clustering methods are used to arrange cell lines and genes in some natural order, with similar cell lines (and/or genes) placed close together. There are two major approaches to clustering: bottom-up and top-down. **Hierarchical clustering** is a bottom-up clustering method that starts with each cell line (gene) in its own cluster. It works by agglomerating the closest pair of clusters at each stage, successively combining clusters until all of the data are in one cluster. A number of methods are in common use for measuring the similarity of the expression profiles, such as ***Euclidean distance***, ***Pearson correlation***, ***Manhattan distance***, and others.[2] The relationships between each sample (or gene) is represented by a hierarchical tree, the **dendrogram**, which can be cut at any level to yield a specified number of clusters. Top-down clustering starts with a preset specified number of clusters and initial positions for the cluster centers. The **K-means** is used to reposition the cluster centers through the following steps: (1) observations are assigned to the closest cluster center to form a partition of the data; (2) the observations in each cluster are averaged, producing new values for the center vector of that cluster. Steps 1 and 2 are iterated, and the process converges to the minimum of total within cluster variance. Tree-structured vector quantization carries out K-means clustering in a top-down, binary manner. It is commonly used in image and signal compression. The **principal components analysis** (PCA), when applied to the genes, finds the linear combinations of gene expressions having the highest variance. Similarly, when applied to cell lines, it finds the highest variance linear combination of the cell lines. The correlation of each gene with the leading principal component provides a way of sorting (clustering) the genes as well as cell lines. The self-organizing map (SOM) is similar to K-means clustering, with the constraint that the cluster centers are restricted to remain in a one- or two-dimensional manifold. An iterative procedure is used to readjust the positions of the centers. There is a similarity between SOMs, multidimensional scaling, and principal components. In a comparative study, it was reported that K-means clustering produces tighter clusters than hierarchical clustering, but the latter tends to produce a greater number of smaller clusters, potentially a valuable feature for discovery.[1]

Unlike K-means clustering, hierarchical clustering produces an ordering of the objects, which can be informative for data display. Because SOMs are constructed from a two-dimensional representation of the data, it is a good idea to check the resulting predicted classes against an algorithm that functions directly in the original dimension of the data set, such as K-means.

The above methods are the one-way clustering techniques; however, the use of two-way clustering, that is, simultaneous clustering both the genes and cell lines, has also been investigated.[1] A simple approach to this problem is to apply a one-way clustering method separately to the genes and to the cell lines. Block clustering, in contrast, uses both gene and cell line information to simultaneously cluster both. The two-way clustering procedures seek a global organization of genes and cell lines. This study reported that these types of procedures are able to discover gross global structure, but may not be effective for discovering finer detail. In response to this finding, a new method called **gene shaving** was proposed. The gene shaving technique can search for sets of genes that optimally separate the cell lines.[3] The algorithm begins by finding a nested sequence of candidate clusters, with all gene clusters in the initial position and one gene cluster in the final position.

Artificial Neural Networks

Artificial neural networks (ANN), modeled after normal brain processes and neurophysiological learning, are powerful computational tools for multifactorial classification and multivariate nonlinear regression. Technological advances and availability of computational power brought by the era of personal computers made ANN a popular method for routine analysis in a wide spectrum of scientific and engineering applications, including automatic target recognition, stock market analysis, expert systems, pattern recognition, medical imaging, and DNA microarray analysis. However, ANN methodologies often represent more art than science. Many decisions related to the choice of ANN structure and parameters are often completely subjective. Theoretical recommendations for the size of training data set are lacking, and an optimum size is almost never available in practice. Special attention also has to be paid to avoid overtraining that would result in memorization instead of generalization of the data. Therefore, there is a considerable uncertainty in the optimal design of the ANN architecture. The final ANN solution may be influenced by a number of factors (e.g., starting weights, number of cases, and their order during the training phase, number of training cycles, etc.).

