Industrial Engineering
& Management Systems
Vol 16, No 1, March 2017, pp.64-79
ISSN 1598-7248 | EISSN 2234-6473 |
https://doi.org/10.7232/iems.2017.16.1.064

Biological Feature Selection and Disease Gene Identification using New Stepwise Random Forests

Wook-Yeon Hwang*
College of Global Business, Dong-A University

(Received: August 13, 2016 / Revised: November 2, 2016; November 7, 2016 / Accepted: January 9, 2017)

ABSTRACT

Identifying disease genes from human genome is a critical task in biomedical research. Important biological features to distinguish the disease genes from the non-disease genes have been mainly selected based on traditional feature selection approaches. However, the traditional feature selection approaches unnecessarily consider many unimportant biological features. As a result, although some of the existing classification techniques have been applied to disease gene identification, the prediction performance was not satisfactory. A small set of the most important biological features can enhance the accuracy of disease gene identification, as well as provide potentially useful knowledge for biologists or clinicians, who can further investigate the selected biological features as well as the potential disease genes. In this paper, we propose a new stepwise random forests (SRF) approach for biological feature selection and disease gene identification. The SRF approach consists of two stages. In the first stage, only important biological features are iteratively selected in a forward selection manner based on one-dimensional random forest regression, where the updated residual vector is considered as the current response vector. We can then determine a small set of important biological features. In the second stage, random forests classification with regard to the selected biological features is applied to identify disease genes. Our extensive experiments show that the proposed SRF approach outperforms the existing feature selection and classification techniques in terms of biological feature selection and disease gene identification.

Keywords: Bioinformatics, Classification, Feature Evaluation and Selection, Modeling and Prediction

* Corresponding Author, E-mail: wyhwang@dau.ac.kr

1. INTRODUCTION

Identifying phenotype-genotype association from human genome is a critical and fundamental objective in biomedical research (Botstein and Risch, 2013). However, an experimental study and validation of phenotype-genotype association are extremely labor intensive, time consuming and costly. Thus, designing cost-effective computational techniques is very useful for prioritizing and identifying candidate disease-causative genes to benefit human beings. While the increased number of genes has been confirmed to be causative to certain diseases (McKusick, 2007), it still remains a daunting challenge to identify new disease genes for particular diseases because of the limited number of phenotype-gene associations (thus less known disease genes for each individual disease), the inaccuracy of biological similarities between the compared genes, the incompleteness of the biological information, genetic heterogeneity of disease and other complications (Giallourakis *et al*., 2005).

In recent years, computational methods have been applied to discover new disease genes, based on the assumption that a particular disease is caused by the genes with similar functions or certain biological linkages. A common strategy of these computational methods is to evaluate the similarities of candidate (unknown) genes to known disease-causative genes in terms of biological information such as gene expression profiles (Qiu *et al*., 2010), protein sequence information (Aerts *et al*., 2006), protein-protein interaction networks (*PPI*) and network topology (Köhler *et al*., 2008), etc. Candidate genes that share *high* similarities with the confirmed disease genes are considered as the *putative* disease genes that can be experimentally validated by biologists or clinicians. Other recent studies have also revealed that similar diseases/phenotypes are likely to be caused by functional related genes (Oti and Brunner, 2007; Ideker and Sharan, 2008). Also, genes associated with similar disorders have been demonstrated to have physical interactions between their gene products, i.e. proteins (Ideker and Sharan, 2008; Goh *et al*., 2007). They also showed that individual genes associated with particular or similar phenotypes are likely to reside in the same biological functional modules and protein complexes, as molecular machines that integrate and coordinate multiple gene products to perform biological functions (Goh *et al*., 2007; Brunner and Van, 2004; Yang *et al*., 2011). The observations of genetic modular organization of human diseases suggest that the common biological characteristics are associated with the disease-causative genes of particular disease phenotypes.

