BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 2 2007, pages 198–206
doi:10.1093/bioinformatics/btl553

Systems biology

Dependence network modeling for biomarker identification

Peng Qiu[1,*], Z. Jane Wang[2], K. J. Ray Liu[1], Zhang-Zhi Hu[3] and Cathy H. Wu[3]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, USA,
[2]Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada and
[3]Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, Washington DC, USA

Received on July 18, 2006; revised on October 17, 2006; accepted on October 24, 2006
Advance Access publication October 31, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Our purpose is to develop a statistical modeling approach for cancer biomarker discovery and provide new insights into early cancer detection. We propose the concept of dependence network, apply it for identifying cancer biomarkers, and study the difference between the protein or gene samples from cancer and non-cancer subjects based on mass-spectrometry (MS) and microarray data.

Results: Three MS and two gene microarray datasets are studied. Clear differences are observed in the dependence networks for cancer and non-cancer samples. Protein/gene features are examined three at one time through an exhaustive search. Dependence networks are constructed by binding triples identified by the eigenvalue pattern of the dependence model, and are further compared to identify cancer biomarkers. Such dependence-network-based biomarkers show much greater consistency under 10-fold cross-validation than the classification-performance-based biomarkers. Furthermore, the biological relevance of the dependence-network-based biomarkers using microarray data is discussed. The proposed scheme is shown promising for cancer diagnosis and prediction.

Availability: See supplements: http://dsplab.eng.umd.edu/~genomics/dependencenetwork/

Contact: qiupeng@umd.edu

1 INTRODUCTION

In genomics studies, great efforts have been made to develop the gene regulatory network using microarray gene expression data, as reviewed in Someren *et al*., 2002. Recently, it is believed that it is the proteomic data and the collective functions of proteins that directly dictate the phenotype of the cell and, thus, are more accurate in interpreting the cause of biological phenomenon. Therefore, proteomics, the large-scale study of protein function and expression, is an emerging field for the discovery and characterization of regulated proteins or biomarkers in different diseases in the post-genome era. During cancer development, the cancerous cells may release unique proteins and other molecules, which may be regarded as early biomarkers. Here biomarkers are defined as the alternations of patterns at the cellular, molecular or genetic level. Caused by the presence of specific diseases, these biomarkers normally serve as the indicators of diseases. Correctly identifying protein biomarkers for cancer holds enormous potential for the early detection of cancer and effective treatments. However, due to the complicate nature of protein functions, it is a research topic with significant challenge.

For the analysis of protein samples, mass spectrometry (MS) technologies have become increasingly important tools (Diamandis, 2004). MS is able to convert proteins or peptides to charged pieces that can be separated on the basis of the mass-to-charge ratio (m/z). There are several types of MS ionization methods currently available, (Budzikiewicz, 2005), including surface enhanced laser desorption ionization (SELDI), electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). The produced protein or peptide spectra are then analyzed for different purposes, such as identifying proteins via peptide mass fingerprints, cancer classification, etc. Until very recently, it has also been applied for cancer biomarker identification, but only simple classification-based approaches were studied. For instance, in Li *et al*. (2002), a panel of three biomarkers were selected using the linear combination based on unified maximum separability analysis (UMSA) to best separate cancer and normal samples.

In Qiu *et al*. (2005), we developed an ensemble dependence model for cancer classification based on microarray gene expression data. It is noted that the proposed method yields very promising classification performance in gene expression data. To further explore the dependence model, in this paper, we present the concept of dependence network, and apply it for biomarker identification. The dependency revealed by the dependence network provides some insight into the functional interaction relationships between genes and proteins. With the dependence network, we can explore the functionalities of the underlying biological system as a whole. The learned dependence network may play an indispensable role in understanding the underlying system, especially as the starting point to further interpret the behaviors and properties of the system. For instance, in this study, as we examine how the dependence network structure evolves when cancer develops in different stages, we could gain insights into the mechanisms of cancer development. In addition, since finding accurate cancer biomarkers is of crucial importance to early diagnosis and effective treatments, to address the problem of identifying biomarkers, we propose to construct the dependence networks under different cases (e.g. cancer, normal, different cancer stages) and identify cancer related biomarkers based on these constructed dependence networks.

This paper is organized as follows. In Section 2, we describe the basic concepts of the dependence model and present the idea of dependence network. Then, in Section 3, the classification-performance-based biomarkers and dependence-network-based

*To whom correspondence should be addressed.

biomarkers are examined based on protein MS datasets. In Section 4, a gene microarray dataset for gastric cancer is examined in detail to show the applicability of the dependence network for genes and to demonstrate the biological evidence which supports the proposed algorithm. The biological relevance of identified biomarkers is discussed. Finally, the conclusions are presented in Section 5.

2 DEPENDENCE MODEL AND DEPENDENCE NETWORK

As mentioned earlier, the concept of ensemble dependence model (EDM) for cancer classification using microarray gene expression data is proposed in Qiu *et al*. (2005). In this section, we first review the dependence model. Then we will discuss the eigenvalue pattern of the dependence model, and focus on the concept of dependence network.

2.1 Dependence model

The dependence model focuses on exploring and modeling the group dependence relationship. Given several gene/protein groups or a set of individual genes/proteins, we regard each group or each individual as one feature. Without any prior knowledge, we assume that each feature is, to some extent, dependent on all the other features. Linear dependence relationship is studied, where each dependence relationship is described by a weight a_{ij}. The so-called self-regulation is assumed to be zero, e.g. $a_{ii} = 0$, $i = 1, 2, 3$.

Suppose there are M features in total, the dependence relationships between the M features can be expressed as the following linear equations, for $i = 1, \ldots, M$:

$$x_i = \sum_{j \in \{1, \cdots, M\}, j \neq i} a_{ij} x_j + n_i, \tag{1}$$

where, a_{ij} form the dependence matrix; x_i, $i = 1, 2, \ldots, M$, are the features' expression data. There is a noise-like term, which could be contributed by the model mismatch and the measurement uncertainty from experiment. The dependence matrix and statistics of the noise term can describe the expression data, and thus be used to distinguish cancer and normal samples.