Bootstrap sampling (random sampling with replacement) to produce a large number of individual neural networks was one proposed approach addressing this problem.[4] Parametric or nonparametric statistical analysis of the resultant distribution of neural networks would yield predictive intervals. ANN is a strongly nonlinear approximation and, therefore, the topology of the objective function in the space of the ANN parameters is usually very complex. In particular, the objective function may contain many local minima, and gradient methods of optimization, such as Newton-Raphson, steepest descent, etc., can easily lead the minimization procedure to one of these local minima, resulting in very suboptimal weights. To avoid this problem, a number of sophisticated optimization algorithms have been developed, such as **genetic algorithm**, **simulated annealing**, and various versions of **stochastic optimization**.

Genetic algorithms are based on the concept of natural selection, survival of the fittest. Using principles of inheritance, mutation and/or cross-over, genetic algorithms generate a series of random potential solutions (population) and use objective fitness function to evaluate each member of the population. The probability of a particular gene being copied in the next generation (reproduction) is determined by its fitness (i.e., its contribution to the objective function). The reiterative process continues with the fitter members until the fittest member of the population is identified as the optimized solution. It has been shown theoretically and computationally that such a process provides a random walk in the space of the ANN parameters toward minimum of the objective function. A fundamental advantage of the genetic algorithm is that in principle it is able to find the global minimum. Genetic algorithms are highly intensive computationally and require millions of readjustments ("generations") of the objective function to reach a convergence.

Simulated annealing is another method of finding a global minimum. The set of arguments of the objective function (the ANN weights, for example) are likened to a thermodynamic system. Objective function is considered as its energy, and the search for the minimum is analogous to the search of thermodynamically stable state with the lowest energy possible. It was shown that simulated annealing has a high probability to converge to a global minimum. Similar to genetic algorithm, simulated annealing is highly computationally intensive.

There are many versions of stochastic optimization. Their common theme is a random walk in the phase space toward the minimum of objective function. The arguments of objective function are perturbed randomly. The new set of parameters are accepted if the objective function is decreasing. The whole process is often likened to the random walk of a drunken person in his/her attempt to find the way home. A surprising feature of this kind of algorithm is that average time required to reach the goal is often smaller than that resulting from exact analysis and prediction of each step. The advantages of stochastic algorithms are especially noticeable for the random walk in high-dimensional space. Generally speaking, this family of algorithms does not guarantee convergence to the global minimum. However, a great advantage of stochastic optimization methods is that, unlike the gradient methods, they generally do not require computation of the objective function gradients. Taking into account high dimension of the parameter space and high computational cost of the gradient evaluation, this feature is highly important and makes the convergence process comparatively fast.

Despite all the obstacles and difficulties in design and training of ANNs, there are numerous examples of highly successful applications of ANNs. Recently, a marked increase in application of ANNs in biomedical areas, especially in cancer research, has been observed. It is currently widely recognized that cancer risk evaluation based on a single or few biomarkers may not be possible. ANN is inherently suited in this regard because of its ability to perform simultaneous analysis of large amounts of diverse information. **ROC** (receiver operating characteristics curve) methodology, which is frequently used as a measure of classification performance, has been adapted to evaluation of the ANN performance.[4,5] The y-axis and x-axis on the ROC curves represent sensitivity and specificity, respectively, and the area under the curve is an indication of how well the independent variable separating two dichotomous classes performs.

Support Vector Machine

An important recent innovation in the statistical learning theory is the support vector machine (SVM).[6] SVM represents a particular instance of a large class of learning algorithms known as kernel machines and is a powerful supervised algorithm for classification. This algorithm projects data into higher dimensional space where two classes are linearly separable. It finds a hyperplane in the space of the data points that separates two classes of data and maximizes the width of a separating band between the data points and the hyperplane. The support vectors are defined as the ones nearest to this margin, and only the support vectors define the model and need to be stored. There are many fundamental advantages of the SVM algorithms compared with other methods. First, unlike ANN, SVM produces a unique solution because it is basically a linear problem and does not have such a pitfall as multiple local minima. Second, SVM is inherently able to deal with very large amounts of dissimilar information. Third, the discriminant function is characterized by only a comparatively small subset of the entire training data set, thus making the computations noticeably faster. SVM is a highly promising tool in genomics and proteomics.