A number of classification methods have been proposed to discover a disease gene related to biological features using different types of biological data. Xu and Li (2006) proposed the K-nearest neighbor (KNN) classifier to identify disease genes via generating topological features of genetic products in *PPI* networks, such as proteins degree and the percentage of disease genes in proteins neighborhood, etc. Adie *et al*. (2005) extracted evolutionary features from genomic sequence, such as evolutionary conservation, presence, coding sequence length, and closeness of paralogs in the human genome, to identify disease related genes by developing a decision tree algorithm. Radivojac *et al*. (2008) built individual support vector machines (SVM) classifiers using each one of three biological data sets, namely *PPI* networks, protein sequence and protein functional information, and then made consensus predictions based on the results from the three individual classifiers. A recent work has applied a positive-unlabeled learning for disease gene identification (PUDI), which treats unknown genes as an unlabeled set *U*, instead of a negative set *N* Yang *et al*. (2012). The PUDI partitioned the negative set into multiple levels based on their likelihoods to be positives on genes affinity networks. Finally, the PUDI used the multi-level weighted SVM with different penalty values on the multi-level sample set. In its feature representation, the PUDI employed diverse biological features, including protein domains (*D*), three sub-ontologies of gene ontologies, i.e. biological processes (*BP*), molecular functions (*MF*) and cellular components (*CC*), as well as the topological features for the genes in *PPI* (Yang *et al*., 2012).

These existing classification methods focused on integrating diverse biological data sources to get more accurate classification models. However, they did not focus on how to select a small subset of useful features from these diverse sources to reduce the problem dimensionality and to remove noise. We can enhance disease gene identification by considering important features only in the classification methods. In addition, those selected features are very important, since they can provide the novel biological knowledge and insights for biologists to further investigate how they are related to the disease phenotypes. Clearly, there are some standard feature selection techniques (Blum and Langley, 1997; Kohavi and John, 1997; Guyon and Elisseeff, 2003) and classification techniques which can automatically select important features from a large amount of input features. For example, some simple and well-known *filter*-based feature selection methods select features based on the relationship between two random variables. These methods include information gain Mitchell (1997), gain ratio Mitchell (1997) , Chi-square statistic (χ^2 test) Greenwood and Nikulin (1996), correlation feature selection (CFS) Hall (1991) and relief Kenji and Rendell (1991). On the other hand, there are some *wrapper* classification approaches that automatically select an optimal feature subset tailored to a particular classification algorithm, e.g., the original SVM (Hastie, 2001). For example, the 1-norm SVM impose the 1-norm penalty function, instead of the 2-norm penalty function, in the objective function, leading to a sparse SVM (Zhu *et al*., 2004). Alternatively, Zhang *et al.* (2006) proposed the smoothly clipped absolute deviation (SCAD) SVM which minimizes the penalized hinge loss function with the non-concave SCAD penalty (Zhang *et al.*, 2006). Liu and Wu (2007) proposed an approach using a combination of 0-norm and 1-norm penalties (Liu and Wu, 2007). Moreover, Zou (2007) considered the adaptive 1-norm penalty for feature selection (Zou, 2007). Fan and Lv (2007) proposed the sure independence screening (SIS) using simple linear regression, where the original features are standardized (Fan and Lv, 2008). However, the feature selection approaches and the *wrapper* classification approaches unnecessarily considered many unimportant biological features.

There is previous research leveraging random forests (RF) for feature selection. Jiang *et al*. (2009) proposed the SWSFS algorithm selecting several important features based on the gini importance index obtained from the random forest classification (Jiang *et al*., 2009). Botta *et al*. (2014) proposed in their work an extension of the ran-