For the purpose of classification, given selected features, the dependence matrices and the statistics of the noise-like terms can be estimated from cancer and normal training samples, respectively. The estimated normal and cancer dependence models form a supervised classifier which can then be used for classification. For each testing sample, the maximum likelihood (ML) decision rule is applied to predict whether it is cancer or normal, that is whether the testing sample fits the cancer model better or fits the normal model better. (Detail descriptions of estimating the dependence model and classification can be found in supplements.)

2.2 Eigenvalue pattern of dependence model

In the previous subsection, the dependence model is described by Equation (1). The dependence matrix and noise statistics are used to model the expression data. However, it is difficult to tell the dependence relationship from the model parameters. Therefore, we propose to use an easy term, the eigenvalues, to describe the dependence relationship.

In the dependence model, the ideal case is defined when the noise-like term is zero in Equation (1), meaning the features' expression profiles are completely linearly dependent. In this ideal case, taking the case of three features, for example, the dependence matrix will have a special structure as

$$\mathbf{A}_{\mathbf{ideal}} = \begin{bmatrix} 0 & \alpha_1 & \alpha_2 \\ \frac{1}{\alpha_1} & 0 & -\frac{\alpha_2}{\alpha_1} \\ \frac{1}{\alpha_2} & -\frac{\alpha_1}{\alpha_2} & 0 \end{bmatrix}, \tag{2}$$

for some non-zero numbers α_1 and α_2. It is proved that the eigenvalues of the above matrix are $1,1,-2$, no matter what are the values of $\alpha_i, i = 1, 2$. For a more general case where we have M features, we note that the eigenvalues of the M-by-M matrix $\mathbf{A}_{\text{ideal}}$ are always $\{1,1,\ldots,1,-(M-1)\}$. (proof in supplements)

Based on the ideal case, we gradually introduce larger and larger random variations, to make the features' expression profiles more and more independent. We examine how the eigenvalue pattern changes as the features' expression become more and more independent. Take the case of three features as example, to simulate the ideal case, an artificial dataset is generated, which contains 3 features and 100 samples. The expression of the three features are linearly dependent. As mentioned above, the corresponding dependence matrix will have a special structure, as in Equation (2), with eigenvalues $1,1,-2$. Then, random variation is added to the ideal expression data, to make them more and more independent. The variation level is controlled by the variation to signal ratio, the energy of the added variation over the energy of the ideal expression. At each variation level, 1000 different realizations of the random variation are added to the ideal case. For each realization, the dependence matrix is estimated, and the eigenvalues of the dependence matrix are calculated. The mean and standard deviation of each eigenvalue (the largest eigenvalue, the smallest eigenvalue and the middle eigenvalue) are calculated at each variation level. In Figure 1, the mean of each eigenvalue is plotted in the solid lines, while the dashed lines represent the mean ± standard deviation.

As shown in Figure 1, from the ideal case, as the features' expression profiles suffer more and more random variations, as the features' expression become more and more independent, the eigenvalues of their dependence matrix will change and follow the trends in Figure 1. It is noticed that, the eigenvalue pattern is closely related to the dependence relationship, especially the smallest eigenvalue. In this three-features example, when the expression profiles are linearly dependent, the smallest eigenvalue is -2. When the dependence relationship become weaker and weaker, the smallest eigenvalue increases, and eventually saturate to ~ -0.7. Furthermore, the small standard deviation indicates that the eigenvalue pattern is a very consistent indicator of the dependence relationship. Therefore, we can use the smallest eigenvalues to describe the strength of dependence relationship among examined features, meaning how dependent they are, or how closely related they are.

2.3 Dependence network

The functionality of a protein is not solely characterized by its own structure. Its surroundings and interacting proteins also play important roles in determining the protein's function. In short, the protein interaction network can provide detailed functional insights of proteins. Moreover, the protein interaction network is also the basis for finding biological signaling pathways for diseases, which are important in understanding the mechanism of diseases (Walhout and Vidal, 2001). In this study, we propose to apply the dependence model for dependence network construction.

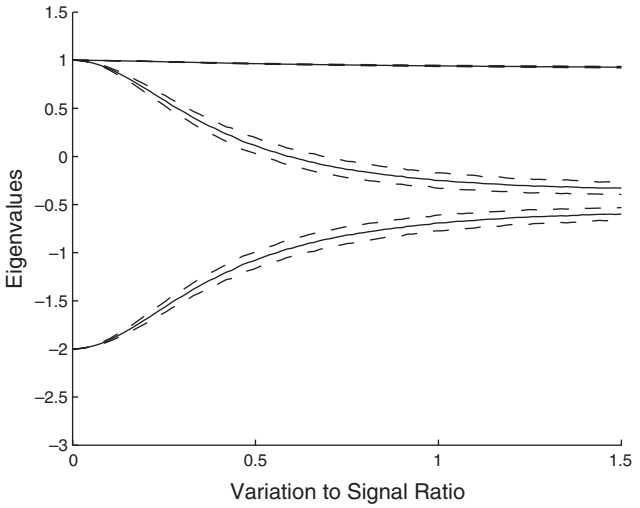

Fig. 1. The horizontal axis is variation level, which indicates how noisy the three cluster expression profiles are. As the features' expression profiles become more noisy, the eigenvalues of the corresponding dependence matrix will change, following the above curves.

A dependence network is a set of components, such as protein MS peak features in our study, and linear dependence interactions among them that collectively carry out specific functions. In the dependence network, each connection represents an inter-component dependence relationship, with an associated weight indicating to what extent the connected components are related. In the following, we describe how a dependence network is constructed.

Since, the eigenvalue pattern is a consistent indicator of the dependence relationship, if we examine three individual MS features at one time, through an exhaustive search, we can find all closely related feature triples, the 'binding triples'. The elements in each binding triple share a strong dependence relationship, which indicates that they have a strong influence on one another in the protein interaction network. Take an ovarian cancer MS dataset as an example. For the normal case, we pick a subset of normal samples, examine all possible feature triples, estimate a dependence matrix and calculate the eigenvalue pattern for each feature triple. A threshold -1.3 is applied. If the smallest eigenvalue of a feature triple is lower than the threshold, there exists a strong dependence relationship within the triple. We call this kind of triples the 'binding triples'. Similar analysis is applied to cancer samples. In the normal case, 512 triples pass the threshold; while in the cancer case, 436 triples pass the threshold. Moreover, there are only 31 triples in the overlap between normal and cancer cases. The results suggest that, from healthy to cancerous, some dependence relationships among proteins are disabled; while some other dependence relationships are activated. The small overlap indicates that, from healthy to cancerous, the overall dependence relationship goes through a major change.

