Sun-Yuan Kung
and Man-Wai Mak

Machine Learning for Multimodality Genomic Signal Processing

One of the main challenges in computational biology is the revelation and interpretation of the rich genomic information underlying cancer biology. Revealing such information can help facilitate classification and prediction of cancers and responses to therapies. Genomic sequencing and gene expression technologies have been widely recognized as vital approaches to modern drug design and disease classification. A collection of microarray experiments can yield a data matrix, whose rows simultaneously reveal the expression level of tens of thousands of genes, a promising advance towards a genome-wide exploration.

Machine learning has been a promising computational approach to genomic studies. A machine learning system comprises two subsystems: 1) feature extraction and 2) adaptive classification or cluster discovery. For adaptive classifiers, both unsupervised and supervised training strategies have been found useful.

In the past, the prevalent approaches to gene classification relied on only a single modality. However, recently there is increasing evidence of substantial performance improvement exacted by combining information from multiple sources. This calls for a new classification approach based on multimodality fusion.

MULTIMODALITY FUSION FOR GENOMIC SIGNAL PROCESSING

There are many ways to generate multiple modalities. One possibility is via sensor diversity, and the other is feature diversity. In terms of sensor diversity, we shall consider both the motif and gene expression modalities. Motifs are short sequences of DNA responsible for regulating gene networks and the expression of genes, whereas gene expression is the process of producing proteins from information coded in genes. A genome-wide study via the pure sequencing approach is computationally prohibitive. Thus, gene expression analysis can play a vital and complementary role in this respect. To further facilitate multimodality fusion, a diversity of features may be extracted from each sensor by computational means. This is called feature diversity.

There are two basic fusion architectures, cascaded and parallel, suitable for genomic signal processing. For simplicity, we shall use only two modalities to illustrate the possible fusion strategies. (The extension to the fusion of multiple modalities is straightforward.) The fusion architectures can be mathematically represented as

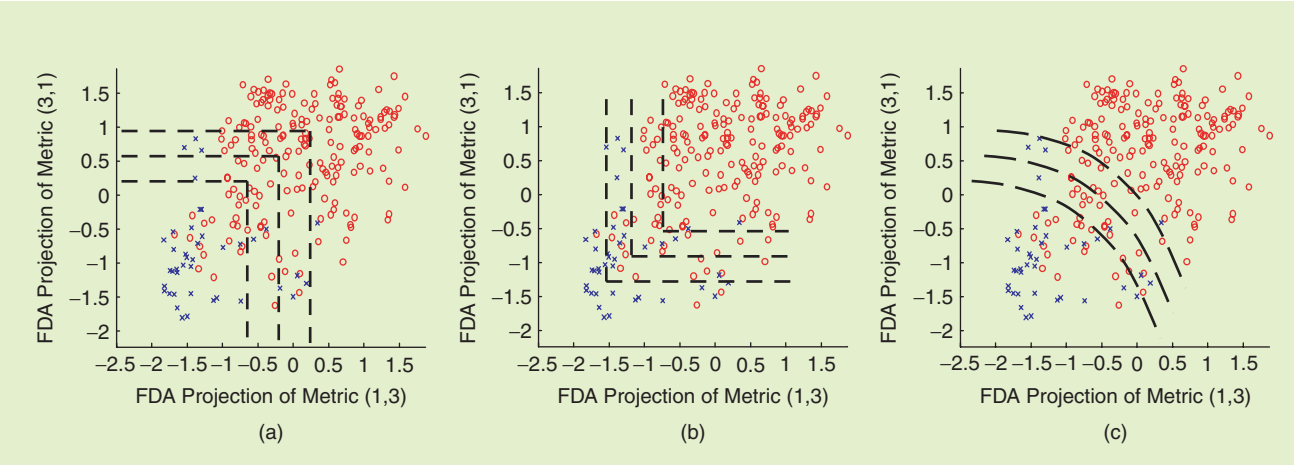
Cascaded:

$$f(\gamma(X), Y) \rightarrow \text{Fusion output} \quad (1)$$

Parallel:

$$g(\phi(X), \psi(Y)) \rightarrow \text{Fusion output}, \quad (2)$$

where X and Y are the features of the two modalities. Typically, f and g are binary functions (i.e., they either output positive or negative class labels), whereas



[FIG1] Various patterns of decision boundaries created by (a) AND logic, (b) OR logic, and (c) nonlinear fusion. The patterns shown here reflect real data: blue crosses (×) represent positive (ribosomal) genes and red circles (○) represent negative (nonribosomal) genes. These examples illustrate that nonlinear fusion not only provides a unified solution but potential improvement in prediction performance as well (see Figure 2).