dom forests algorithm tailored for structured GWAS data based on the variable importance (Botta *et al*., 2004). Wang *et al*. (2016) explored the performance of random forests based on a feature screening procedure to emphasize the SNPs that have complex effects for a continuous phenotype (Wang *et al*., *2016*). In this paper, we propose a new *stepwise* random forests (SRF) approach which consists of two stages: a *forward* feature selection based on *random forests regression* and *random forests classification* with the selected features. Roughly speaking, the SRF approach is a combination of the forward selection with random forest regression and random forest classification. Note that the forward selection with random forest regression is adopted for the SRF approach. The proposed SRF approach has several advantages over the existing filter- and wrapper-based techniques. While the existing techniques simultaneously select important features associated with a disease, the SRF approach selects features *sequentially*, in a forward selection fashion, making our algorithm more efficient. Particularly, in the first stage, we use simple random forests *regression*, which enables us to select a feature iteratively and to reduce the residuals in the previous step. The reduced residuals represent the portion unexplained by the multiple random forests regression model, where more features can be added to reduce the residuals to lead to a better multiple random forests regression model. As a result, a feature highly correlated with the newly added feature at current iteration may not be added to the multiple random forest regression model at next iteration. On the other hand, since the existing *filter*-based feature selection methods only consider correlations between the original response vector and all the features at one time, the highly correlated features can be simultaneously selected for the predictive models. However, the correlated features are redundant and ineffective in explaining the response vector. Moreover, the RF adopted in the SRF can not only capture a linear pattern as well as a non-linear pattern between features and diseases, but also provide the test set error rates monotonically decreasing (Breiman, 2001). Therefore, our SRF approach is expected to select effective features which can result in better prediction performance than the existing approaches.

Our extensive experiments show that the proposed algorithm outperforms the existing feature selection and classification approaches significantly. As a result, we can not only identify and prioritize useful features associated with specific diseases appropriately but also improve the performance of disease gene identification. For the phenotype-genotype data analysis, we consider the various diseases despite that the real data sets for the phenotype-genotype association from human genome are quite limited. Moreover, since we randomly sample the unlabeled data 5 times and perform 3 fold cross validation for each disease, we assume that many real examples are considered. The downside of the SRF is its computation time because random forest regression takes a lot of time. Regarding the computational complexity of the SRF, we mainly need to consider two parts. First, the overall computational complexity of random forests is O(ntree·mtry·(n) log(n)), where ntree represents the number of trees, mtry is the number of features considered at each node and n is the number of samples. Second, $O(p)$ calculations are significantly needed in the one-dimensional regression step of the SRP, where p is the number of features. $O(p)$ calculations in the one-dimensional regression step of the SRP also increases as the number of features increases.

2. EXISTING METHODS

We first describe the various gene features that are used in this research and preliminary work. Next, we introduce some state-of-the-art feature selection and classification methods.

2.1 Gene Features and Preliminary Work

To generate comprehensive biological features to represent the characteristics of genes, we leverage the following biological evidences, namely, protein domains, gene ontology and human protein interactions. Next, we conduct phenotype-gene association and preliminary feature selection.

2.1.1 Protein Domain Features

A protein domain is an evolutionary conserved part of a protein sequence or tertiary structure that can function and evolve independently with the rest of the protein chain. Protein domains often form functional units that participate in transcriptional activities, etc. Protein domains (D) (http://www.sanger.ac.uk/Software/Pfam) can be downloaded from the Pfam domain database that comprises comprehensive domain information about various proteins (Brown and, Jurisica, 2015). To ensure feature accuracy and avoid the noise information in our study, we only select Pfam-A, which is a collection of manually curated and functionally assigned domains and is thus more reliable and accurate. Define the Pfam-A domain set as $\{D_1, \cdots, D_j, \cdots, D_{|\text{Pfam-A}|}\}$, where D_j is a protein domain feature. Note that |Pfam-A| represents the number of the Pfam-A terms. Given a gene g_i, its domain representation will be denoted as a binary feature vector $D(g_i) = (d_{i1}, \cdots, d_{ij}, \cdots, d_{i|\text{Pfam-A}|})$, where d_{ij} ($1 \leq i \leq n$, $1 \leq j \leq$|Pfam-A|) is equal to 1 if protein sequence of g_i contains the domain j; 0 otherwise.