The dependence network is constructed from the binding triples. As in graph theory, the topology of an n-node network can be represented by an $n \times n$ adjacency matrix D. If feature i and feature j both appear in a binding triple, it is suggested by the dependence model that feature i and feature j are closely related. And we will count once for D_{ij}, the connection between feature i and feature j. Basically, we look at the binding triples, count the appearance of all feature pairs, and form an adjacency matrix D. Then, the adjacency matrix D is normalized by the total number of binding triples. Each element D_{ij} is a confidence value, which indicates the importance and strength of the connection between feature i and feature j. We call network D the dependence network. For the purpose of biomarker identification, we propose a simple way below to detect biomarkers which represent the largest topology change under different situations. The dependence networks of normal case (D_{normal}) and cancer case (D_{cancer}) can be built based on normal samples and cancer samples, respectively. Note that each element of the matrix $D_{\text{normal}} - D_{\text{cancer}}$ corresponds to the change of weight between a pair of nodes, and each column (or row) corresponds to the change of all the connections related to one node. Therefore, mathematically, by examining the norm of all the columns (or rows) of $D_{\text{normal}} - D_{\text{cancer}}$, we are able to see which features go through a large topology change from normal to cancer and, thus, are potentially biomarkers. We call them the dependence-network-based biomarkers. For the purpose of visualization, the dependence networks can be presented as shown in Figure 2, where strong dependence relationship is reflected in small distance between connected nodes. The length of each connection is defined to be inversely proportional to the confidence value. Because the confidence values are normalized, through $1/D_{ij}$, features with strong dependence relationship will locate close to each other, while features with weak dependence relationship will be far apart. From Figure 2, we are able to visually identify important core nodes which are indicated by drawing circles around. In the following section, we can see that both the mathematical identification criterion and the visual inspection yield similar biomarkers.

3 BIOMARKER IDENTIFICATION

In this section, there are three protein mass spectrum datasets under investigation, one ovarian cancer dataset, with 25 normal samples and 24 cancer samples (Tibshirani *et al*., 2004), one prostate cancer dataset, with 81 normal samples, 84 early stage cancer samples and 84 late stage cancer samples (Adam *et al*., 2002), and one liver cancer dataset, with 176 cancer samples and 181 normal samples (Ressom *et al*., 2005). (Data is available in supplemental website). All three MS datasets are examined in detail. However, due to page limit, only the results for the ovarian cancer dataset and prostate cancer dataset are shown in this section. Other results can be found in the supplements. Because of the noisy nature of mass spectrum (MS) datasets, proper preprocessing of MS data is needed before analysis. The details of preprocessing is available in the supplements. After preprocessing, peaks in the mass spectra are identified as features for further analysis. Since not all peak features are informative in understanding the difference between cancer and normal samples, feature selection is performed to exclude irrelevant peaks. We apply the selection criterion used in Gloub *et al*., 1999. In the following discussion, the 50 top-score peak features are examined for biomarker identification.

3.1 Classification-performance-based biomarkers

In our early works, the concept of ensemble dependence model was applied to classify microarray gene expression data, yielding excellent classification performance. In this study, we apply the dependence model to build a supervised classifier, examine individual

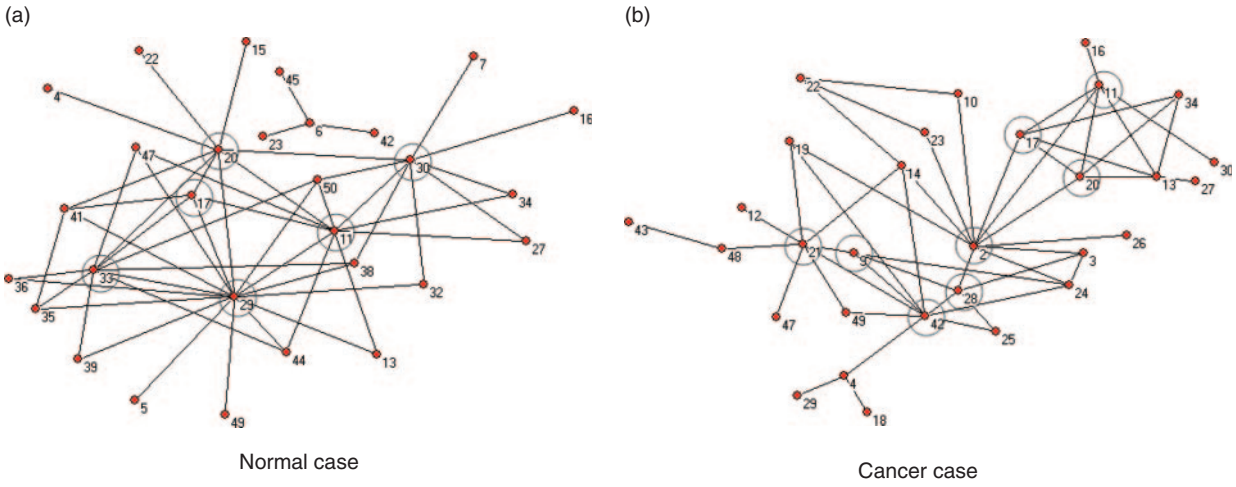

Fig. 2. Dependence networks for normal and cancer cases in the ovarian cancer MS dataset. (Isolated nodes are omitted.) For the purpose of illustration, the circles are used to indicate the core features, which are obtained through visual inspection.

protein mass features, and use their classification performance as biomarker identification criterion. (details available in supplements)

We examine three features at one time, and apply the dependence model for classification. Through an exhaustive search, all possible feature triples are examined, and the classification performance is recorded as a metric. Triples with classification accuracy >95% are considered to be informative triples. Features that frequently appear in the informative triples are regarded as important cancer biomarkers. These are biomarkers identified based on the criterion of classification performance. We call them the classification-performance-based biomarkers.