Boosting

Abundance of exploratory tools, each possessing their pros and cons, creates a difficult problem of selecting the best of them. It seems to be a good idea to try to combine their strengths for creating an even more powerful tool. To a certain extent, this idea has been implemented in a new family of classification algorithms known under the general term “boosting”. **Boosting** was proposed in a series of groundbreaking works.[7] Boosting is a general method for combining many weak classifiers to produce a stronger classifier. Boosting sequentially applies a classification algorithm to reweighed versions of the training data and then takes a weighted majority vote of the sequence of classifiers thus produced. For many classification algorithms, this simple strategy results in a dramatic improvement in performance. This seemingly mysterious phenomenon can be understood in terms of well-known statistical approaches, such as additive models and maximum likelihood.[8]

Bagging

Another technique that has evolved as a mechanism for improving existing classification algorithms is “**Bagging**”, an acronym for (B)ootstrap (Agg)regation.[9] Given a particular classifier and a data set, bagging proceeds by drawing B bootstrap samples from the data set (random sample with replacement of equal size). Each bootstrap sample trains a classifier. Since sampling with replacement tends to pick from those already sampled about a third of the time, a bootstrap sample of size n contains roughly $2n/3$ unique samples. Consequently, $n/3$ of the original sample is left out. The validation step is carried out by predicting class membership, for each of the n elements of the original sample, using the (roughly) $B/3$ classifiers that element did not train. Final class membership is predicted using the most popular vote. The point of bagging a classifier is to pick a middle way between overfitting (low variance, but high bias) and oversmoothing (low bias, but high variance). A very promising new tool that incorporates bagging in a very clever way is **random**

forests.[10–12] This tool performs bagging on classification (or regression) trees with the added novel idea of random feature set selection each time a node is split during the training process. This has the effect of decorrelating the ensemble of classification/regression trees and helps to strengthen the divide between training and validation. An especially nice additional property of random forest is that it performs so well with practically no real tuning parameters. The random forest algorithm has recently been successfully applied in the analysis of proteomics data.[13]

Fuzzy Logic

The real world, including that of medicine, is imprecise, vague, and ambiguous, that is, fuzzy. Lotfi Zadeh, the founder of fuzzy logic, proposed that one could exploit tolerance for imprecision and partial truths to achieve tractability, robustness, interpretability, and decreased computational cost. Fuzzy logic deals with ambiguity and vagueness, as opposed to probability that involves uncertainty and likelihood. A distinction between fuzzy and binary or crisp logic is that the former involves concepts of more or less or degree of membership (partial set membership) or continuity as opposed to yes or no or absence or presence or discreteness. Fuzzy logic uses the linguistic variable (i.e., computing with words instead of numbers). It provides a mathematical tool for representing and manipulating information in a way that resembles human communication and reasoning processes. It embeds existing structured human knowledge into workable mathematics. A typical fuzzy inference system includes fuzzification (classifying numeric data into fuzzy sets), knowledge/rule base (employing linguistic reasoning with “if … then” rules mapping the input into output variables), inference engine (applying the rule base to the fuzzy set to obtain a fuzzy outcome), and defuzzification process (converting the fuzzy outcome to a crisp one).

Nowadays, fuzzy logic devices are present in many everyday consumer products (e.g., automobile brakes, camera and camcorder autofocus, meteorology instrumentation, intravenous infusion pumps, kitchen devices, etc.). Theory and applications of fuzzy logic in medicine have been reviewed in several recent publications.[14–18] Some of the fuzzy techniques that have been employed for biomedical data analysis include fuzzy clustering, fuzzy classification, and hybrid systems, such as combinations of fuzzy logic and neural networks (neuro-fuzzy networks), genetic algorithm, evolutionary algorithm, or discrete wavelet transforms. For example, the advantage of some hybrid methods like neuro-fuzzy systems is that they combine the advantages of fuzzy systems that deal with explicit knowledge (understood and explainable) and neural networks that deal with implicit knowledge (acquired by learning).

