Neurocomputing 92 (2012) 9–17

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Computational analysis of muscular dystrophy sub-types using a novel integrative scheme

Chen Wang [a], Sook Ha [a], Jianhua Xuan [a,*], Yue Wang [a], Eric Hoffman [b]

[a] Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, Virginia, USA
[b] Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC, USA

ARTICLE INFO

Available online 27 February 2012

Keywords:
Gene expression
Classification
Muscular dystrophy
Affinity propagation clustering
Biomarker discovery

ABSTRACT

To construct biologically interpretable gene sets for muscular dystrophy (MD) sub-type classification, we propose a novel computational scheme to integrate protein–protein interaction (PPI) network, functional gene set information, and mRNA profiling data. The workflow of the proposed scheme includes the following three major steps: firstly, we apply an affinity propagation clustering (APC) approach to identify gene sub-networks associated with each MD sub-type, in which a new distance metric is proposed for APC to combine PPI network information and gene–gene co-expression relationship; secondly, we further incorporate functional gene set knowledge, which complements the physical PPI information, into our scheme for biomarker identification; finally, based on the constructed sub-networks and gene set features, we apply multiclass support vector machines (MSVMs) for MD sub-type classification, with which to highlight the biomarkers contributing to sub-type prediction. The experimental results show that our scheme can help identify sub-networks and gene sets that are more relevant to MD than those constructed by other conventional approaches. Moreover, our integrative strategy improves the prediction accuracy substantially, especially for those 'hard-to-classify' sub-types.

1. Introduction

The muscular dystrophy (MD) [1] is a group of inherited muscle diseases characterized by progressive muscle wasting and weakness, consisting of several sub-types with different severity. Although many MD-related defective genes and proteins have been identified, no effective treatments are known yet for many sub-types of MD as their disease pathways are not clearly understood. The availability of high throughput gene expression data provides us the opportunity to elucidate disease pathways involved in MD progression, which is an important task in computational biology aiming for disease biomarker discovery.

Traditional disease biomarker discovery is usually performed by individual gene based classification approaches [2], which ignore the internal relationship among genes, and thus encounter the curse-of-dimensionality problem [3]. Many computational efforts have been put in to address this problem by incorporating biological knowledge. For examples, several supervised approaches [4–6] were proposed to identify phenotype-specific PPI sub-networks so as to reveal related genetic pathways or predict clinical outcomes. Functional gene set categorization was also combined with clinical information to classify disease samples [7]. However, these methods, which are based on supervised learning, could easily overlook many important biomarkers that only mildly correlate with phenotype label only but have strong relevance to the disease status.

To address the aforementioned drawbacks of conventional approaches, we propose an integrative scheme in this paper to fully utilize available biological knowledge such as protein–protein network and functional gene set information to construct biologically interpretable features for sub-type classification. The workflow of the proposed scheme is shown in Fig. 1. Specifically, we use a modified affinity propagation clustering (APC) approach [8] for sub-network identification, incorporating both topological adjacency and expression similarity into the calculation of distance between genes. By doing so, we aim to identify sub-networks comprising genes with consistent activities in the local regions of PPI network. Besides the physical interaction information from PPI, we also use functional gene set knowledge to argument biomarker features, since functional interactions among genes also play important roles in cellular systems. Using both sub-network and functional gene set as features, we then construct classifiers to predict the MD sub-types in a biologically interpretable way, i.e, sub-type specificities are reflected in the abnormal activities of differentially expressed sub-networks and functional gene sets. We have applied

* Corresponding author.
E-mail addresses: topsoil@vt.edu (C. Wang), sook@vt.edu (S. Ha), xuan@vt.edu (J. Xuan), yuewang@vt.edu (Y. Wang), ehoffman@cnmcresearch.org (E. Hoffman).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.08.037

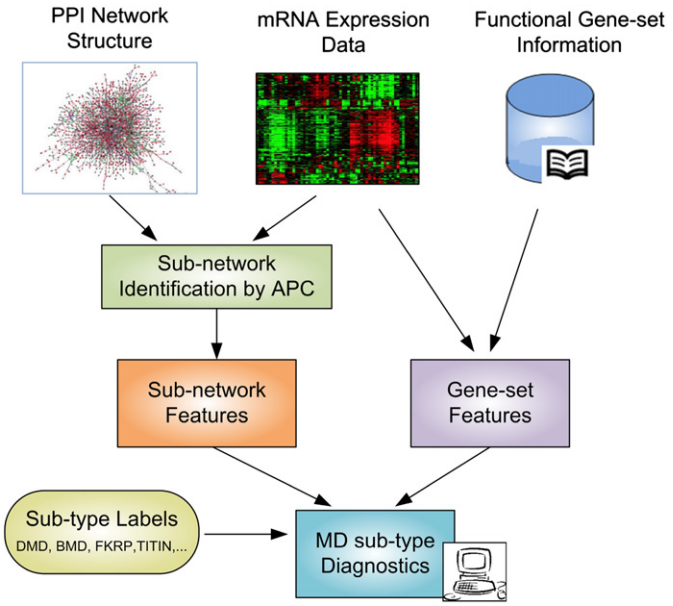


Fig. 1. Workflow of the proposed integrative analysis scheme.

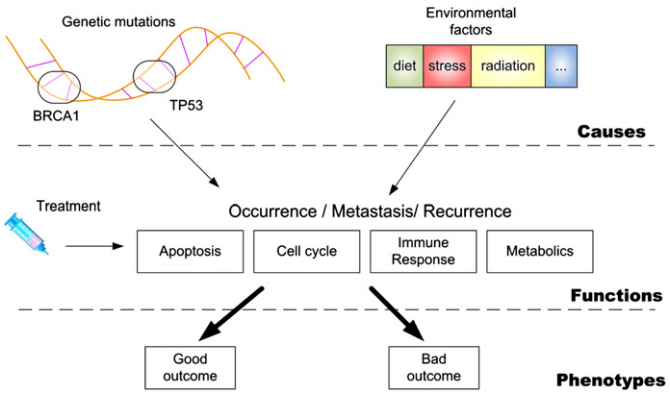


Fig. 2. Different levels in the development of disease.