First, we examine the ovarian cancer MS dataset. To ensure reproducibility of the identified biomarkers, we apply a strategy similar with 10-fold cross-validation, where the ovarian cancer dataset is divided into 10 parts; 9 parts are used for model learning (training) and the one left is used for validation (testing). In each of the 10 iterations, we search for biomarkers based on different choices of training and testing samples. For each iteration, through an exhaustive search, the classification performance of all possible feature triples are examined to find informative triples, and the top 10 highest frequently appeared features are considered as biomarkers. Therefore, in each of the 10 iterations, based on different training and testing set, 10 biomarkers are identified. We examine the biomarkers identified by different data to assess the consistency of the identification criterion. The result is that, only three features are commonly identified as biomarkers by $\geq$7 out of the 10 iterations. Figure 3a shows the histogram of the identified biomarkers, where the horizontal axis is the feature indexes, and the vertical axis shows how many times one feature is identified during the 10-fold iterations. From the widely spread histogram, we can conclude that the result is not quite consistent.

We further examine the prostate MS dataset for two cases: normal samples versus early stage cancer samples, and normal samples versus late stage cancer samples. Our main purpose in analyzing this dataset is to examine the possible difference between dominant biomarkers in early cancer stage and late cancer stage. Then we examine the liver cancer MS dataset. Similar with above analysis, 10-fold cross-validation is applied. Again, every iteration, top 10 features that most frequently appeared in informative triples are considered as biomarkers. The histograms of identified biomarkers are shown in supplements. From the results, we again observe that the classification-performance-based criterion lacks consistency under 10-fold cross-validation.

3.2 Dependence-network-based biomarkers in ovarian cancer dataset

In the ovarian cancer dataset, we examine the selected 50 features three at one time. Through an exhaustive search, the dependence relationship of all feature triples are examined to find binding triples. From normal samples, the binding triples of normal case are found, and we build a dependence network for the normal case D_{normal}. From cancer samples, the binding triples of cancer case are found, and we build a dependence network for the cancer case D_{cancer}. By examining the norm of all the columns of the matrix $D_{\text{normal}} - D_{\text{cancer}}$, we are able to see which features go through a large topology change from normal to cancer, and identify them as dependence-network-based biomarkers. Similar to the previous subsection, 10-fold cross-validation is conducted. For each iteration, D_{normal} and D_{cancer} are calculated and compared, and 10 features with large topology changes are considered as biomarkers. Ten features are commonly identified as biomarkers by $\geq$7 out of the 10 iterations. They are features 2, 11, 17, 20, 21, 28, 29, 33, 42 and 49. Figure 3b shows histogram of the identified biomarkers. From this figure, we can see that the dependence-network-based criterion yields much more consistent results, compared with the classification-performance-based criterion. Another observation is that, if we apply a simple differential method, such as *t*-test, for biomarker identification, the identified biomarkers will be features with indexes $\sim$40–50 (since the pre-selection 50 features are based on *t*-test). From Figure 3, we can see that the classification-performance-based biomarkers have high correlation with the simple differential method. However, the dependence-network-based criterion identifies many biomarkers that are not simply the most differentially expressed features. The results indicate that, the dependence-network-based biomarker identification criterion

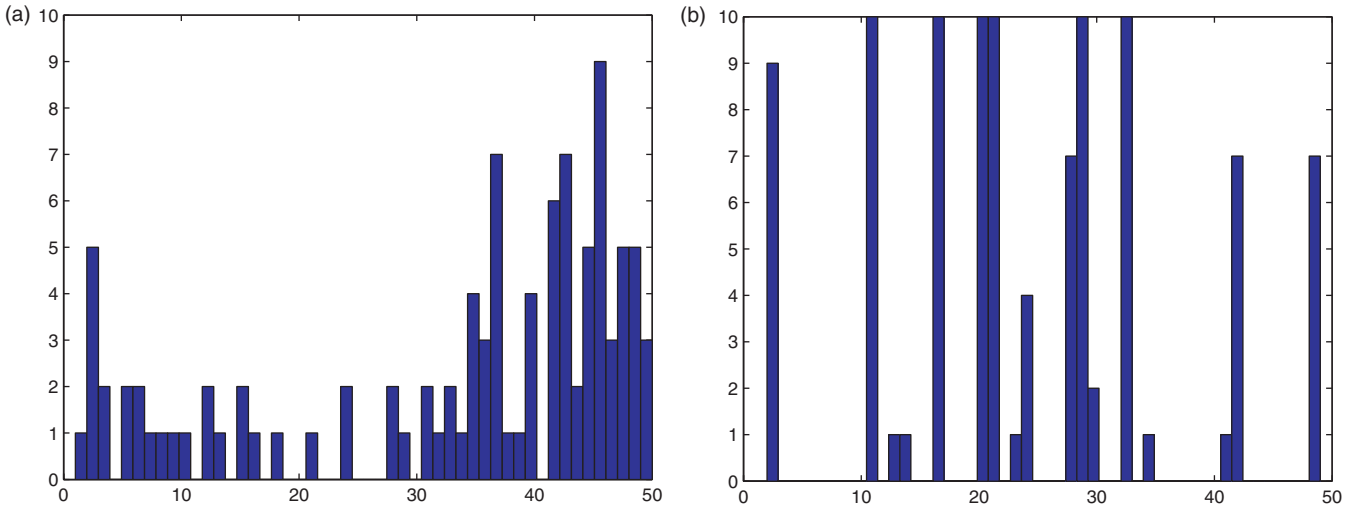

Fig. 3. (a) Shows the histogram of performance-based biomarkers in the ovarian cancer dataset. (b) Shows the histogram of network-based biomarkers of the ovarian cancer dataset. In both figures, the horizontal axis is the feature indexes, and the vertical axis shows how many times one feature is identified during the 10-fold iterations. From these figures, we can see that the network-based criterion yields more consistent results than the performance-based criterion.

yields much more information than the simple differential method and the performance-based criterion.