2.1.2 Gene Ontology Features

Gene ontology (GO, http://www.geneontology.org/) database is a very useful source to annotate genes and

their corresponding gene products, where three sub-ontologies, namely biological processes (*BP*) (a series of molecular events with defined beginning and end, e.g., cell division), molecular functions (*MF*) (biological reactions or activities of a gene product at the molecular level, e.g., monooxygenase activity) and cellular components (*CC*) (a part of a cell or its extracellular environment in which a gene product is located, e.g., inner membrane) are provided (Gene Ontology Consortium, 2004).

For a gene g_i , its *MF* GO vector, using its molecular functions annotations from GO, can be represented as $MF(g_i) = (mf_{i1}, \cdots, mf_{ij}, \cdots, mf_{i|MF|})$, where mf_{ij} denotes GO term similarity between g_i's *MF* annotations and the GO feature MF_j. Likewise, we can define a *BP* GO vector as $BP(g_i) = (bp_{i1}, \cdots, bp_{ij}, \cdots, bp_{i|BP|})$ and a *CC* GO vector as $CC(g_i) = (cc_{i1}, \cdots, cc_{ij}, \cdots, cc_{i|CC|})$ respectively, based on *BP* and *CC* where bp_{ij} and cc_{ij} respectively denote GO term similarity in terms of *BP* and *CC*. Note that $|MF|$ represents the number of the GO terms under molecular functions (*MF*) ($|BP|$and $|CC|$ are the number of GO terms in terms of *BP* and *CC* respectively). To compute the GO term similarity between two GO terms, we employ a computational method introduced in (Wang *et al*., 2007) which takes the GO's DAG (Directed Acyclic Graph) structure into consideration.

2.1.3 PPI Networks Features

PPI networks are denoted as $G_{PPI} = (V_{PPI}, E_{PPI})$ where V_{PPI} represents the set of vertices (gene protein products) and E_{PPI} denotes all edges (detected pairwise interactions between proteins). The *PPI* networks used in the paper contain 143,939 *PPI*s among a total of 13,035 human proteins, which are downloaded from human protein reference database (HPRD) (Prasad *et al*., 2009) and online predicted human interaction database (OPHID) (Finn *et al*., 2010). Following Xu and Li's approaches, we build four topological features for a gene g_i, $PPI(g_i) = (degree_i, 1N_i, 2N_i, Cluster_i)$ for the genes in the *PPI* networks, which respectively represent node degrees of genes, the percentage of disease genes in their 1 hop neighborhood (1N-index), the percentage of disease genes in their 2 hop neighborhood of genes (2N-index) and the clustering coefficient of genes (Xu and Li, 2006), which measures the degree to which nodes in a graph tend to cluster together.

2.1.4 Phenotype-Gene Association

4,260 phenotype-gene association data, spanning 2,659 known disease genes and 3200 disease phenotypes, are obtained from the latest version of OMIM (http://omim.org/) (McKusick, 2007). 3,200 disease phenotypes in the OMIM database are categorized into 22 disease phenotype classes based on the physiological system (Goh *et al*., 2007). For example, the endocrine disease phenotype class comprises 62 OMIM phenotypes, including OMIM 241,850 (Bamforth-Lazarus syndrome) and OMIM 304,800 (Diabetes insipidus, nephrogenic). Given a disease phenotype class, genes associated with any phenotypes in the class are treated as a disease gene set *P*, while the remaining genes are treated as a non-disease gene set *N*, which can be used to perform feature selection as well as to build a binary classification model.