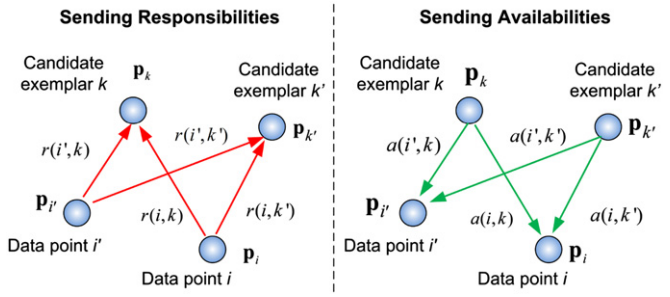


Fig. 3. Message passing of APC.

the proposed scheme to a gene expression dataset with six different MD sub-types for their improved diagnostics. Experimental results show that the sub-networks identified by our scheme are comprised of multiple important pathways related to MD. Moreover, the prediction accuracy has been substantially improved, especially for those sub-types that are difficult to classify.

2. Methods

2.1. Sub-network construction using affinity propagation clustering (APC)

2.1.1. Protein–protein interaction (PPI) information

Proteins collaborate with each other to perform various types of molecular functions and PPI network structure provides potential interaction information of proteins. As the alternation of protein interactions could contribute to diseases onset or progression, a better understanding of disrupted protein sub-networks is essential for the study of disease systems and biomarker discovery. However, there are limitations associated with available PPI information. First, current PPI measurements are quite noisy and every existing technique for PPI information acquisition has its own limitations [9]. Second, PPI only provides the static information of protein interactions and cannot reflect the dynamics of protein interactions in cellular systems. Therefore, it is necessary to incorporate other data types such as gene expression in order to identify condition-specific sub-networks.

Current computational approaches using PPI information can be categorized into three types: The first type is to identify protein complexes, by extracting densely connected modules [5]; the second type is to reveal condition specific gene modules utilizing both phenotype label information and gene expression data [4,6]; the third type is to define gene modules by using unsupervised clustering approaches [10].

Supervised learning is a common approach to discover biomarkers that differentiate phenotypes. However, such an approach is mainly focused on the disease outcomes, and may easily overlook the disease mechanisms underneath. As shown in Fig. 2, human diseases such as cancers are usually caused by genetic and environmental factors through multiple intertwined biological functions. If we focus only on the difference in clinical outcomes, we may lose the important information about the coherence of gene activities and their functional roles. For example, in tumor progression, metabolic activities are the most differentiable signals associated with clinical outcomes but provide limited information for us to understand the underlying mechanism of disease. Another example can be found in our MD study, where the muscle degeneration activity can be successfully used for the diagnostic purpose but hard to be used for the treatment purpose. Aiming to identify biologically informative sub-network biomarkers, we propose to construct sub-networks without using clinical label information directly. Instead, we will use clinical information later in classifiers to highlight MD sub-type specific sub-networks.

2.1.2. Affinity propagation clustering (APC)

Before we describe our sub-network construction method, we will briefly explain the affinity propagation clustering (APC) algorithm in this section. Given a set of data points $\mathbf{P}=\{\mathbf{p}_1,\ldots,\mathbf{p}_N\}$ and function $S(i,j)$ calculating the similarity between $\mathbf{p}_i$ and $\mathbf{p}_j$, the goal of affinity propagation clustering is to find a mapping function $g(\cdot)$ that maximizes the energy function E_g defined as

$$E_g=\sum_{i=1}^{N}S(i,g(i))-\sum_{i=1}^{N}\chi_i(g). \quad (1)$$

The second term in Eq. (1) represents a consistency constraint such that if one data point is an exemplar for other data points, it has to be its own exemplar [8].

The energy function can be optimized through message passing among different data points, and there are two types of messages (as shown in Fig. 3): "availability" $a(i,k)$ represents the accumulated evidence for $\mathbf{p}_k$ to be selected as the exemplar for $\mathbf{p}_i$; "responsibility" $r(i,k)$ tells that how suitable $\mathbf{p}_k$ acts as the exemplar of $\mathbf{p}_i$. The values of these two messages are iteratively

updated as follows:

$$r(k,k) = S(k,k) - \max_{k',k' \neq k} \{S(k,k')\}, \tag{2}$$

$$r(i,k) = S(i,k) - \max_{k',k' \neq k} \{a(i,k') + S(i,k')\}, \tag{3}$$

$$a(i,k) = \min\left\{0, r(k,k) + \sum_{i',i' \neq i,k} \max\{0, r(i',k)\}\right\}, \tag{4}$$

$$a(k,k) = \sum_{i',i' \neq k} \max\{0, r(i',k)\}. \tag{5}$$

Once the algorithm is converged, the index of the most appropriate exemplar for i-th data point is determined by the following formula:

$$g(i) = \arg\max_k \{r(i,k) + a(i,k), k = 1 \ldots N\}. \tag{6}$$

The message passing algorithm of APC involves pair-wise distance calculations, which can incur high computational complexity if the number of data points N is large, thus hinder its applicability to gene clustering; note that APC has been used for microarray sample grouping [11] but not for gene clustering. But with the help of PPI data, the computation load of APC will be greatly reduced since the interactions between proteins are sparse even when the indirectly connected interactions are considered.

In APC, every data point within one cluster can be "represented" by a common exemplar, which is also a data point. Such exemplar-member relationship resembles the gene module network, where a hub gene interacts with other genes in a module. The hub gene can be a key regulator affecting or coordinating the activities of other genes. Such resemblance motivates us to exploit APC to reveal gene modules by incorporating PPI into the gene-gene relevance calculations.

Let $\mathbf{p}_i = [p_{1i}, \ldots, p_{Li}]^T$ be the expression vector of i-th gene across L microarray samples and p_{li} is its gene expression level in l-th microarray sample. Then the correlation coefficient $\rho(\mathbf{p}_i, \mathbf{p}_j)$ between expression vectors of i-th and j-th genes can be defined as follows:

$$\rho(\mathbf{p}_i, \mathbf{p}_j) = \frac{\sum_{k=1}^{L} (p_{ki} - \mu_{p_i})(p_{kj} - \mu_{p_j})}{(L-1)\sigma_{p_i}\sigma_{p_j}}. \tag{7}$$

Here, μ_{p_i}, μ_{p_j}, σ_{p_i} and σ_{p_j} are the means and standard deviations of i-th and j-th expression vectors, respectively. If we only focus on the similarity of expression vectors regardless of up or down regulation of genes, we can measure the relevance $S(i,j)$ between two genes i and j using the following formula:

$$S(i,j) = \frac{|\rho(\mathbf{p}_i, \mathbf{p}_j)|}{d_{ij}^{\gamma}}. \tag{8}$$

Here, d_{ij} can be any topological distance metric between i-th and j-th genes based on PPI network structure [12], and γ is a weight to control the influence of distance to $S(i,j)$. In this paper, we adopt the shortest distance to calculate d and set $\gamma = 1$ for simplicity. If one wishes to tell up- from down-regulated genes, the relevance in (8) can be modified as following:

$$S(i,j) = \frac{\rho(\mathbf{p}_i, \mathbf{p}_j) + 1}{2d_{ij}^{\gamma}}. \tag{9}$$

In both (8) and (9), the relevance is bounded between 0 and 1, with 1 indicating the highest relevance and 0 the lowest. Notice that (9) is more favorable in practice if we need to further combine expression patterns to construct features, as there is no ambiguity of signs.