APPLICATIONS

Genomics in Early Cancer Detection and Classification

The innovative technologies of gene expression analysis are providing promising tools for the identification of cancer cell signatures and cancer molecular targets, thereby facilitating early detection of cancer and intervention. Computational methods used for microarray data analysis have been reviewed.[19] Perou *et al.*[20] used hierarchical clustering for human breast tumor classification and identification of molecular portraits. Alizadeh *et al.*[21] studied gene expression in the three most prevalent adult lymphoid malignancies. Two previously unrecognized types of diffuse large B-cell lymphoma, with distinct clinical behaviors, were identified based on gene expression data and were shown to have markedly different median survival, even within a low-risk profile, according to the currently accepted diagnostic criteria. Two-way hierarchical clustering on cell lines and on genes was used to identify the two tumor subclasses, as well as to group genes with similar expression patterns across the three different samples. However, these results were obtained after thresholding, which is a step fraught with problems over validity. Using a gene shaving technique, it was possible to duplicate two subtypes with differing median survival in an unsupervised manner, without thresholding.[3] Hierarchical clustering has several drawbacks (non-uniqueness, inversion problems, grouping based on local decision, lack of an opportunity to reevaluate the clustering, etc.) and other approaches have also been employed. Unsupervised methods, such as the Kohonen SOMs, have been used for gene clustering in promyelocytic leukemia.[22]

Use of supervised methods for microarray data analysis has also been recently reviewed.[23] Supervised ANNs have been used to classify estrogen receptor status in human breast tissue following PCA to reduce the dimensionality[24] and correctly classify the small, round blue-cell tumor subtypes and identify possible gene targets for therapy.[25] Others have successfully applied ANN to distinguish among subtypes of neoplastic colorectal lesions and showed that ANN outperformed hierarchical clustering in classification power and was able to distinguish between 27 different subtypes of neoplastic colorectal lesions.[26] SVMs have also been successfully applied in microarray gene expression analysis,[27] tumor classification,[28,29] cancer diagnosis,[30] and prognosis.[31] A group at Stanford developed supervised learning software for genomic expression data mining and made it available from their Web site (http://www-stat.stanford.edu/~tibs/SAM/index.html/).[32] It is in the form of an *Excel* add-in and is applicable to cDNA, oligo, SNP, and protein array data. It correlates expression data to clinical parameters. A fuzzy logic approach to identify connected networks of genes describing how the genes interrelate has also been proposed.[33]

Proteomic Profiling, Bioimaging, and Pattern Recognition

Proteomics is the analysis of the proteome, which is a term applied to the proteins expressed by the genome of a species. Importance of the problem was recently emphasized by founding the Human Proteome Organization (HUPO) (http://www.hupo.org/) with the aim of elucidating the human proteome. In general, there is a poor correlation between mRNA and protein levels.[34] In addition, genomics does not provide information regarding posttranslational events (such as phosphorylation, acetylation, lipidation, glycosylation, or ubiquitination). Proteome imparts cellular functionality as proteins carry out most of the work of the cell. The majority of drug targets are proteins. Proteomic fingerprinting provides complementary information to genomic fingerprinting. Proteins can serve as markers and targets of chemoprevention. Newer technological advances have enabled growth and interest in proteomics. Some of the tools presently available or under development include **surface-enhanced**

laser desorption/ionization time-of-flight (SELDI-TOF) and **matrix-assisted laser desorption/ionization time-of-flight** (MALDI-TOF), mass spectrometric (MS) techniques, **surface plasmon resonance** (SPR) technology, **laser capture microdissection** (LCM), and a number of protein, antibody, and tissue microarrays.