2.1.5 Preliminary Feature Selection

Each gene is represented by the biological data categories, *D, BP, MF, CC,* and *PPI* and labeled as a disease gene or a non-disease gene with respective to one disease phenotype class. Then a feature selection strategy is proposed to prioritize the features from the numerous feature set. Because our objective is to prioritize biological features that uniquely identify the given disease phenotype classes, selected features are frequently shared by the disease-related genes in *P* but seldom occur in the non-disease gene set *N* (Yang *et al*., 2012). The feature selection score is computed as follows:

$$FS(f) = \big(F(f, P)+F(f, N)\big) \times log\left(\frac{|P|}{F(f, P)}+\frac{|N|}{F(f, N)}\right), \qquad (1)$$

where $F(f, P)$ is affinity frequency of a feature *f* in *P*. When we want to select important features only from the *MF* feature set, $\{MF_1, \cdots, MF_j, \cdots, MF_{|MF|}\}$, for example, given the GO feature MF_j and a gene g_i, its affinity fre quency in *P* is formatted as $F(MF_j, P) = \sum_{g_i \in P} mf_{ij}$ where a gene set $G = \{g_1, \cdots, g_i, \cdots, g_{|G|}\}$ and the *MF* GO vector $MF(g_i) = (mf_{i1}, \cdots, mf_{ij}, \cdots, mf_{i|MF|})$ are considered. Note that $|P|$ is the number of the *P* terms and $|N|$ is the number of the *N* terms.

2.2 Existing Feature Selection and Classification Methods

We consider well-known feature selection methods, such as information gain (Mitchell, 1997), gain ratio (Mitchell, 1997), Chi-square statistic (χ^2 test) Greenwood and Nikulin (1996) and Relief (Kenji and Rendell, 1992). The feature selection methods belong to *filter-based* approaches that simultaneously select features. In general, they are based on the correlation measures between the features to be selected and the response to be predicted. For binary classification problems in data mining and machine learning research, feature selection methods are firstly employed to select a set of important features. Next, we can apply a classification algorithm to the selected feature subset. After selecting features based on the well-known feature selection methods, we apply the selected features to the random forests classification in this paper.

2.2.1 The 1-norm SVM and SCAD SVM

Unlike the well-known feature selection methods, there are some *wrapper-based* classification approaches which aim to automatically find an optimal feature subset. Note that the SVM variants make the regression coefficients of unimportant features as exactly zeros. As such, we can identify important features whose coefficients are not zeros. Specifically, the 1-norm SVM can be solved by linear programming represented as follows (Zhu *et al*., 2004).

$$\min_{\boldsymbol{\beta},\beta_0,\epsilon_1,\ldots,\epsilon_1,\boldsymbol{s}} \frac{1}{2}\boldsymbol{e}^T\boldsymbol{s} + \gamma\sum_{i=1}^{n}\epsilon_i$$

$$\text{subject to} \quad Y_i\left(\boldsymbol{X}_i^T\boldsymbol{\beta}+\beta_0\right) \geq 1-\epsilon_i,\ \forall_i$$
$$-\boldsymbol{s} \leq \boldsymbol{\beta} \leq \boldsymbol{s},$$
$$\boldsymbol{s} \geq 0,\ \epsilon_i \geq \quad \forall 0,$$

$$\text{where } \gamma > 0, \boldsymbol{e} = (1,\ldots,1)^T \in \mathbb{R}^p, \boldsymbol{s} = \left(s_1,\ldots,s_p\right) \in \mathbb{R}^p \quad (2)$$

For the SCAD SVM, Zhang *et al.* (2006) improved the 1-norm penalty function of the 1-norm SVM by considering a steep penalty for large coefficients (Zhang *et al*., 2006). Particularly, they considered a *constant* penalty for *large* coefficients as well as the 1-norm penalty function for *small* coefficients. Then, the SCAD SVM can be solved by the successive quadratic algorithm, which can be described as follows:

$$\min_{\beta,\beta_0} \sum_{i=1}^{n}\left[1-Y_i(\boldsymbol{X}_i^T\boldsymbol{\beta}+\beta_0)\right]_+ + \sum_{j=1}^{p} p_\lambda(\beta_j),$$
$$\text{where} \quad p_\lambda\left(\beta_j\right) = \lambda\left|\beta_j\right|, \text{ if } \left|\beta_j\right| \leq \lambda,$$
$$p_\lambda\left(\beta_j\right) = \frac{-\left|\beta_j\right|^2 - 2a\lambda\left|\beta_j\right| + \lambda^2}{2(a-1)}, \text{ if } \lambda < \left|\beta_j\right| \leq a\lambda$$
$$p_\lambda\left(\beta_j\right) = \frac{(a+1)\lambda^2}{2}, \text{ if } \left|\beta_j\right| > a\lambda,$$
$$a > 2,\ \lambda > \quad (3)$$