In Figure 2, the dependence networks for normal and cancer cases are drawn, where we can see the important features in the normal and cancer dependence networks through visual inspection. In the normal case, features 11, 17, 20, 29, 30 and 33 are important core features. They have rich dependence relationships with lots of other features. However, in the cancer case, there are more core features 2, 9, 11, 17, 20, 21, 28 and 42. From normal case to cancer case, some unimportant features in normal case become core features in cancer case, especially features 2, 21, 28 and 42; while some core features in normal case become deactivated in cancer case, such as features 29, 30 and 33. These core features are strongly suggested to be biomarkers in ovarian cancer. It is our intention to investigate the origin and identity of these features.

3.3 Dependence-network-based biomarkers in prostate cancer dataset

We further examine the prostate MS dataset. From binding triples of samples from normal, early cancer stage, and late cancer stage, we build dependence networks D_{normal}, D_{early} and D_{late}, respectively. Based on D_{normal} and D_{early}, we identify biomarkers for early stage cancer samples; based on D_{normal} and D_{late}, we identify biomarkers for late stage cancer samples. In supplements, we show the histograms of the identified biomarkers under 10-fold cross-validation. Consistent with the results in the ovarian cancer dataset, the dependence-network-based criterion gives more consistent results for both early stage cancer case and late stage cancer case than the classification-performance-based criterion. Compared with a simple differential method, such as *t*-test, the dependence-network-based criterion yields more information than the classification-performance-based criterion.

The dependence networks for normal, early cancer stage and late cancer stage are drawn in Figure 4. From this figure, we can see some interesting behaviors of the identified dependence-network-based biomarkers through visual inspection. For example, feature 34 is not important in normal stage. However, in cancer stages, it plays a more important role in the dependence network. Features 20 and 24 are more interesting. They are important network nodes in both normal stage and late cancer stage. However, they are deactivated in early cancer stage. Features 12, 13 and 16 behave oppositely: they are activated in early cancer stage only. These features might be the key to early stage cancer development, and deserve to be further investigated.

4 DEPENDENCE NETWORK FOR BIOMARKER IDENTIFICATION IN MICROARRAY DATA

In this section, to illustrate the generality of the proposed schemes and to demonstrate the biological significance and evidence of the identified biomarkers in caner pathogenesis and clinical applications, we examine the gene microarray expression data. In this section, a gastric cancer microarray dataset (Chen *et al*., 2003) and a liver cancer microarray dataset (Chen *et al*., 2002) are studied to examine the performance of the two proposed biomarker identification schemes. Similar with the analysis of the protein MS datasets, 50 gene features are selected by the selection criterion in Golub *et al*. (1999). Biomarkers are identified from the 50 top-score genes.

For the gastric cancer microarray dataset, in order to identify the classification-performance-based biomarkers, we exhaustively examine all possible feature triples, and apply the dependence model for classification. Triples with classification accuracy >95% are considered to be informative triples. Gene features that frequently appear in the informative triples are regarded as cancer biomarkers. 10-fold cross-validation is applied to examine the consistency of the identified biomarkers. In each of the 10 iterations, 10 biomarkers are identified based on different training and testing sets. The histogram of the identified biomarkers are shown in Figure 5a. Only one feature is commonly identified as biomarkers by $\geq$7 out of the 10 iterations. The widely spread histogram shows the lack of consistency of classification-performance-based criterion in the gastric gene microarray data.

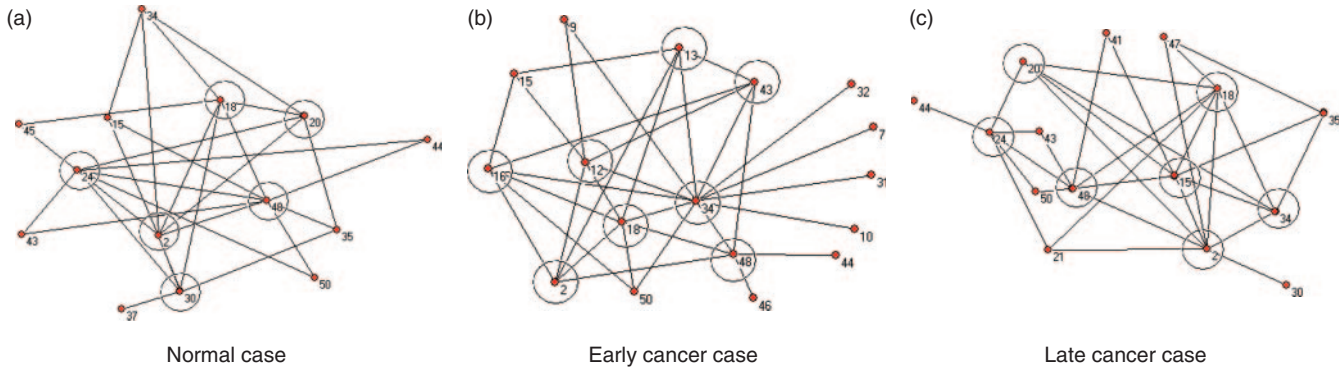

Fig. 4. Dependence networks for the prostate cancer dataset: normal, early and late cancer cases. Isolated nodes are omitted for simplicity. For the purpose of illustration, the circles are used to indicate the core features, which are obtained through visual inspection.