2.1.3. Significance analysis of identified sub-networks

Unlike conventional clustering methods, sub-networks learned by the proposed scheme can be statistically evaluated using significance analysis. Without label information, it is infeasible to design significance analysis for traditional clustering, and the confidence of resulting clusters cannot be statistically evaluated. In contrast, our proposed scheme is semi-supervised by PPI information, therefore we can shuffle the PPI and gene corresponding relationship to assess the reliability of identified sub-networks. Let us define a statistic to measure the compactness of one sub-network as follows:

$$c_e = \frac{1}{M-1} \sum_{i=1}^{M} S(i,e), \tag{10}$$

where e is the exemplar gene index, M is the number of genes within a sub-network, and $S(i,e)$ measures the relevance between i-th gene and its hub (or, "exemplar"). Using randomly shuffled PPI information, we construct sub-networks and calculate their compactness. A sufficiently large number of random shuffling (e.g. 10,000) is required to construct the null distribution. Based on the null distribution, we can calculate the significance value, i.e, p-value, as follows. Letting $\{c_1^*, \ldots, c_R^*\}$ be the compactness measurements generated by R times of random shuffling, the empirical null distribution $F_R(t)$ can then be defined by the following equation:

$$F_R(t) = \frac{\text{number of elements } \leq t}{R} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\{c_r^* \leq t\}, \tag{11}$$

in which, $\mathbf{1}\{A\}$ is the indication of event A. Based on the empirical null distribution, we define the p-value of an observed compactness measurement c_e as follows:

$$p\text{-value}(c_e) = 1 - F_R(c_e). \tag{12}$$

2.2. Feature construction and classification

2.2.1. Feature constructions

As the PPI information has been exploited for sub-network construction, what we eventually have are multiple gene sub-sets based on identified sub-networks. We first standardize expression level of each gene as z-score [13]

$$z_{li} = \frac{p_{li} - \mu_{p_i}}{\sigma_{p_i}}, \tag{13}$$

where μ_{p_i} and σ_{p_i} are the mean and standard deviations of i-th expression vector, respectively. For the m-th gene sub-set $\mathcal{G}_m$ with N_m gene members, we compute the activity of this gene sub-set in the l-th microarray sample as the aggregated expression of gene members [6,4]

$$act_{lm} = \frac{1}{\sqrt{N_m}} \sum_{i \in \mathcal{G}_m} z_{li}. \tag{14}$$

These gene sub-set activities will be calculated for each individual microarray sample and treated as the features for classification. We also incorporate functional gene sets, as defined from other biological knowledge databases, into the features to take into account the functional interactions between genes. Instead of using all the genes within a functional gene set, we apply a variance based filtering to eliminate the genes with less variance that are more likely to have low signal quality. For each functional gene set, we map gene symbols to probe set ids, and select the sub-set of probe set ids that have relatively large expression variation across all microarray samples. We define the activity of each new gene set by taking the average of the

standardized expressions of all genes belonging to the same set, just like what the activity for our sub-networks is calculated.

2.2.2. Classification techniques

For our MD prediction study, we used three commonly used classification techniques: K-Nearest-Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM). KNN is a non-parametric method that can describe nonlinear decision boundaries for classification, and we include it to investigate whether there is any nonlinearity among different MD sub-types. DT is an approach that can be used to establish tree-like models to for classification or prediction. Since clustering analysis of MD microarray data [14] has already revealed the hierarchical structure among different MD sub-types, we want to further investigate if tree based models can also facilitate classification of MD sub-types. We also use SVM classifier for this study since it is less prone to the curse-of-dimensionality problem intrinsic to the high dimensional microarray data [3]. While KNN and DT algorithms can naturally handle multiclass prediction, the SVM algorithm was originally designed to perform binary classification, and later extended to handle multiclass prediction as well, using one-versus-one (OVO) or one-versus-all (OVA) strategy. In our study, we use the OVA strategy to construct multiclass SVM (MSVM), because the OVA strategy has been reported to perform better than the OVO strategy for classifying microarray datasets with small number of samples [15]. The performance difference can be partially explained by the fact that OVO-MSVM only uses a portion of the training data to construct each binary classifier, thus the resulting classifiers can be more subject to the over-fitting problem.

3. Experiments

3.1. Muscular dystrophy

Before we explain the microarray gene expression data used in this muscular dystrophy (MD) study, we will briefly describe some clinical background of MD diseases. Muscular dystrophy refers to a group of more than 30 genetic muscle diseases characterized by progressive skeletal muscle weakness, defects in muscle proteins, and the death of muscle cells and tissue. The onset of some MD types is in infancy or childhood, while others in middle age or later. The disorders differ in terms of the distribution and extent of muscle weakness, rate of progression, and pattern of inheritance. Among them, Duchenne Muscular Dystrophy (DMD) is known as the most common and fatal form primarily affecting boys, while myotonic MD is the most common form affecting adults. Becker MD (BMD) is similar to DMD but the symptom is less severe. There are no known cures and no specific treatments for any form of MD, and thus the goal of this MD profiling study is to gain a better understanding of MD sub-types so as to enable the development of novel techniques to diagnose, treat, prevent, and ultimately cure this disorder. In this paper, we will focus on computational analysis of six MD sub-types consisting DMD, BMD, dysferlin deficiency (DYS), dystrophy related with fukutin-related protein defect (FKRP), dystrophy related with the TITIN protein encoded by mutated TTN gene (TITIN), and amyotrophic lateral sclerosis (ALS) (see Table 1).

3.2. Dataset description

We analyze a microarray dataset acquired by Children's National Medical Center (CNMC). The dataset consists of 68 microarray samples based on Affymetrix U133-plus2 platform. The disease group consists of 62 samples of six MD sub-types, and the control group consists of six 'normal' samples. A brief summary of the dataset is given in Table 1. PPI information comprising 9303 proteins and 35,000 protein interactions is collected from the Human Protein Reference Database (HPRD) [16], which contains manually curated physical interactions among proteins. 639 functional gene sets are retrieved from Molecular Signatures Database (MSigDB) (http://www.broadinstitute.org/gsea/msigdb/) to take into account the functional interactions between genes.