2.2.2 Sure independence Screening (SIS)

Fan and Lv (2007) introduced a property that all the important variables survive after applying a variable screening with probability tending to one. In the sure screening method using simple linear regression, the original features are standardized. Let $\boldsymbol{\beta} = (\beta_1,\ldots\ \beta_p{}^{\mathrm{T}}$, be a p-vector obtained by the simple linear regression. For any given $\delta \in (0,1)$, we sort the p magnitudes of the vector β in a decreasing order and choose a model

$$\{1 \leq \mathrm{i} \leq \mathrm{p} : \left|\beta_i\right| \text{ is among the first } \left|\delta \mathrm{n}\right| \text{ largest of all}\} \quad (4)$$

where $\left|\delta n\right|$ is the integer part of δn and n represents the number of observations. This shrinkage approach is called sure independence screening (SIS) (Fan and Lv, 2008).

3. STEPWISE RANDOM FORESTS

Given a large number of features, it will be extremely time-consuming or impossible to directly find an optimal subset of features which can perform the best classification fit. This involves a NP-hard optimization problem that can only be approximated by heuristic search for a "good" feature subset. Our proposed method aims to iteratively select features one by one in a forward selection manner and thus is more efficient. Overall, the SRF approach consists of the two stages: 1) SRF feature selection to choose a subset of effective biological features based on random forests regression, and 2) disease gene identification by building a random forests classification model considering all the features selected in Stage 1. For stage 1, we build two different types of regression models in each iteration: 1) a *one-dimensional* random forests regression model to find a single best feature to account for the unexplained portion or residual, and 2) a random forests regression model to evaluate the feature subset, i.e. the newly selected best feature, as well as those features already selected from the previous steps to best enhance the regression model performance. We present the detailed SRF algorithm as follows.

Stage 1: SRF feature selection

Without loss generality, given n genes $\{g_1, \cdots, g_i, \cdots, g_n\}$ and p biological features $(f_1, \cdots, f_j, \cdots, f_p)$, we define a regression model $\boldsymbol{Y} = f(\boldsymbol{X})$, where the response $\boldsymbol{Y} = (y_1, \cdots, y_1, \cdots, y_n)^T$ is a vector of disease gene responses, i.e. $y_i = 1$ if g_i is a disease gene, $y_i = -1$ otherwise. The predictors $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_j, \cdots, \boldsymbol{X}_p) = (f_1, \cdots f_j \cdots f_p)$ $= \left(\left(x_{ij}\right)\right), (i = 1, \cdots, n,\ j = 1, \cdots, p)$ is a $n \times p$ matrix with n rows representing the given n genes and p columns denoting the p biological features. In, $\boldsymbol{X}$ each feature can be represented as a column vector, i.e. $\boldsymbol{X}_j = (x_{1j}, \cdots, x_{ij}, \cdots, x_{nj})^T$ For instance, when $\boldsymbol{X} = D$, the j-th protein domain feature $\boldsymbol{X}_j = f_j \in D$ can be considered as one column in $\boldsymbol{X}$, where $x_{ij} = 1$ represents that the i-th gene g_i contains the j-th protein domain feature f_j. Otherwise $x_{ij} = 0$. We define $\boldsymbol{R_m} = (r_{1m}, \cdots, r_{im}, \cdots, r_{nm})^T$ as the updated continuous residual vector at the m-th iteration. Figure 1 illustrates the detailed SRF feature selection procedure in Stage 1. As shown in the initialize step of Figure 1, after we initialize the iteration variable m as 1, the initial residual vector $\boldsymbol{R_m}$ as $\boldsymbol{Y}$, the selected feature set FS_m as the empty set, and the remaining feature set $\boldsymbol{X^m}$ as $\boldsymbol{X}$, we build a one-dimensional random forests regression model for each of the remaining features. In the one-dimensional regression step of Figure 1, one-dimensional