1053-5888/06/$20.00©2006IEEE

γ, ϕ, and ψ can be either binary functions or score functions.

Fusion can be viewed as a classification problem in which optimal decision boundaries (see Figure 1) are defined to separate the positive and negative patterns in the score space. The effect of fusion can be illustrated by the receiver operating characteristics (ROCs) [4] as shown in Figure 2. The performance of cluster discovery and the improvement gained by fusion can be evaluated via its goodness for gene prediction. Specifically, a test gene t will be predicted as positive if

$$\log p\left(s(t)|\Lambda^{(p)}\right) > \log p\left(s(t)|\Lambda^{(n)}\right) + \alpha, \tag{3}$$

where $p(s(t)|\Lambda^{(p)})$ and $p(s(t)|\Lambda^{(n)})$ denote the distribution of the feature values $s(t)$s being generated by the positive and negative models $\Lambda^{(p)}$ and $\Lambda^{(n)}$, respectively. It is common to adopt the following performance measures: Precision $=$ TP/(TP $+$ FP), Sensitivity $=$ TP/(TP $+$ FN), and Specificity $=$ TN/(FP $+$ TN), where TP, TN, FP, and FN are the numbers of true-positives, true-negatives, false-positives, and false-negatives, respectively. ROCs display the tradeoff of two of these measurements over their entire range. To produce the entire spectrum of sensitivity-precision-specificity ROCs, we can gradually adjust the running variable α in (3) to change from small to large values. For example, by setting $\alpha < 0$ (respectively, $\alpha > 0$), we can obtain a higher sensitivity (respectively, precision).

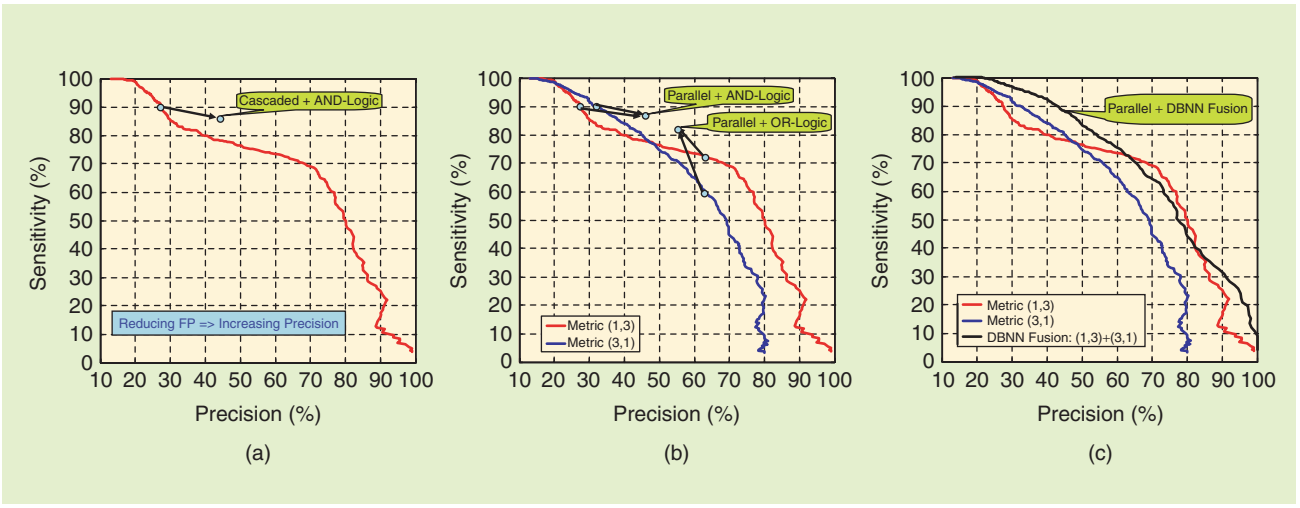
Let us illustrate an application example of cascaded fusion. Suppose that a gene is classified to be positive if and only if it is admitted by both modalities. Such a classification can be implemented by a cascaded architecture with the AND-logic in Figure 1(a). Suppose the first modality, adopting an exploded search space, yields a result with high sensitivity (due to a large number of true positives) at the expense of poor precision (due to a large number of false positives). Now, according to (1), the second modality (represented by f) can be used to further screen all the genes that are admitted to be positive by the first modality (represented by γ). Suppose a far greater number of false positives than true positives can be weeded out, then the precision will be substantially improved while the sensitivity remains very much intact. This process is manifested by the boldfaced arrow depicted in Figure 2(a).