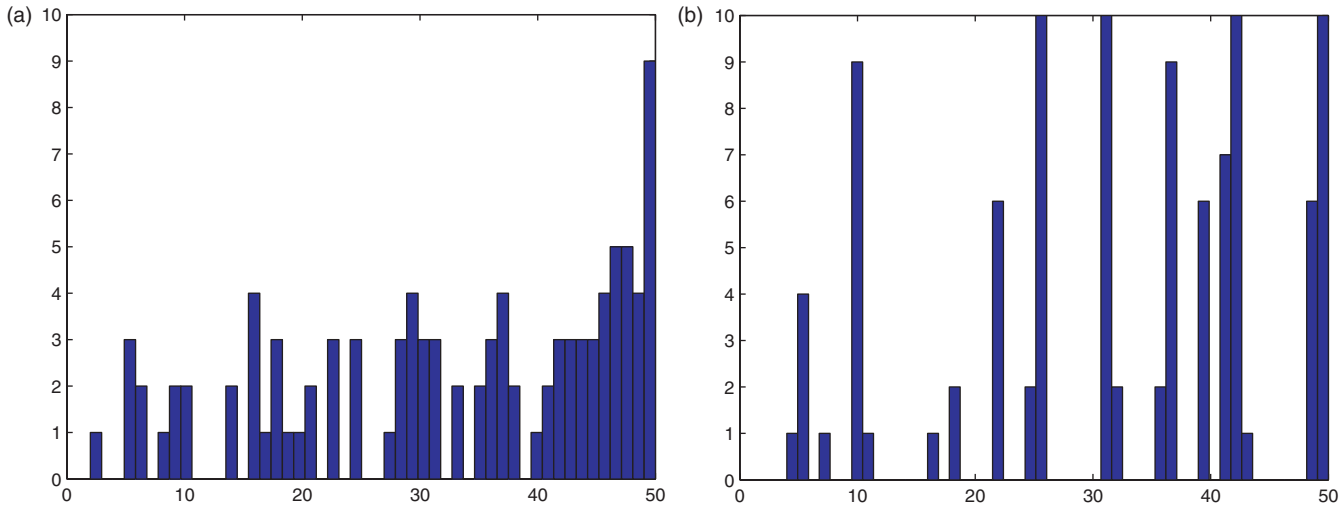

Fig. 5. (a) Shows the histogram of performance-based biomarkers in the gastric cancer microarray dataset. (b) Shows the histogram of network-based biomarkers of the gastric cancer microarray dataset. From this figure, we can see that the network-based criterion yields more consistent results than the performance-based criterion.

The dependence-network-based criterion is also examined under 10-fold cross-validation. For each of the 10 iterations, from a subset of normal samples, we build a dependence network for normal case D_{normal}; from a subset of cancer samples, we build a dependence network for cancer case D_{cancer}; then, biomarkers are identified based on the difference between D_{normal} and D_{cancer}. As shown in Figure 5b, seven features are commonly identified as biomarkers by ≥ 7 out of the 10 iterations. They are features $10, 26, 31, 37, 41, 42$ and 50. From this figure, we again observe that the network-based criterion yields much more consistent results than the classification-performance-based criterion. Also, the dependence-network-based criterion yields more information than the classification-performance-based criterion, with respect to the simple differential method *t*-test. Compared with the results from protein MS data, the results from microarray show less consistency. This may be because gene microarray experiments have larger noise than the protein MS experiments.

From Figure 6, we can see the important features in the normal and cancer dependence networks through visual inspection. In the normal case, features 8 and 50 are important core features. However, in the cancer case, there are much more core features 10, 26, 31, 37, 41 and 42. From normal case to cancer case, some unimportant features in normal case become core features in cancer case, while some core features in normal case become deactivated in cancer case. These gene features are strongly suggested to be biomarkers in gastric cancer.

Below we compare the biomarkers we identified with the hierarchical clustering result of the original study (Chen *et al*., 2003) and discuss their biological significance in gastric caner pathogenesis and clinical applications.

The 50 top-score genes we analyzed represent the most significant changes of gene expression patterns across different cancer pathological types, and correspond to four distinct gene clusters in the hierarchical clustering result (Chen *et al*., 2003). Of the 50 genes, 7 are consistently identified as biomarkers during the 10-fold cross-validation in our study. Table 1 summarizes the function of the seven genes, six with significantly increased expression levels and one with decreased expression. Interestingly, the six up-regulated genes all correspond to the same ECM (extracellular matrix) cluster, which has highly similar expression pattern across

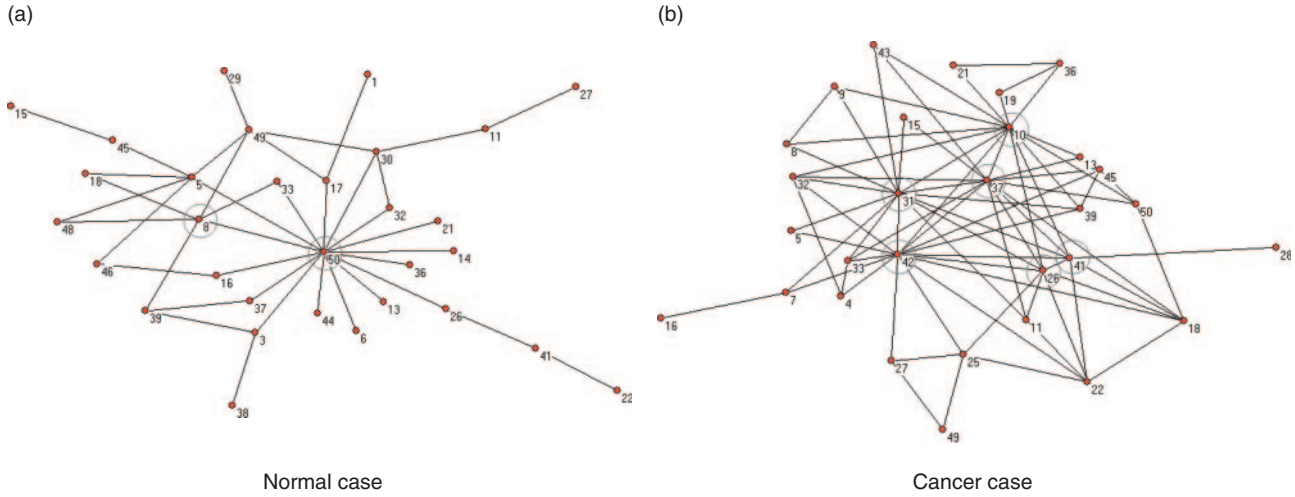

Fig. 6. Dependence networks for normal and cancer cases in the gastric cancer microarray dataset. (Isolated nodes are omitted.) For the purpose of illustration, the circles are used to indicate the core features, which are obtained through visual inspection.