3.3. Differentially expressed sub-networks and gene sets

By applying our proposed scheme to the MD dataset, we identified 122 sub-networks for this MD study. For a comparison, we also applied PinnacleZ, the software implementation of Chuang's algorithm [4] PinnacleZ (http://chianti.ucsd.edu/~slotia/pinnaclez/help.html), to the same dataset for sub-network identification. PinnacleZ uses a phenotype label guided approach to identify sub-networks, and follows a heuristic strategy to search for phenotype associated sub-networks. It starts from a sub-network consisting of only one selected seed gene, and gradually includes the adjacent genes in PPI network by examining whether including additional genes will increase the association score (i.e., mutual information), which is measured by the relationship between averaged gene expression pattern and phenotype labels. It keeps growing the network until the association score stops increasing or its increasing falls below a certain threshold. Afterwards, statistical assessments are performed extensively to filter out irrelevant sub-networks with non-significant association scores. With the same p-value cut-off used in our proposed approach, PinnacleZ only finds 34 sub-networks which is only 28% of the 122 sub-networks identified by our approach. In addition, the sizes of the individual sub-networks constructed by the PinnacleZ method are smaller than those constructed by our proposed approach (i.e. APC). 41 (34%) APC identified sub-networks have six to ten genes and 46 (37%) have eleven or more gene members. But 79% of PinnacleZ identified sub-networks have six to ten genes, and none has more than ten genes. The summary of comparison is given in Table 2. The difference in the sub-network size (constructed by the two

Table 1
Six MD sub-types and control in the MD dataset.

Class index	Types of muscular dystrophy	No. of samples
1	**CTRL**—Control	6
2	**BMD**—Becker muscular dystrophy	14
3	**DMD**—Duchenne muscular dystrophy	17
4	**DYS**—Dysferlin deficiency; also known as limb-girdle muscular dystrophy 2B (LGMD 2B)	10
5	**FKRP**—Dystrophy related with fukutin-related protein defect	9
6	**TITIN**—Dystrophy related with the TITIN protein encoded by mutated TTN gene	5
7	**ALS**—Amyotrophic lateral sclerosis; also known as Lou Gehrig's disease	7
All	Total number of samples	68

Table 2
Comparison of sub-network size identified by the proposed APC scheme and PinnacleZ method.

Methods	No. of genes in sub-networks			
	2–5	6–10	≥11	Total
PinnacleZ	7 (21%)	27 (79%)	0 (0%)	34 (100%)
APC	35 (29%)	41 (34%)	46 (37%)	122 (100%)

Table 3
MD related pathways captured by (A) the APC identified sub-networks, and (B) PinnacleZ identified sub-networks.

KEGG pathway term	No. of genes/*p*-value
(A)	
Cell adhesion molecules (CAMs)	24/9.15E−06
ECM–receptor interaction	17/4.88E−04
Hematopoietic cell lineage	16/9.51E−04
Focal adhesion	25/1.89E−03
Fc epsilon RI signaling pathway	14/1.90E−03
Natural killer cell mediated cytotoxicity	19/2.23E−03
B cell receptor signaling pathway	12/8.60E−03
Leukocyte transendothelial migration	16/1.50E−02
(B)	
Dentatorubropallidoluysian atrophy	5/1.77E−02
Calcium signaling pathway	13/3.14E−02
Leukocyte transendothelial migration	19/3.32E−02

approaches) could be partly explained by the fact that the heuristic search scheme would limit PinnacleZ to discover complex sub-networks with a large number of gene nodes.

To objectively assess the biological relevance of genes selected by both methods, we conduct functional enrichment analysis using online bioinformatics tools DAVID [17]. The enrichment *p*-value provides us a statistical confidence measure of a specific number of genes falling into specific functional categories, taking the random case as the reference. All the presented *p*-values are corrected by Benjamini technique to handle the multiple hypothesis testing problem [18]. Also, to fairly compare the resulting sub-networks, we selected the same number of sub-networks constructed by both methods according to mutual information score.

Overall, APC identified sub-networks reveal more biological relevance to MD disease by capturing eight MD related pathways, while PinnacleZ identified sub-networks have only captured three pathways with relatively lower statistical significance. Particularly, three most statistically significant pathways captured by APC, namely *Cell adhesion molecules*, *ECM-receptor interaction*, and *Hematopoietic cell lineage* are not included in PinnacleZ identified sub-networks. Specifically, Hematopoietic cell lineage is a canonical pathway involved in self-renewal or differentiation of blood–cell development from Hematopoietic stem cells, which might be related to the muscle loss and resulting systematic compensations. Actually, stem cell based therapy is one of the most promising approaches to treat MD [19]. It has also been documented that cell adhesion molecules and ECM-receptor moleculars all have essential links with various of muscular dystrophy sub-types [1,20].

Table 3 summarizes the KEGG pathway term, the number of genes, and the *p*-value for each MD related pathway captured by APC identified sub-networks (A), and PinnacleZ identified sub-networks (B).

Table 4 presents biological process enrichment analysis results for the APC identified sub-networks (A) and the PinnacleZ identified sub-networks (B). Again, cell adhesion, an important MD related biological process, is enriched only in the genes from the APC identified sub-networks, but not in the genes from the PinnacleZ identified sub-networks.

Table 4
Gene ontology (GO) terms captured by (A) the APC identified sub-networks, and (B) PinnacleZ identified sub-networks.

GO ID: biological process	No. of genes/*p*-value
(A)	
0022610: biological adhesion	72/1.32E−13
0007155: cell adhesion	72/1.32E−13
0032502: developmental process	173/1.81E−11
0048856: anatomical structure development	125/9.91E−10
0048518: positive regulation of biological process	79/3.03E−09
0009605: response to external stimulus	56/4.82E−09
0006952: defense response	52/5.72E−09
0009611: response to wounding	44/6.16E−09
0007049: cell cycle	68/6.53E−09
0002253: activation of immune response	18/8.13E−09
(B)	
0065007: biological regulation	106/1.54E−08
0050789: regulation of biological process	96/4.41E−07
0032502: developmental process	75/5.60E−07
0007242: intracellular signaling cascade	46/1.25E−06
0050790: regulation of catalytic activity	25/1.85E−06
0007165: signal transduction	79/3.01E−06
0030154: cell differentiation	50/3.29E−06
0048869: cellular developmental process	50/3.29E−06
0016043: cellular component organization and biogenesis	64/3.34E−06
0016265: death	32/3.62E−06