A similar argument can be applied to parallel fusion. Suppose that a gene is predicted to be positive as long as it is admitted by either of the two modalities. This can be implemented by using the parallel architecture with the OR-logic in Figure 1(b). The potential improvement of such a fusion scheme is manifested by the boldface arrows in Figure 2(b).

The general scheme is that with a parallel architecture, nonlinear fusion leads to nonlinear decision boundaries as illustrated in Figure 1(c). All the previously mentioned scenarios are just special cases of nonlinear fusion. Moreover, nonlinear fusion is often implemented via a machine learning or neural network approach. Therefore, the machine learning approach provides a unified framework for multimodality fusion.

SENSOR DIVERSITY: A BIOLOGICAL PERSPECTIVE

The genome is not just a collection of genes working in isolation, but rather it encompasses the global and highly coordinated control of information to carry out a range of cellular functions. Therefore, it is imperative to conduct genome-wide studies so as to facilitate 1) effective identification of correlated genes and 2) better understanding of the mechanisms underlying gene transcription and regulation.



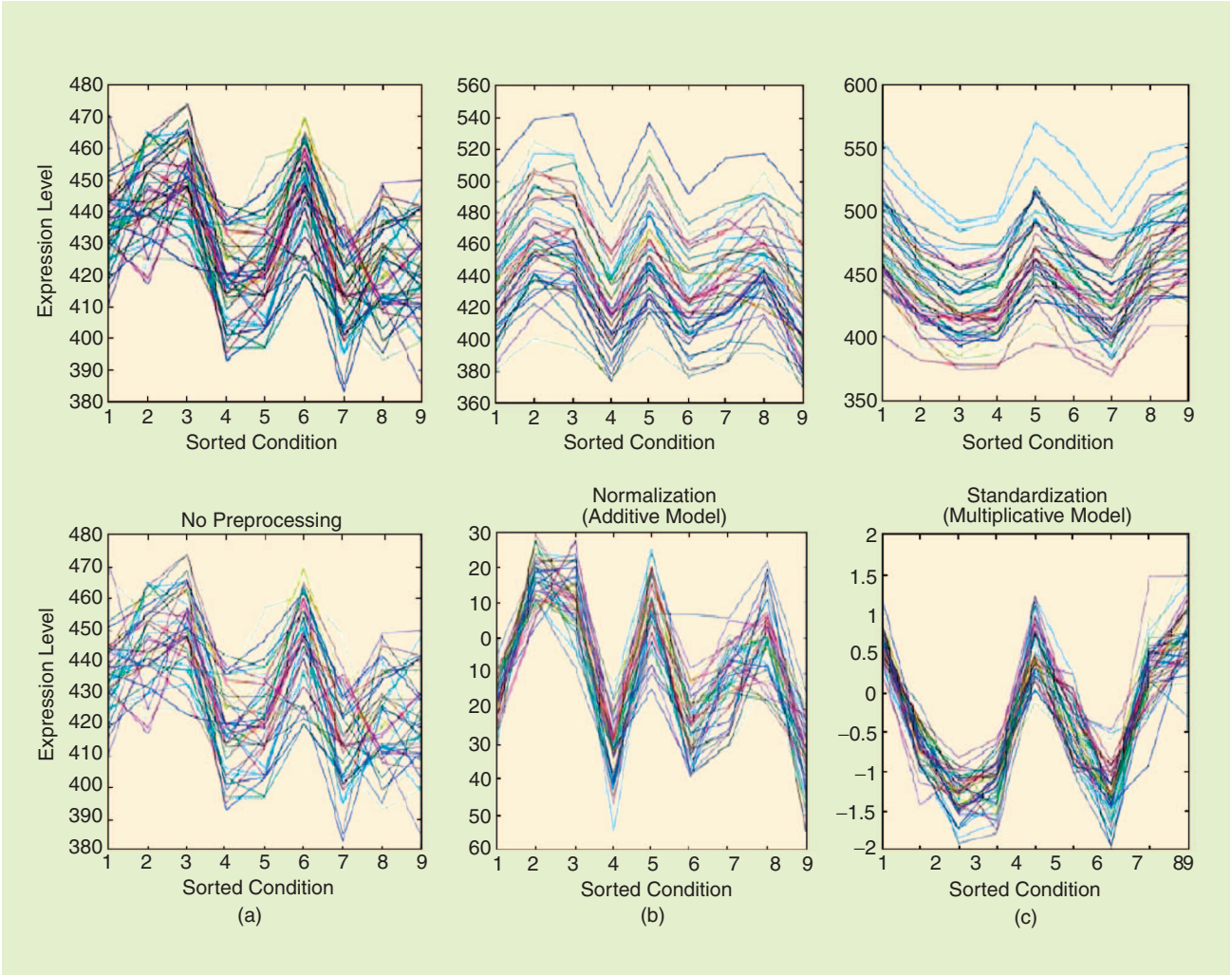
[FIG2] Figures illustrating the benefits of cascaded and parallel fusion architectures in terms of ROCs. (a) The ROC represents the prediction performance of the first stage of cascaded fusion and the arrow indicates the possible improvement due to the AND logic in Figure 1(a). For example, the AND logic in cascaded fusion (1) can be used to reduce the number of false positives, thus increasing the precision with a small sacrifice in sensitivity. (b) The ROCs represent the prediction performance of two modalities in parallel fusion (2) and the arrows indicate the possible improvement due to the AND or OR logic in Figure 1(a) and (b). (c) The ROC of nonlinear fusion which represents the most general scenario.

It is known that within the DNA sequences, there are short segments of DNA called motifs that are responsible for regulating the gene networks and the expression of genes. Therefore, the knowledge of motifs provides important information for biologists to verify and explain experimental results. However, the downside of relying only on gene sequencing (or motif discovery) for genomic scale studies lies in its extremely high computational complexity.