Table 1. Identified biomarkers based on dependence network modeling for gastric cancer

Gene name	Protein name [UniPortKB accession]	Feature (node)	Expression level in cancer samples	Function
SPARC	Osteonectin, SPARC precursor [P09486]	42	Up	Regulate cell growth through interactions with the extracellular matrix and cytokines
COL3A1	Type III collagen alpha-1 chain precursor [P02461]	26	Up	Components of most soft connective tissues along with type I collagen
SULF1	Extracellular sulfatase Sulf-1 precursor [Q8IWU6]	50	Up	Exhibits arylsulfatase activity and highly specific endoglucosamine-6-sulfatase activity
YARS	Tyrosyl-tRNA synthetase, cytoplasmic (TyrRS) [P54577]	10	Up	Protein synthesis; N- and C- terminal fragments exert cytokine activities
ABCA5	ATP-binding cassette A5 [Q8WWZ7/Q9NY14]	41	Up	A member of ABC transporters, reside in lysosome, its knockout mice develop lysosomal disease-like symptoms
THY1	Thy-1 membrane glycoprotein precursor [P04216]	31	Up	May play a role in cell-cell or cell-ligand interactions during synaptogenesis; also involved in maintenance of T cell homeostasis and T cell responses
SIDT2	SID1 transmembrane family member 2 precursor [Q8NBJ9]	37	Down	Multi-transmembrane proteins, involved in siRNA uptake into cells

The marker genes are mapped to the protein accession numbers in UniProt Knowledgebase (UniProtKB) (Wu *et al*., 2006).

most pathological types. The down-regulated *SIDT*2 gene, on the other hand, belongs to a cluster with no assigned function (see supplements).

The ECM cluster of genes, including many that encode extracellular matrix components, tends to be more highly expressed in tumors of the diffuse histological type than in those of the intestinal type. This is consistent with greater propensity of this group of tumors for invasive growth, often provoking a dense fibrous reaction, and a reflection of reciprocal interactions between tumor and stromal cells that play important roles in tumor biology (Chen *et al*., 2003). In fact, three of the six biomarker genes we identified (*SPARC*, *COL3A*1 and *THY*1) encode proteins of extracellular matrix component or of mediating cell-matrix interactions. In addition, *SULF*1 and *YARS* are either extracellular sulfatase or secreted cytokine and both are implicated in tumor growth and progression.

Osteonectin, also known as SPARC, is a non-structural component of extracellular matrix-associated matricellular glycoprotein. Matricellular proteins mediate interactions between cells and their extracellular environment. Osteonectin is involved in the regulation of tumor cell growth, differentiation and metastasis. It is produced at high levels in many types of cancers, especially by cells associated with tumor stroma and vasculature (Framson *et al*., 2004). Osteonectin was suggested as a prognostic marker for several cancers, including invasive differentiated stomach adenocarcinoma (Maeng *et al*., 2002b), gastric cancer (Inoue *et al*., 2002), and

malignant melanoma (Bosserhoff, 2006), and was correlated with metastasis in prostate cancer (Thomas *et al*., 2000). Furthermore, osteonectin and type III collagen alpha-1, another marker gene predicted by the dependence network, were highly expressed in gastric cancer tissue (Hippo *et al*., 2002). Marked increases in expression of osteonectin and six other extracellular matrix proteins, including collagen type III, were also observed in rat gastric cancer models (Maeng *et al*., 2002a).

SULF1 is an extracellular endosulfatase that desulfates cell surface heparan sulfate proteoglycans (HSPG), thus regulating the cellular signaling cascades. Dynamic regulation of HSPGs by sulfatases within the tumor microenvironment can have a dramatic impact on the growth and progression of malignant cells. SULF1 has been implicated in promoting cell proliferation in bladder cancer and repression of differentiation in the muscle-invasive tumors, and was suggested as one of the top predictors for the bladder cancer outcome (Blaveri *et al*., 2005). SULF1 was also shown to inhibit tumor growth in hepatocellular carcinoma (Lai *et al*., 2006).

The human tyrosyl-tRNA synthetase (TyrRS) is a synthase that produces two distinct cytokines from the N- and C-terminal fragments (Wakasugi and Schimmel, 1999). It may be involved in a coordinated mechanism for regulating angiogenesis with a related synthetase, tryptophanyl-tRNA synthetase (TrpRS), which also generates two fragments in a similar fashion. While fragments of TyrRS stimulate angiogenesis, those of TrpRS inhibit this process (Tzima and Schimmel, 2006). TyrRS and TrpRS are proinflammatory cytokines with multiple activities during apoptosis, angiogenesis and inflammation. They also play important roles in cancer progression, modulating tumor angiogenesis and its escape from surveillance by immune system (Ivakhno *et al*., 2004).

ABCA5 is a transmembrane protein in the ABC transporter family, and has been shown to reside in lysosomes. *ABCA*5 gene knockout mice develop lysosomal disease-like symptoms (Kubo *et al*., 2005). ABCA5 was also identified as a tissue and urine diagnostic marker for prostate intraepithelial neoplasia.

Thy-1 (CD90) is a small GPI-anchored protein abundant on the surface of mouse thymocytes and peripheral T cells. Thy-1 is involved in the maintenance of T cell homeostasis in the absence of TCR triggering, as well as potentiating antigen-induced T cell responses induced through TCR (Haeryfar and Hoskin, 2002). Thy-1 is also an important regulator of cell–cell and cell–matrix interactions, with important roles in nerve regeneration, metastasis, inflammation and fibrosis (Rege and Hagood, 2006).

The only down-regulated marker gene is *SIDT*2, which is a cell membrane protein that enhances cell uptake of small interfering RNA (siRNA) (Duxbury *et al*., 2005), resulting in increased siRNA-mediated gene silencing efficacy. However, its cellular functions and roles in cancer are unclear. As a central node in the dependence network (node 37 in Fig. 6), the cellular functions and roles of SIDT2 in gastric cancer are worth further investigation.

Taken together, the seven gastric cancer biomarker genes that are consistently identified by the dependence network modeling approach have been shown to be biologically relevant in gastric and other cancers. Of special note is that both SPARC and COL3A1 are concurrently observed in this study (as connected core nodes 42 and 26 in Fig. 6) as well as in several other studies as valuable biomarkers for gastric cancers. We therefore conclude that our network modeling approach have provided a novel and consistent mathematic model to define potential cancer biomarkers, which imply functional associations or interactions that are important for the underlying cancer biology.

The liver cancer microarray dataset is also examined, with the detail results given in the supplements. The learned dependence networks for normal and cancer cases are shown, and we are investigating the biological significance and evidence of the identified liver-cancer biomarkers.