To further compare the capability of both methods to detect sub-networks enriched with biological functions, we defined the significance score for each biological function term T with given gene sub-set Q as follows:

$$\text{signf}(T,Q) = -\log_{10}(p\text{-value}(T,Q)), \tag{15}$$

in which, p-value (T,Q) is the DAVID enrichment p-value of biological function T for given gene sub-set Q. The score function signf$(\cdot)$ ranges from 0 to ∞ and the higher score indicates the better enrichment. Thus, we can compute the significance difference of biological enrichment between the gene sub-sets constructed by the proposed scheme and PinnacleZ, based on individual biological function term T

$$\text{Significance difference}(T) = \text{signf}(T,Q_{\text{APC}}) - \text{signf}(T,Q_{\text{PinnacleZ}}), \tag{16}$$

where a positive value of which indicates that our proposed scheme is better to capture the corresponding functional term T, and a negative value suggests PinnacleZ is better. There are totally 647 biological functional terms enriched in the gene sets from both methods, and we draw the significance difference for each term in Fig. 4. We can see that in overall our proposed scheme has much better capability than PinnacleZ to capture biological enriched functions. There is no biological function term with significance difference less than -5, while there are 22 terms associated with significance difference larger than 5.

3.4. Prediction performance

As summarized in Table 5, the prediction accuracy of MSVM based on selected sub-network features is 68%. It is striking to observe a huge contrast between the 100% accuracy for DMD and the 1% accuracy for TITIN. Such a large difference of prediction accuracy could be explained by several reasons including: (i) Clinically, DMD is the most rapidly worsening MD sub-type accompanied by highly varied expression patterns, and thus serves as the easiest diagnostic case. (ii) The number of DMD samples in the dataset is much larger than that of TITIN, and consequently the training of classifier is biased towards DMD.

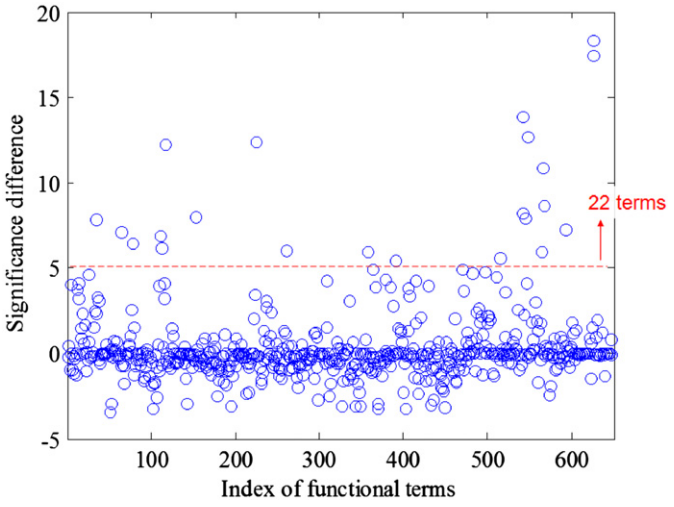


Fig. 4. Significance difference for different biological terms, between APC and PinnacleZ.

Table 5
Prediction accuracy rates measured by MSVM classifier for each MD sub-types and control of features selected from the sub-networks and the gene sets combined.

MD sub-types	Prediction accuracy rates		
	Sub-network features (%)	Combined features (%)	Prediction improvement (%)
CTRL	52	76	24
BMD	68	90	22
DMD	100	99	−1
DYS	61	91	30
FKRP	35	70	35
TITIN	1	42	41
ALS	86	97	11
Average	68	86	18

(iii) PPI sub-network based prediction incorporates only physical interaction information, and it may not be sufficient to tell the sub-type differences by using PPI alone.

As functional interaction could also play vital roles in the onset and progression of MD diseases, we have further added functional gene set features into our prediction analysis. Surprisingly, the results show that the accuracy for TITIN is dramatically improved from 1% to 42%, and the accuracies for DYS, and FKRP are also improved by 30% or more. Fig. 5 shows the prediction performances based on selected sub-network features, and selected combined features (sub-networks and gene sets). Notice that the prediction accuracy of MSVM classification results based on selected sub-network features is only 72% at best, while the accuracy based on combined features is mostly higher than 72% and increases up to 90%. The fact that prediction accuracy is dramatically improved when functional gene set features is added may suggest that the functional interactions play an essential roles in some of the MD sub-types such as DYS, FKRP and TITIN.

We performed KNN classification on our MD microarray data using three different numbers of neighbors ($k=1, 2, 3$). The results for different k value are very similar, and so we will present only the result of $k=2$ case. As we can observe from Fig. 5(A) and (B), the prediction performance of Decision Tree (DT) is the worst, while that of MSVM is the best among the three. The poor performance of Decision Tree can be explained, at least in part, by its complexity in training a tree structure. It also suggests that even though certain MD sub-types may exhibit hierarchical relationship, it is still very