The expression of several thousand genes can be measured simultaneously by DNA microarrays. The upside is that gene expression analysis is computationally less demanding than sequencing. Furthermore, recent advances in machine learning tools for expression profiling have become more mature and cost effective. However, microarray data are very noisy and contain artifacts, making gene prediction very difficult.

There already exists strong evidence supporting the correlation of motifs and expression profiles [1], [8], [9]. This suggests that these two modalities are complementary to each other, making them legitimate candidates for fusion. Park et al. [8] found genes that have similar promoter regions and determined the extent to which these genes have similar expression profiles. It was concluded that, on the average, genes with similar promoter regions have significantly higher correlation, although the correlation can vary widely depending on genes.

A more promising result supporting the prediction of gene expression from DNA sequences was reported by Beer and Tavazoie [1]. In their work, gene expression patterns were predicted by looking only at their regulatory sequences. For thousands of genes, they found the set of



[FIG3] Example of feature diversity. Effect of applying preprocessing on raw expression data, including normalization (additive coherence model) and standardization preprocessing (multiplicative coherence model). In the graphs, each line represents the expression level of one gene across nine conditions in the yeast data set. (a)–(c) Three time-course gene groups. By the naked eye, it would not be very persuasive to claim that (b) forms a tighter family than (a), because the former exhibits a wider spread data structure. However, if we probe into the underlying data structure, it can then be revealed that (b) has a wider spread only because there is a large variation of the additive shifts. If such an additive variation can be properly compensated by additive preprocessing [(b), lower], the remaining spread becomes relative smaller than (a). Similarly, (c) will be tightened up if the underlying multiplicative variation is also compensated in addition to additive normalization [(c), lower].