There are a number of difficulties inherent in proteomic research. For example, estimated number of proteins in the proteome exceeds that of genes in the genome by more than an order of magnitude. There is no PCR equivalent to amplify protein signal. Proteins are in a continual dynamic flux depending on cellular status and activity. Proteomic research will lead to new developments in identification and detection of biomarkers, recognition of new drug targets, individualized patient therapy, and enhancements in rational drug design. These new technologies have facilitated disease detection and diagnosis based on protein fingerprinting, relying on multiple instead of single protein biomarkers. It is not simply a matter of presence or absence of number of proteins, but rather their relative amounts to each other. This realization led to a considerable improvement in predictability. Bioinformatics tools are critical in the analysis of the huge amount of data being generated by the newer parallel analytical techniques.

Appropriate combinations of analytical and bioinformatics tools have been used to define an optimum discriminatory proteomic pattern in women without sign of disease, early-stage ovarian cancer, late-stage ovarian cancer, and benign diseases.[35] In this study, genetic algorithm and unsupervised SOM were used to analyze SELDI-TOF data from serum proteins applied to a hydrophobic interaction protein chip. In order to identify 5 proteins with different relative abundances between two training sets, on the order of 10^{20} combinations would be required and would be overwhelming even with today's computer technology. However, the problem can be greatly simplified by use of a genetic algorithm. Petricoin *et al.*[35] report that this approach yielded 100% sensitivity, 95% specificity, and 94% predictability.

Another study presented promising preliminary results in using neural network with a back-propagation algorithm for the tumor classification and biomarker identification of human astrocytoma based on tissue protein data.[36] Qu *et al.*[37] applied a modified AdaBoost algorithm proposed by Freund and Schapire.[7] Using this algorithm, 97% sensitivity and specificity have been achieved in discrimination between healthy men and those with prostate cancer and benign prostate hyperplasia based on serum protein data. The same group also obtained satisfactory results (positive predictive value of 91% for general population) for early detection of prostate cancer based on serum protein fingerprinting with a decision tree classification algorithm.[38]

ANN is also a powerful tool for **image analysis** and **pattern recognition**. It is the technique behind the FDA-approved computer-assisted diagnosis instrumentation ImageChecker® that is used to analyze digital mammograms and draws the physician's attention to suspicious features that may be indicative of cancer (http://www.r2tech.com/prd/prd001.html/).

ANNs have been used to classify patterns of subcellular structures in fluorescence microscope images of HeLa cells.[39] Images of subcellular structures were parameterized using an elaborate system of 37 geometric and texture features. The 37 vectors of these features were used as input vectors for the back-propagation ANN. The ANN was able to successfully recognize all 10 subcellular structures used in the training process. This method allows monitoring of dynamic protein properties and relates them to changes with disease states and therapeutic intervention.

Usefulness of fuzzy logic was demonstrated in **computer-aided cancer detection** in the areas of bioimaging, classification, and pattern recognition. A fuzzy logic approach was used in detection of lobulated and microlobulated masses in digital mammography,[40] and fuzzy-neural and feature extraction techniques were employed for detecting and diagnosing microcalcifications on digital mammograms.[40] A breast cancer diagnosis system (BCDS) combined a fuzzy microcalcification detection algorithm with a feature extraction method and a back-propagation neural network (BPNN) for classification of benign or malignant microcalcifications with 89% classification rates.[41]

Multifactorial Analysis of Early Detection, Risk Identification, Risk Assessment, and Risk Reduction of Cancer

Cancer is a complex, multifactorial collection of diseases. The goal of chemoprevention is to identify risks, assess risks, detect early, and intervene to reduce risk of cancer prior to the appearance of clinical signs and pathological abnormalities. Individual variables (biomarkers and indicators of cancer) are not adequately predictive or discriminatory. However, simultaneous consideration of multiple factors (**composite medical index** or **panel of markers**) should provide a more useful indication into the initiation, progression, and reversal of carcinogenesis. A further complication is that the predictability of an outcome is not based on presence or absence of several biomarkers or their linear summation, but on a complex, nonlinear relationship between them. The challenge is to identify suitable composite medical indices that would provide acceptable sensitivity, specificity, and predictability.