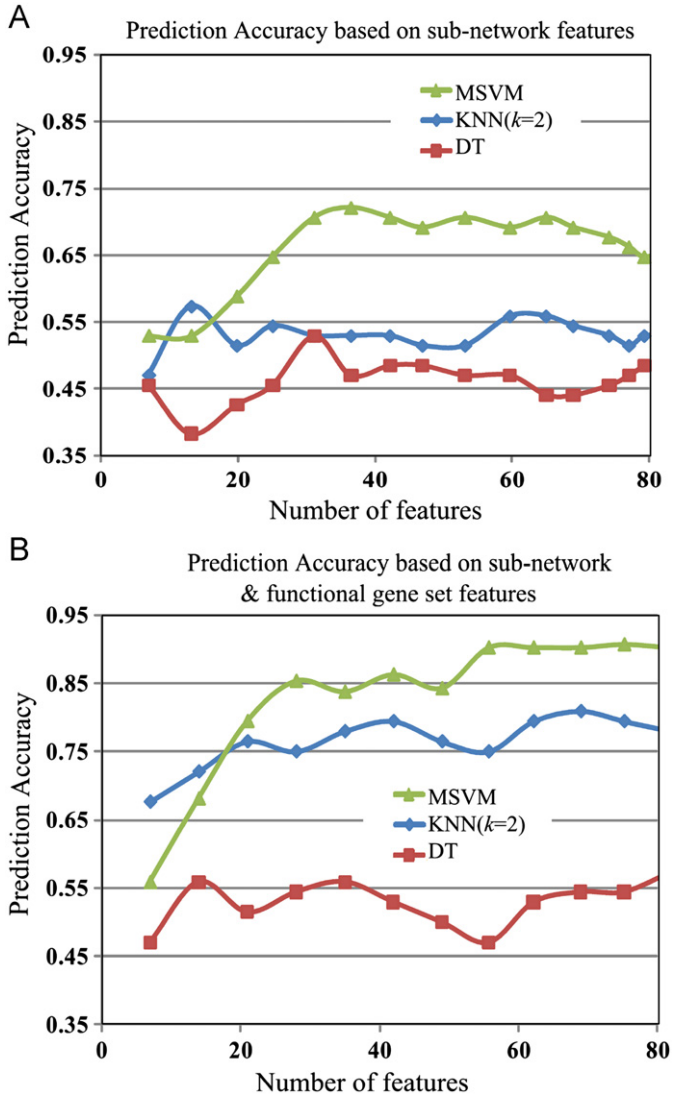


Fig. 5. Prediction accuracy of up to 80 selected sub-network features (A), and sub-network and gene set combined features (B), of MSVM, KNN($k=2$) and DT classifiers.

risky to use classification only scheme to discover such relationship, since the number of samples in the microarray data is usually too small to fully support such relationship, and thus additional clinical information may be needed to overcome such limitation.

3.5. *Some representative sub-networks*

3.5.1. *Sub-network features*

We have presented four representative sub-networks in Fig. 6. From the figure, we can observe that most of the gene nodes are directly connected through protein interactions, and some indirectly related genes can also be identified by our proposed APC scheme. Specifically, sub-network A consisting of 50 genes is dominantly enriched in cell cycle biological process (GO: 0007049, p-value$=2.75E-14$) and cytoskeleton cellular component (GO: 0005856, p-value$=4.66E-6$), indicating that the muscle regeneration activity is vigorous in MD in order to compensate its muscle loss. It is also very interesting to see that all the 10 genes in sub-network B are belonging to glycoprotein category, as it has been reported that the mutation genes of several MD sub-types can interact with glycoprotein to form protein complex

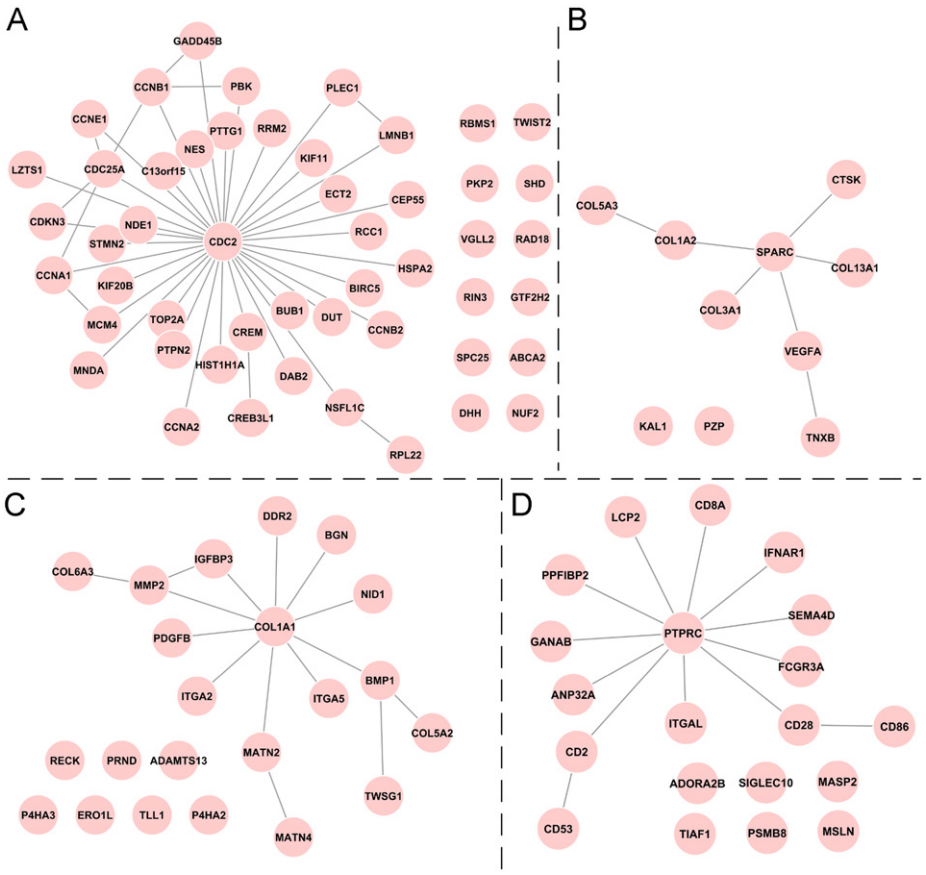


Fig. 6. Four representative sub-networks constructed by APC. The nodes are genes and edges are the protein interactions. Notice that some isolated nodes are also included as proposed APC scheme could identify indirectly related gene nodes.

Table 6
Some representative functional gene sets.

MSigDB gene set name	Descriptions
KEGG_MAPK_SIGNALING_PATHWAY	MAPK signaling pathway
BIOCARTA_STRESS_PATHWAY	TNF_Stress Related Signaling
REACTOME_INSULIN_SYNTHESIS_AND_SECRETION	Genes involved in Insulin Synthesis and Secretion
KEGG_ETHER_LIPID_METABOLISM	Ether lipid metabolism
KEGG_CELL_CYCLE	Cell cycle

[20,21]. These genes are also highly enriched in extracellular matrix cellular component (GO: 0031012, p-value = 3.04E − 9), which is also closely related to MD as we mentioned in the previous section. Sub-network C comprising 22 genes shows similar enrichment in terms of extracellular matrix cellular component (p-value = 3.46E − 5), and it is also enriched in the skeletal muscle growing biological processes (GO: 0001501, p-value = 4.84E − 4) closely related to MD. Unlike all the other sub-networks, sub-network D containing 20 genes emphasizes on the leukocyte activation (GO: 0045321, p-value = 2.20E − 6) and regulation of immune system process (GO: 0002684, p-value = 5.90E − 4), reflecting the active immune response evoked by muscle injures and repairs. The discovery of these enriched biological processes in the constructed sub-networks coincides with the inflammatory pathway activations in MD [22]; anti-inflammatory treatment is also developed to delay the progress of diseases [23]. In summary, our proposed scheme has effectively prioritized the sub-networks closely related to MD disease mechanisms. Note that additional in-depth biological experiments are required to clarify the specific relationships of those features with MD onset and progression.