[TABLE 1] NINE TYPES OF COHERENCE MODELS THAT CAN BE OBTAINED BY APPLYING DIFFERENT PREPROCESSING METHODS TO GENES AND CONDITIONS.

		CONDITION COHERENCE		
		NO ADJUSTMENT	NORMALIZATION	STANDARDIZATION
GENE COHERENCE	NO ADJUSTMENT	CONSTANT-VALUE	MEAN-NORMALIZATION	Z-NORM
	NORMALIZATION	MEAN-NORMALIZATION	CHENG AND CHURCH TYPE [2]	GENERALIZED Z-NORM
	STANDARDIZATION	Z-NORM	GENERALIZED Z-NORM	GENERALIZED Z-NORM

DNA sequence elements most likely to be responsible for their expression. It was concluded that 1) to a very high degree the gene expression is determined by local sequences and 2) the inferred regulatory rules can correctly predict expression patterns for 73% of genes in yeast.

Conlon et al. [3] proposed a cascaded fusion architecture in which microarray expression values are used to help screen out false positive motifs found by a motif discovery searching algorithm called Motif Discovery Scan (MDscan). The strength of correlation between gene expressions and motifs (represented by motif matching scores) is captured by linear regression coefficients, which are determined by a selective regression procedure. A larger nonzero regression coefficient suggests that the corresponding gene expression is more correlated with the upstream sequences containing the motif. This approach is able to increase both the sensitivity and specificity of the discovered motifs. Naturally, the selected motifs and the regression equations can be used to predict other genes in the same group.

In contrast, Kasturi and Acharya [5] adopted a parallel fusion architecture to combine promoter sequences of genes/DNA binding motifs and gene expression profiles. In their work, similarity scores based on the frequency of motif occurrences and expression intensities were probabilistically combined during gene cluster identification. To fuse the information embedded in sequence data and expression data, the Kohonen feature map was adopted to simultaneously update the expression profile clusters and sequence clusters.

FEATURE DIVERSITY: AN ALGORITHMIC PERSPECTIVE

Note that a crucial aspect is not only how to cluster data but also how to find an appropriate way of looking at data. Thus, extracting correct features is essential for gene prediction. We now introduce a comprehensive set of coherence models to better capture the biologically relevant features of genes to improve prediction performance. Two popular coherence models are the additive coherence model and multiplicative coherence model.

1) *Additive coherence model:* A scaling relation between mRNA_a and mRNA_b is expressed as $\text{mRNA}_b = \lambda(\text{mRNA}_a)$, where λ is a scaling factor. Denoting $a = \log(\text{mRNA}_a)$ and $b = \log(\text{mRNA}_b)$, we have $b = \lambda' + a$ where $\lambda' \equiv \log(\lambda)$.

2) *Multiplicative coherence model:* An exponential relation between mRNA_a and mRNA_b is expressed as $\text{mRNA}_b = (\text{mRNA}_a)^{\gamma}$. Now the logarithm converts the exponential changes of the relative abundance into multiplicative factors, leading to a "multiplicative model" governing dependence between a and b: $b = \gamma \times a$.

It is common to assume that the multiplicative variation is imposed on top of the additive variation. This leads to the "additive-multiplicative" coherence model: $b = \lambda' + \gamma \times a$. Figure 3 shows the effect of applying normalization preprocessing (additive coherence model) and standardization preprocessing (additive-multiplicative coherence model) on gene expression data.