5 CONCLUSION

In this study, we propose to construct dependence networks between protein or gene features. In building the dependence network, the dependence relationship among features can be indicated by the eigenvalue pattern. From binding triples found via the desired eigenvalue pattern, the dependence networks for both cancer and normal cases are built. From the results of the protein MS datasets and the gene microarray datasets, we can see clear difference between the dependence networks for cancer and normal cases. Biomarkers are identified based on the difference between dependence networks for normal and cancer cases.

In conclusion, we developed a dependence modeling and network framework to identify cancer biomarkers using protein MS data and microarray data. The proposed framework provides two schemes (i.e. classification-performance-based and dependence-network-based) to identify biomarkers. Based on results from both protein and gene expression data, we observed that the dependence-network-based approach provides much more consistent results in identifying biomarkers, as shown in Figures 3 and 5. This interesting consistency motivates us to further explore the idea of dependence network. In Section 4, in the gastric cancer microarray dataset, the identified biomarkers are examined with respect to their biological significance. Several identified biomarkers have been shown to be valuable biomarkers for gastric cancers in several other studies. The encouraging results reported above demonstrate that the proposed dependence modeling and network framework can facilitate discovery of better biomarkers for different types of cancer.

Conflict of Interest: none declared.

REFERENCES

Adam,B. *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.

Antoniadis,A. *et al*. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563–570.

Blaveri,E. *et al*. (2005) Bladder cancer outcome and subtype classification by gene expression. *Clin. Cancer Res*., **11**, 4044–4055.

Bosserhoff,A. (2006) Novel biomarkers in malignant melanoma. *Clin. Chim. Acta*., **367**, 28–35.

Budzikiewicz,H. (2005) Selected reviews on mass spectrometric topics. *Mass Spectrom. Rev*., **24**, 611–612.

Chen,X. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*., **13**, 1929–1939.

Chen,X. *et al*. (2003) Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell*., **14**, 3208–3215.

Diamandis,E. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell Proteomics*, **3**, 367–378.

Duxbury,M. *et al*. (2005) RNA interference: a mammalian SID-1 homologue enhances siRNA uptake and gene silencing efficacy in human cells. *Biochem. Biophys. Res. Commun*., **331**, 459–463.

Fisher,R. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**, 179–188.

Framson,P. and Sage,E. (2004) SPARC and tumor growth: where the seed meets the soil. *J. Cell. Biochem*., **92**, 679–90.

Furey,T. *et al*. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Haeryfar,S. and Hoskin,D. (2004) Thy-1: more than a mouse pan-T cell marker. *J. Immunol*., **173**, 3581–3588.

Hippo,Y. *et al*. (2002) Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res*., **62**, 233–240.

Inoue,H. *et al*. (2002) Prognostic score of gastric cancer determined by cDNA microarray. *Clin. Cancer Res*., **8**, 3475–3479.

Ivakhno,S. and Kornelyuk,A. (2004) Cytokine-like activities of some aminoacyl-tRNA synthetases and auxiliary p43 cofactor of aminoacylation reaction and their role in oncogenesis. *Exp. Oncol*., **26**, 250–255.

Kubo,Y. *et al*. (2005) ABCA5 resides in lysosomes, and ABCA5 knockout mice develop lysosomal disease-like symptoms. *Mol. Cell. Biol*., **25**, 4138–4149.

Lai,J. *et al*. (2006) SULF1 inhibits tumor growth and potentiates the effects of histone deacetylase inhibitors in hepatocellular carcinoma. *Gastroenterology*, **130**, 2130–2144.

Li,J. *et al*. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin. Chem*., **48**, 1296–1304.

Liu,J. and Li,M. (2004) Finding cancer biomarkers from mass spectrometry data by decision lists. In: *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*.

Liu,Q., Krishnapuram,B., Pratapa,P., Liao,X., Hartemink,A. and Carin,L. (2003) Identification of differentially expressed proteins using MALDI-TOF mass spectra. In: *ASILOMAR Conference: Biological Aspects of Signal Processing, November 2003*, pp. 1323–1327.

Maeng,H. *et al*. (2002a) Appearance of osteonectin-expressing fibroblastic cells in early rat stomach carcinogenesis and stomach tumors induced with *N*-methyl-*N*′-nitro-*N*-nitrosoguanidine. *Jpn J. Cancer Res*., **93**, 960–967.

Maeng,H. *et al*. (2002b) Osteonectin-expressing cells in human stomach cancer and their possible clinical significance. *Cancer Lett*., **184**, 117–121.

Qiu,P. (2005) Ensemble dependence model for classification and predication of cancer and normal gene expression data. *Bioinformatics*, **21**, 3114–3121.

Rege,T. and Hagood,J. (2006) Thy-1 as a regulator of cell-cell and cell-matrix interactions in axon regeneration, apoptosis, adhesion, migration, cancer, and fibrosis. *FASEB J*., **20**, 1045–1054.

Ressom,H. *et al*. (2005) Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics*, **21**, 4039–4045.

Steinhoff,C. *et al*. (2003) Gaussian mixture density estimation applied to microarray data. *Lecture Notes in Computer Sciences (LNCS)*, **2810**, 418–429.

Steve Fu,X., Hu,C., Jie Chen,J., Wang,Z. and Ray Liu,K.J. (2005) Cancer genomics, proteomics, and clinic applications. *Genomic Signal Processing and Statistics*. Hindawi Publishing Corporation.

Thomas,R. *et al*. (2000) Differential expression of osteonectin/SPARC during human prostate cancer progression. *Clin. Cancer Res*., **6**, 1140–1149.

Tibshirani,R. *et al*. (2004) Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, **20**, 3034–3044.

Tzima,E. and Schimmel,P. (2006) Inhibition of tumor angiogenesis by a natural fragment of a tRNA synthetase. *Trends Biochem. Sci*., **31**, 7–10.

van Someren,E. *et al*. (2002) Genetic network modeling. *Pharmacogenomics*, **3**, 507–525.

Wakasugi,K. and Schimmel,P. (1999) Two distinct cytokines released from a human aminoacyl-tRNA synthetase. *Science*, **284**, 147–151.

Walhout,A. and Vidal,M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol*., **2**, 55–62.

Wu,C. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **1** 34(Database issue), D187–D191.