3.5.2. *Some representative functional gene-sets*

In Table 6, we present a few representative functional gene sets. As MSigDB has various functional gene sets collected from multiple knowledge databases (KEGG, BIOCARTA, REACTOME, etc.) [24], it provides alternative angles for us to investigate MD sub-types. While cell cycle activities are also detected in the gene sets, several different functional pathways are highlighted. Among them, MAPK, TNF and Insulin signaling pathways are known to play important roles in skeletal muscle remodeling and regeneration [25]. Specifically, the activation of MAPK pathway has been reported to be linked with the mutation gene of another MD sub-type named EDMD (Emery–Dreifuss muscular dystrophy) [26]; experimental observations of MAPK and TGFβ1 networks in muscle-wasting pathway also have been reported to contribute to the early onset of DMD [22]. Another study discussed that TNF pathway has links to pro-inflammatory activity and its disrupted signaling may cause exaggerated injury response in Dysferlin sub-type patients [27]. Although biological validations by additional experiments are required to come to any specific conclusion, we can see that those similar biological

process enrichments could be retrieved from both physical sub-networks and functional gene sets information. The proposed integrative approach can provide us with multiple levels and different angles to delineate the complex functional mechanisms of diseases.

4. Discussions and conclusions

In general, analysis of genetic data should be done within a biological context in order to gain a full understanding of complex disease mechanisms. However, commonly used single gene based machine learning approaches are unable to uncover the full picture of complex cellular systems. Different from traditional classification applications mainly focusing on accuracy, microarray based classification usually requires the classification features to be biologically interpretable. The merit to utilize priori-knowledge such as pathways collected in knowledge databases is we can interpret biological context towards resulted features, as well as classification model. Such interpretability can also facilitate the design of follow-up experimental validation to determine how abnormal molecular activities contribute to the distinction between disease sub-types. The weakness is these well studied pathways may be not as effective as some less studied and even unknown pathways to accurately describe sub-type differences. That is also our motivation to integrate PPI information, which is not limited to the context of known pathways, since the identification of PPI sub-networks can potentially reveal some novel pathways in the disease. We have showed an improvement in the prediction results using the selected features constructed from both knowledge sources. More importantly, we have identified many potential sub-network/gene-set biomarkers through feature selection and classification procedures.

Clinically, DMD is the most severe MD sub-type characterized by rapid progression of muscle degeneration [1], and its expression profiles highly vary. Therefore, it is relatively easy for classifiers to differentiate DMD from other sub-types. However, it makes difficult to classify some less severe sub-types with lower expression variations, such as TITIN and FKRP. In addition, since all MD sub-types share the common biological processes such as immune response, apoptosis and cell cycle responding to muscle loss, it is even harder to identify sub-type specific biomarkers. Due to such difficulties, supervised approaches can be biased by dominant expression signals from DMD samples, and fail to capture the gene expression signatures of other weakly distinguishable MD sub-types. In an effort to address such problem, we have proposed a semi-supervised approach, which can be used to identify more biologically interpretable features than conventional clinical label guided approaches [4]. As the discovery of new MD biomarkers could contribute to revealing disruption of genetic pathways in MD diseases [28], our identified sub-network and gene set features may also imply disrupted interactions in related sub-types and provide clues for biological study. As an extension to the proposed computational analysis, we will continue to carry out comparative study on normal muscle recovery experiments [29] for a better understanding of the failed muscle regeneration processes in MD.

Since the identification of condition-specific sub-network has been proved as a NP-hard problem [6], heuristic approaches such as simulated annealing [6] and greedy searching [4] are usually utilized to seek sub-networks associated with large differentiation scores. Instead of directly utilizing sub-type information, we proposed a heuristic scheme to highlight co-expressed sub-networks, considering topological adjacency in PPI network and expression similarity. One weakness of the proposed approach is that the given PPI information could be very general and may consequently degrade the performance of sub-network identification. For the future research, we will study the refinement of PPI topology through combining other information such as co-evolution evidence and topological features [30], and further investigate how to solve PPI refinement and sub-network identification algorithms jointly. In our future research, other biological knowledge such as protein–DNA interaction network structure would also be incorporated into our computational analysis for a deeper understanding of MD diseases. However, biological knowledge contains errors and noises, since it is collected from different sources, such as biological experiments, automatic text-mining results, and manually curated annotations. Therefore, it is essential to carefully examine the reliability or specificity of biological knowledge prior to its use and evaluate its impacts on computational analysis [29]. The limitation of existing biological knowledge poses a challenge for computation approaches to discover meaningful and true biomarkers. Therefore, computational approaches should try to utilize further available biological knowledge while minimizing adverse impact of the biological knowledge due to its incompleteness. In other words, computational approaches that utilize but not restricted by biological knowledge are more desirable for biomarker discovery [31].

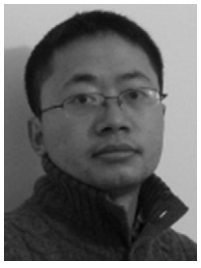
Acknowledgments

This research was supported in part by NIH Grants (R01NS29525-13A1, R01NS29525-18A1, CA139246 and CA149147).

References

[1] A.E. Emery, The muscular dystrophies, Lancet 359 (2002) 687–695.
[2] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, T.R. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Natl. Acad. Sci. USA 98 (2001) 15149–15154.
[3] R. Clarke, H.W. Ressom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nat. Rev. Cancer 8 (2008) 37–49.
[4] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Mol. Syst. Biol. 3 (2007) 140.
[5] E. Georgii, S. Dietmann, T. Uno, P. Pagel, K. Tsuda, Enumeration of condition-dependent dense modules in protein interaction networks, Bioinformatics 25 (2009) 933–940.
[6] T. Ideker, O. Ozier, B. Schwikowski, A.F. Siegel, Discovering regulatory and signalling circuits in molecular interaction networks, Bioinformatics 18 (Suppl 1) (2002) S233–S240.
[7] E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker, D. Lee, Inferring pathway activity toward precise disease classification, PLoS Comput. Biol. 4 (2008) e1000217.
[8] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[9] N. Blow, Systems biology: untangling the protein web, Nature 460 (2009) 415–418.
[10] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer, Co-clustering of biological networks and gene expression data, Bioinformatics 18 (Suppl 1) (2002) S145–S154.
[11] M. Leone, Sumedha, M. Weigt, Clustering by soft-constraint affinity propagation: applications to gene-expression data, Bioinformatics 23 (2007) 2708–2715.
[12] C. Lin, Y. Cho, W. Hwang, et al., Clustering methods in a protein–protein interaction network, in: X. Hu, Y. Pan (Eds.), Knowledge Discovery in Bioinformatics, John Wiley and Sons, Inc., 2007, pp. 319–355.
[13] C. Cheadle, M.P. Vawter, W.J. Freed, K.G. Becker, Analysis of microarray data using z score transformation, J. Mol. Diagn. 5 (2003) 73–81.
[14] Y. Zhu, H. Li, D.J. Miller, Z. Wang, J. Xuan, R. Clarke, E.P. Hoffman, Y. Wang, Cabig visda: modeling, visualization, and discovery for cluster analysis of genomic data, BMC Bioinformatics 9 (2008) 383.
[15] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics 21 (2005) 631–643.
[16] T.S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D.S. Somanathan, A. Sebastian, S. Rani, S. Ray, C.J. Harrys Kishore, S. Kanth, M. Ahmed, M.K. Kashyap, R. Mohmood, Y.L. Ramachandra, V. Krishna, B.A. Rahiman, S. Mohan,