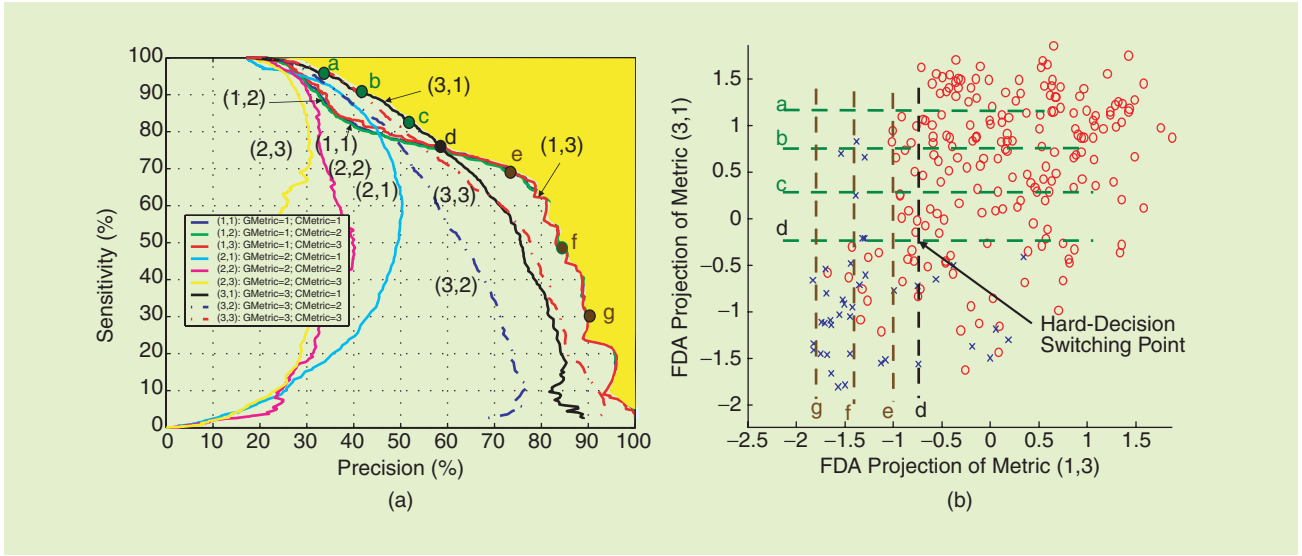
A collection of microarray experiments produces a matrix with each row representing the expression profile of a gene across different time course or independent conditions. To maximize the flexibility of discovering clusters in the matrix, it is suggested that all possible combinations of coherence models on genes and/or conditions be allowed [7]. This leads to a total of nine coherence models, as shown in Table 1.

A gene may be coexpressed via more than one coherence model. This calls for a fusion strategy that combines features corresponding to different coherence models in order to improve the prediction performance. ROCs can provide a very clear indication of which coherence models are most advantageous to fuse. For example, Figure 4(a) shows that model (3, 1) has a relatively higher sensitivity in the low-precision region but a relatively lower sensitivity in the high-precision region. In contrast, model (1, 3) has just the opposite performance. In this case, these two coherence models are truly complementary to each other and can serve as ideal fusion candidates.

A modest fusion objective is to deliver a consistent fusion result [7], which is at least as good as any of the single model results in the entire sensitivity/specificity region. As long as the sources are complementary to each other (with respect to the ROC), consistent fusion is always possible and it will yield improvement as long as certain statistical conditions are met. Such a goal can be achieved by a hard-switching fusion scheme as illustrated in Figure 4(b).

Mathematically denote the fusion score as $Z = \alpha\phi(X) + \beta\psi(Y)$ such that the fusion output is $g(Z)$ [as in (2)]. In the hard-switching scheme, we have either $\alpha = 1,\ \beta = 0$ or $\alpha = 0,\ \beta = 1$. In contrast, one may adopt a linear soft fusion scheme based on a new fusion score $Z' = \alpha'\phi(X) + \beta'\psi(Y)$, where $\alpha' + \beta' = 1$. In many cases, such a soft fusion scheme can lead to better-than-lower-bound performance. The optimal values of α' and β' can better be derived via prominent machine learning techniques, such as Fisher classifiers and support vector machines (SVMs) with a linear kernel [10]. Unfortunately, it is known that linear classifiers often have limited discriminating power.