Different multivariate analytical tools have been employed to identify a composite variable for early detection, risk identification, risk assessment, and risk reduction of cancer. ANNs have been frequently used in cancer detection, cancer classification, and prognosis.[4,25,26,36,42,43] Inputs into ANNs can include data from any or all of the following: clinical findings, clinical chemistry, gene microarrays, protein microarrays, biomarkers, genetic factors, environmental factors, etc. The output variable represents a **composite variable** or a predicted outcome (in terms of a cancer risk or prognosis) for the individual patient.[4]

Other computational methodologies that have been shown to be useful in the area of diagnosis, classification, and prediction based on multifactorial analysis include SVMs,[30,31] genetic algorithms and SOMs,[35] and fuzzy logic.[44]

The fuzzy logic approach in conjunction with a panel of biomarkers demonstrated accuracy, sensitivity, and specificity in the diagnosis and classification of lung cancer.[44] For example, this study was able to distinguish malignant versus benign cases with a sensitivity of 88% and a specificity of 86%. It was also able to discriminate between non-small-cell carcinoma (NSCLC) versus small-cell carcinoma (SCLC) with sensitivity and specificity of 91% and 91% and squamous versus adenocarcinoma with sensitivity and specificity of 77% and 79%, respectively. This approach was especially effective in early stages of cancer and in patients with all marker levels in the gray area. Another study[45] employed genetic algorithm to automatically produce a fuzzy BCDS in relation to the Wisconsin Breast Cancer Diagnosis (WBCD) database. This **fuzzy-genetic** approach provided a high classification performance and interpretability. Subsequently, these same authors[46] introduced a combination of a fuzzy system and a **cooperative coevolutionary** approach in the area of breast cancer diagnosis. Evolutionary methods are search or optimization techniques and are especially useful in search of large and complex spaces. Cooperative coevolutionary approach involves two coevolving cooperative species (database membership function and a rule base). Two genetic algorithms (subset of evolutionary algorithms) were used to control the evolution of two populations. The advantages of this approach were that it provided higher classification performance with a lower computational cost than other systems.

In all cases, these multifactorial computational methodologies have enabled or significantly improved detection, classification, or prognosis of cancer over evaluations based on a single variable.

Drug Discovery

Pharmaceutical companies recognize the value of the huge amounts of data being generated by new **-omics** and high-throughput technologies, and are trying to leverage these data into drug discoveries with the help of evolving bioinformatics tools. In an effort to streamline, expedite, and optimize drug discovery and development, pharmaceutical companies are actively incorporating bioinformatics into their practices. As a result, new fields such as **chemogenomics**, **chemical genomics**, and **chemical genetics** have emerged. Although definitions for these fields vary and overlap, these areas encompass new approaches to drug discovery and therapeutic target identification/validation.[47–49] The idea behind them is that drugs can be used to identify new therapeutic targets, and targets can be used to identify new drugs in the context of genomics/proteomics. In a sense, chemical genomics integrates chemical structure space and biological structure space and provides an *in silico* approach to drug discovery and optimization. There are a number of new companies with a primary focus on chemical genomics that have emerged recently and are developing their versions of the chemical genomic mousetraps. In addition, another related **-omic** discipline has emerged: pharmacogenomics.[50] **Pharmacogenomics** is a discipline examining an individual's response to drugs based on an individual's genetic makeup. It promises to enable optimized, personalized therapy for each patient. Advances in biological, analytical, and computational technologies have allowed for emergence of innovative **computer-aided drug design** (CADD). Genetic algorithms are frequently used in CADD.[51,52] New data mining techniques and visualization tools can be used to characterize effects of chemopreventive intervention and thereby facilitate drug discovery. The expectation is that they will be useful in identifying appropriate targets, suitable biomarkers, and more fitting drug agents.