P. Ranganathan, S. Ramabadran, R. Chaerkady, A. Pandey, Human protein reference database—2009 update, Nucleic Acids Res. 37 (2009) D767–D772.
[17] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (2009) 44–57.
[18] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1995) 289–300.
[19] E. Gussoni, Y. Soneoka, C.D. Strickland, E.A. Buzney, M.K. Khan, A.F. Flint, L.M. Kunkel, R.C. Mulligan, Dystrophin expression in the mdx mouse restored by stem cell transplantation, Nature 401 (1999) 390–394.
[20] M. Durbeej, K.P. Campbell, Muscular dystrophies involving the dystrophin–glycoprotein complex: an overview of current mouse models, Curr. Opin. Genet. Dev. 12 (2002) 349–361.
[21] V. Straub, K.P. Campbell, Muscular dystrophies and the dystrophin–glycoprotein complex, Curr. Opin. Neurol. 10 (1997) 168–175.
[22] Y.W. Chen, K. Nagaraju, M. Bakay, O. McIntyre, R. Rawat, R. Shi, E.P. Hoffman, Early onset of inflammation and later involvement of TGFbeta in duchenne muscular dystrophy, Neurology 65 (2005) 826–834.
[23] J.G. Tidball, Inflammatory processes in muscle injury and repair, Am. J. Physiol. Regul. Integr. Comput. Physiol. 288 (2005) R345–R353.
[24] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.
[25] R. Bassel-Duby, E.N. Olson, Signaling pathways in skeletal muscle remodeling, Annu. Rev. Biochem. 75 (2006) 19–37.
[26] A. Muchir, P. Pavlidis, V. Decostre, A.J. Herron, T. Arimura, G. Bonne, H.J. Worman, Activation of MAPK pathways links LMNA mutations to cardiomyopathy in Emery–Dreifuss muscular dystrophy, J. Clin. Invest. 117 (2007) 1282–1293.
[27] K. Nagaraju, R. Rawat, E. Veszelovszky, R. Thapliyal, A. Kesari, S. Sparks, N. Raben, P. Plotz, E.P. Hoffman, Dysferlin deficiency enhances monocyte phagocytosis: a model for the inflammatory onset of limb-girdle muscular dystrophy 2b, Am. J. Pathol. 172 (2008) 774–785.
[28] M. Bakay, Z. Wang, G. Melcon, L. Schiltz, J. Xuan, P. Zhao, V. Sartorelli, J. Seo, E. Pegoraro, C. Angelini, B. Shneiderman, D. Escolar, Y.W. Chen, S.T. Winokur, L.M. Pachman, C. Fan, R. Mandler, Y. Nevo, E. Gordon, Y. Zhu, Y. Dong, Y. Wang, E.P. Hoffman, Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration, Brain 129 (2006) 996–1013.
[29] C. Wang, J. Xuan, L. Chen, P. Zhao, Y. Wang, R. Clarke, E. Hoffman, Motif-directed network component analysis for regulatory network inference, BMC Bioinformatics 9 (Suppl 1) (2008) S21.
[30] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, T. Ideker, Conserved patterns of protein interaction in multiple species, Proc. Natl. Acad. Sci. USA 102 (2005) 1974–1979.
[31] C. Wang, J. Xuan, H. Li, Y. Wang, M. Zhan, E.P. Hoffman, R. Clarke, Knowledge-guided gene ranking by coordinative component analysis, BMC Bioinformatics 11 (2010) 162.



Chen Wang received his Bachelor and Master Degrees in Department of Electronic Engineering and Information Science, at University of Science and Technology of China (USTC) in 2003 and 2006, respectively. He is currently a Ph.D. candidate in Electrical and Computer Engineering of Virginia Tech. His research interests include signal processing and system biology.

Sook Ha received her Bachelor degree in computer science and Master degree in information technology from Virginia Tech in 1994 and 2005, respectively. She is currently a PhD candidate in Electrical and Computer Engineering of Virginia Tech. Her research interests include pathway analysis and visualization of biological data.

Jianhua Xuan received his Ph.D. degree in electrical engineering and computer science from the University of Maryland in 1997. He received his B.S., M.S., and Ph.D. degrees from University of Zhejiang, China, in 1985, 1988, and 1991, respectively, all in electrical engineering. Currently, he is an Associate Professor of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University. His research interests include biomedical image analysis, cellular and molecular imaging, computational bioinformatics, systems biology, intelligent information systems, visual intelligence, computer vision, information visualization, and machine learning.

Yue Wang received his B.S. and M.S. degrees in electrical and computer engineering from Shanghai Jiao Tong University in 1984 and 1987, respectively. He received his Ph.D. degree in electrical engineering from University of Maryland Graduate School in 1995. Currently, he is a Professor of electrical, computer, and biomedical engineering at Virginia Polytechnic Institute and State University. His research interests focus on intelligent computing, machine learning, pattern recognition, statistical visualization, and advanced imaging and image analysis, with applications to molecular analysis of human diseases.

Eric Hoffman received his Ph.D. degree in Biology/Genetics from Johns Hopkins University in 1986. He received B.A. degrees in both Biology and Music from Gettysburg College in 1982. From 1986 to 1988 he was a postdoctoral fellow with Louis Kunkel at Harvard Medical School, and Boston Children's Hospital. Since 1990, he has been Professor of Pediatrics at George Washington University School of Medicine and Health Sciences, and Director of the Research Center for Genetic Medicine at Children's National Medical Center in Washington DC. His research interests include molecular pathogenesis of muscle disease, exercise physiology, development of novel therapeutics, and bioinformatics.