The most flexible decision boundaries to represent a general version of (2) can be implemented by neural networks such as SVMs or decision-based neural networks (DBNNs) [6]. For the DBNN fusion classifier, there are two processing phases: 1) a local expert uses a Gaussian mixture model to



[FIG4] (a) Performance of predicting ribosomal genes in yeast using nine different combinations of preprocessing methods for the conditions and genes. In the legend, "GMetric = *m*; CMetric = *n*" means that coherence models *m* and *n* were assigned to the genes and conditions, respectively. In other words, it corresponds to Box (*m*, *n*) in Table 1. The shading area represents the region of consistent fusion. (b) Decision boundaries that lead to consistent fusion. The (horizontal) decision boundaries based on model (3, 1)—boundaries a, b, and c—have relatively higher sensitivity, while the (vertical) decision boundaries based on model (3, 1)—boundaries e, f, and g—have a relatively higher precision. Therefore, the boundaries are switched from horizontal to vertical around the crossover point. At the crossover point, boundary d can be either horizontal or vertical as they deliver exactly the same performance.

represent the patterns of the positive (or negative) class and 2) a "gating agent" fuses the local scores to reach a Bayesian optimal decision. Figure 2(c) shows the ROCs based on such a fusion scheme. Similar improvement can be found in the sensitivity-specificity ROC and the (more difficult) cluster discovery of molecular activity genes [7]. The MATLAB programs that produce the results in this paper can be found at http://www.eie. polyu.edu.hk/~mwmak/microarray.htm.

FUTURE WORK

It is our hope that this article will inspire more research into machine learning for genomic signal processing, particularly via fusion of multiple biological or algorithmic modalities. In the near term, it should be promising to look into various machine learning techniques combining these modalities to improve prediction performance. As an example, one could first apply a parallel fusion scheme to combine coherence models (1, 3) and (3, 1), with the results being further enhanced via a cascaded fusion scheme using motifs as the feature. From a long-term perspective, the proposed fusion schemes do not have to be limited to genomic data analysis. They may be further applied to, for example, the fusion of genomic, proteomic, and transcriptomic data. Before concluding, it has to be reiterated that any computational prediction of gene clusters must be ultimately verified in the laboratory. Nevertheless, it is the authors' belief that the future advance in bioinformatics will continue to depend on a close partnership with machine learning communities.

ACKNOWLEDGMENTS

This work was supported in part by the Burroughs Wellcome Fund Fellowship and the Research Grant Council of Hong Kong SAR (Project PolyU 5230/05E).

AUTHORS

Sun-Yuan Kung is with the Department of Electrical Engineering, Princeton University. *Man-Wai Mak* is with the Department of Electronic and Information Engineering, Hong Kong Polytechnic University.

REFERENCES

[1] M.A. Beer and S. Tavazoie, "Predicting gene expression from sequence," *Cell*, vol. 117, pp. 185–198, Apr. 2004.

[2] Y. Cheng and G.M. Church, "Biclustering of expression data," in *Proc. ISMB*, 2000, vol. 8, pp. 93–103.

[3] E.M. Conlon, X.S. Liu, J.D. Lieb, and J.S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis," *PNAS*, vol. 100, no. 6, pp. 3339–3344, Mar. 2003.

[4] D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics*. Los Altos, CA: Peninsula Pub., 1988.

[5] J. Kasturi and R. Acharya, "Clustering of diverse genomic data using information fusion," *Bioinform.*, vol. 21, no. 4, pp. 423–429, 2005.

[6] S.Y. Kung, M.W. Mak, and S.H. Lin, *Biometric Authentication: A Machine Learning Approach*. Upper Saddle River, NJ: Prentice-Hall, 2005.

[7] S.Y. Kung, M.W. Mak, and I. Tagkopoulos, "Symmetric and asymmetric multi-modality biclustering analysis for microarray data matrix," *J. Bioinform. Computat. Biol.*, vol. 4, no. 3, June 2006.

[8] P.J. Park, A.J. Butte, and I.S. Kohane, "Comparing expression profiles of genes with similar promoter regions," *Bioinform.*, vol. 18, no. 12, pp. 1576–1584, 2002.

[9] F.P. Roth, J.D. Hughes, P.W. Estep, and G.M Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnol.*, vol. 16, pp. 939–945, Oct. 1998.

[10] V.N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

SP