random forests regression is implemented. In the selection step of Figure 1, we aim to select a most effective feature $\boldsymbol{X}_{j_m^*}$ from the remaining features in $\boldsymbol{X^m}$. We then choose a feature which minimizes the sum of the squares of the residuals $\{\sum_{i=1}^{n}[r_{im}\hat{f}_j(x_{ij})]^2\}$. In the feature set step of Figure 1, the remaining biological features at the m-th iteration are denoted by $\boldsymbol{X}^m = \boldsymbol{X} - S_{m-1}$ where the selected feature set $FS_{m-1} = \{X_{j_1^*}, X_{j_2^*}, \cdots, X_{j_{m-1}^*}\}$ are excluded from $\boldsymbol{X}$. We remove the selected feature $X_{j_m^*}$ from $\boldsymbol{X^m}$, which will not be considered again in the one dimensional regression step of the next iteration. Reversely, we add the selected feature $X_{j_m^*}$ into the selected feature set FS_m. The multiple regression step of Figure 1 builds a random forests regression model using all the features selected until the current iteration. The updated multiple random forests regression model is denoted by $\boldsymbol{R_m} = f\left(\boldsymbol{X}_{j_1^*}, \boldsymbol{X}_{j_2^*}, \cdots, \boldsymbol{X}_{j_m^*}\right)$ where features $\boldsymbol{X}_{j_1^*}, \boldsymbol{X}_{j_2^*}, \cdots \boldsymbol{X}_{j_m^*}$ are selected until the m-th iteration. The residual step of Figure 1 obtains the residual vector $\boldsymbol{R}_{m+1}$ by subtracting the residual vector $\hat{R}_m$ estimated by the random forests regression model from the current residual vector $\boldsymbol{R_m}$. The newly updated $\boldsymbol{R}_{m+1}$ represents the remaining unexplained portion (as response) after we add the newly selected feature. When we iteratively perform the algorithm, we expect that the unexplained portion will become smaller because we can build a better random forests regression model than the random forests regression model built in the previous iteration by adding the newly selected feature to FS_m. Finally, in the stopping step of Figure 1, the iteration stops if the difference between the old residual vector and the new residual vector is small enough which implies that no more features need to be added to the multiple random forests regression model. Note that the shrinkage parameter ϵ plays a role in determining how many features are selected for Stage 2. Too many selected features may lead to overfitting and introduce more noise signals, whereas only a few selected features may lead to underfitting and are thus not good enough for classification. Therefore, the shrinkage parameter ϵ can be decided by cross-validation experiments. Additionally, Figure 1 also illustrates the related data sets for the SRF steps. The arrows in Figure 1 represent inputs and outputs for the SRF steps. First, the residual vector and remaining feature set are flowed into one-dimensional regression step. After the selection step, the remaining feature set and selected feature set are updated considering the selected feature. The selected feature set and residual vector are considered at the multiple regression step. At the residual step, the residual vector is updated. If the stopping conditions are satisfied, the SRF algorithm stops. Otherwise, the iteration variable is updated.