REFERENCES

1. Tibshirani, R., T. Hastie, M. Eisen *et al.* 1999. Clustering methods for the analysis of DNA microarray data. Technical report. Stanford University. Stanford, CA.
2. Jagota, A. 2001. Microarray data analysis and vizualization. Department of Computer Engineering, University of California, Santa Cruz. Santa Cruz, CA.
3. Hastie, T., R. Tibshirani, M.B. Eisen *et al.* 2000. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. **1**(2)**:** 0003.1–0003.21 (http://genomebiology.com/2000/1/2/research/0003/).
4. Dayhoff, J.E. & J.M. DeLeo. 2001. Artificial neural networks: opening the black box. Cancer **91:** 1615–1635.

5. DeLeo, J.M. & J.E. Dayhoff. 2001. Medical applications of neural networks: measures of certainty and statistical tradeoffs. *In* International Joint Conference on Neural Networks (IJCNN '01), pp. 3009–3014. IEEE Press. New York.
6. Vapnik, V.N. 1998. Statistical Learning Theory. Wiley. New York.
7. Freund, Y. & R.E. Schapire. 1996. Experiments with a new boosting algorithm. *In* Machine Learning: Proceedings of the Thirteenth International Conference, pp. 148–156. Morgan Kaufmann. San Francisco.
8. Friedman, J., T. Hastie & R. Tibshirani. 1996. Additive statistical regression: a statistical view of boosting. Ann. Stat. **28:** 337–407.
9. Hastie, T., R. Tibshirani & J. Friedman. 2001. The Elements of Statistical Learning, Springer Pub. New York.
10. Breiman, L. 2001. Random forests. Mach. Learn. **45**(1)**:** 5–32.
11. Breiman, L. 2002. A Manual on Setting Up, Using, and Understanding Random Forests V3.1 (http://oz.berkeley.edu/users/breiman/using_random_forests_v3.1.pdf/).
12. Liaw, A. & M. Weiner. 2003. The Random Forest Package in R (http://cran-us-r.project.org/).
13. Xiao, Z., B. Luke, G. Izmirlian *et al.* 2003. Submitted.
14. Vitez, T.S., R. Wada & A. Macario. 1996. Fuzzy logic: theory and medical applications. J. Cardiothorac. Vasc. Anesth. **10:** 800–808.
15. Steimann, F. 1997. Fuzzy set theory in medicine. Artif. Intell. Med. **11:** 1–7.
16. Kuncheva, L.I. & F. Steimann. 1999. Fuzzy diagnosis. Artif. Intell. Med. **16:** 121–128.
17. Abbod, M.F., D.G. von Keyserlingk, D.A. Linkens *et al.* 2001. Survey of utilization of fuzzy technology in medicine and healthcare. Fuzzy Sets Syst. **120:** 331–349.
18. Mahfouf, M., M.F. Abbod & D.A. Linkens. 2001. A survey of fuzzy logic monitoring and control utilisation in medicine. Artif. Intell. Med. **21:** 27–42.
19. Quackenbush, J. 2001. Computational analysis of microarray data. Nat. Rev. **2:** 418–427.
20. Perou, C.M., T. Sorlie, M.B. Eisen *et al.* 2000. Molecular portraits of human breast tumours. Nature **406:** 747–752.
21. Alizadeh, A.A., M.B. Eisen, R.E. Davis *et al.* 2000. Distinct types of large B-cell lymphoma identified by gene expression profiling. Nature **403:** 503–511.
22. Tamayo, P., D. Slonim, J. Mesirov *et al.* 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA **96:** 2907–2912.
23. Ringner, M., C. Peterson & J. Khan. 2002. Analyzing array data using supervised methods. Pharmacogenomics **3:** 403–415.
24. Gruvberger, S., M. Ringner, Y. Chen *et al.* 2001. Estrogen receptor in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. **61:** 5979–5984.
25. Khan, J., J.S. Wei, M. Ringner *et al.* 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7:** 673–679.
26. Selaru, F.M., Y. Xu, J. Yin *et al.* 2002. Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. Gastroenterology **122:** 606–613.
27. Brown, M.P.S., W.N. Grundy, D. Lin *et al.* 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA **97:** 262–267.
28. Furey, T.S., N. Cristianini, N. Duffy *et al.* 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16:** 906–914.
29. Yeang, C-H., S. Ramaswamy, P. Tamayo *et al.* 2001. Molecular classification of multiple tumor types. Bioinformatics **17**(suppl. 1)**:** S316–S322.
30. Ramaswamy, S., P. Tamayo, R. Rifkin *et al.* 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA **98:** 15149–15154.
31. Shipp, M.A., K.N. Ross, P. Tamayo *et al.* 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. **8:** 68–74.
32. Tusher, V.G., R. Tibshirani & G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98:** 5116–5121.

33. Woolf, P.J. & Y. Wang. 2000. A fuzzy logic approach to analyzing gene expression data. Physiol. Genomics **3:** 9–15.
34. Pratt, J.M., J. Petty, I. Riba-Garcia *et al.* 2002. Dynamics of protein turnover, a missing dimension in proteomics. Mol. Cell. Proteomics **1:** 579–591.
35. Petricoin, E.F., A.M. Ardekani, B.A. Hitt *et al.* 2002. Use of proteomic patterns in serum to identify ovarian cancer. Lancet **359:** 572–577.
36. Ball, G., S. Mian, F. Holding *et al.* 2002. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics **18:** 395–404.
37. Qu, Y., B-L. Adam, Y. Yasui *et al.* 2002. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin. Chem. **48:** 1835–1843.
38. Adam, B-L., Y. Qu, J.W. Davis *et al.* 2002. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res. **62:** 3609–3614.
39. Boland, M.V. & R.F. Murphy. 2001. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics **17:** 1213–1223.
40. Kovalerchuk, B., E. Triantaphyllou, J.F. Ruiz *et al.* 1997. Fuzzy logic in computer-aided breast cancer diagnosis: analysis of lobulation. Artif. Intell. Med. **11:** 75–85.
41. Verma, B. & J. Zakos. 2001. A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. IEEE Trans. Inf. Technol. Biomed. **5:** 46–54.
42. Errejon, A., E.D. Crawford, J. Dayhoff *et al.* 2001. Use of artificial neural networks in prostate cancer. Mol. Urol. **5:** 153–158.
43. Xu, Y., F.M. Selaru, J. Yin *et al.* 2002. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Res. **62:** 3493–3497.
44. Keller, T., N. Bitterlich, S. Hilfenhaus *et al.* 1998. Tumor markers in the diagnosis of bronchial carcinoma: new options using fuzzy logic–based tumour marker profiles. J. Cancer Res. Clin. Oncol. **124:** 565–574.
45. Peña-Reyes, C.A. & M. Sipper. 1999. A fuzzy-genetic approach to breast cancer diagnosis. Artif. Intell. Med. **17:** 131–155.
46. Peña-Reyes, C.A. & M. Sipper. 2000. Evolutionary computation in medicine: an overview. Artif. Intell. Med. **19:** 1–23.
47. Lenz, G.R., H.M. Nash & S. Jindal. 2000. Chemical ligands, genomics, and drug discovery. Drug Discovery Today **5:** 145–156.
48. Dean, P.M. & E.D. Zanders. 2002. The use of chemical design tools to transform proteomics data into drug candidates. Biotechniques Suppl. **32:** S28–S33.
49. Zheng, X.F. & T.F. Chan. 2002. Chemical genomics: a systematic approach in biological research and drug discovery. Curr. Issues Mol. Biol. **4:** 33–43.
50. McLeod, H.L. & W.E. Evans. 2001. Pharmacogenomics: unlocking the human genome for better drug therapy. Annu. Rev. Pharmacol. Toxicol. **41:** 101–121.
51. Pegg, S.C., J.J. Haresco & I.D. Kuntz. 2001. A genetic algorithm for structure-based *de novo* design. J. Comput. Aided Mol. Design **15:** 911–933.
52. Terfloth, L. & J. Gasteiger. 2001. Neural networks and genetic algorithms in drug design. Drug Discovery Today **6**(suppl.)**:** S102–S108.