Stage 2: disease gene identification

After we select a subset of features FS_{m+1} in Stage 1, we consider the random forests classification with all the

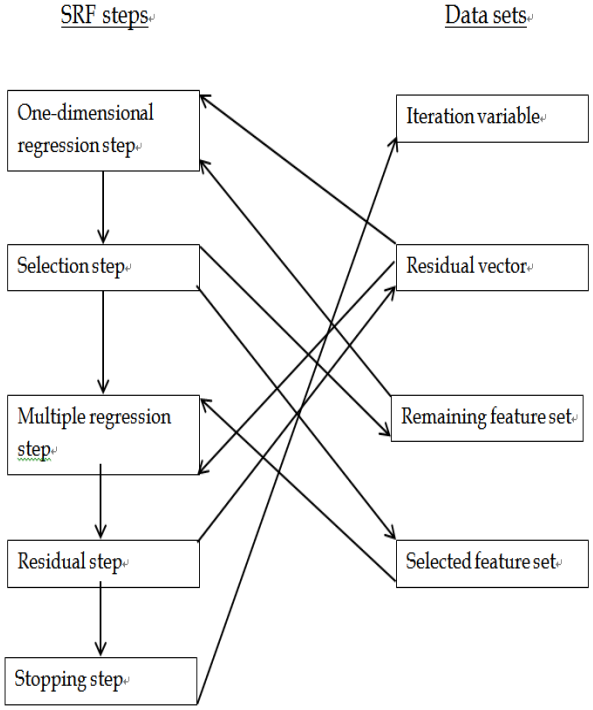

Figure 1. SRF feature selection in Stage 1.

1. Initialize step:
   The iteration variable $m = 1$;
   Residual vector $\boldsymbol{R_1} = \boldsymbol{Y}$;
   Selected feature set $FS_1 = \emptyset$;
   Remaining feature set $\boldsymbol{X^1} = \boldsymbol{X}$;
2. One-dimensional regression step: Adopt one-dimensional random forests regression $\hat{f}_j$: $\boldsymbol{R_m} = f_j(\boldsymbol{X_j})$ for each feature $\boldsymbol{X_j} \in \boldsymbol{X^m}$;
3. Selection step: Select $\boldsymbol{X}_{j_m^*}$ as the most effective feature at the m-th iteration, where $j_m^* = \arg_j \min\{\sum_{i=1}^{n}[r_{im} - \hat{f}_j(x_{ij})]^2\}$;
4. Feature set step: $\boldsymbol{X^{m+1}} = \boldsymbol{X^m} - \{\boldsymbol{X}_{j_m^*}\}$, $FS_{m+1} = FS_m \cup \{\boldsymbol{X}_{j_m^*}\}$;
5. Multiple regression step: Estimate a multiple random forests regression model $\boldsymbol{R_m} = \hat{f}(FS_{m+1}) = \hat{f}(\boldsymbol{X}_{j_1^*}, \boldsymbol{X}_{j_2^*}, \ldots, \boldsymbol{X}_{j_m^*})$ where both previously selected features $\boldsymbol{X}_{j_1^*}, \boldsymbol{X}_{j_2^*}, \ldots, \boldsymbol{X}_{j_{m-1}^*}$ and the newly selected feature $\boldsymbol{X}_{j_m^*}$ are applied for model building;
6. Residual step: Update the residual vector (representing the unexplained portion) with the difference between the current residual and predicted residual, i.e. $\boldsymbol{R_{m+1}} = \boldsymbol{R_m} - \hat{\boldsymbol{R}}_{\boldsymbol{m}}$;
7. Stopping step:
   If $\left| \|\boldsymbol{R_m}\|^2 - \|\boldsymbol{R_{m+1}}\|^2 \right| < \epsilon$,

Then
output FS_{m+1};
Algorithm **STOP**.

Else m++;
Goto 2. One-dimensional regression step.

selected features included in FS_{m+1} in Stage 2. Random forests are a collection of decision trees based on a re-sampling technique (Breiman, 2001). A bootstrap sample for a tree is randomly drawn from the training data. The tree then casts equally one vote for predicting the final response or class. The Law of Large Numbers ensures the convergence (Breiman, 2001). As the number of trees increases, the test set error rates are monotonically decreasing and converge to a limit, without leading to overfitting. The key to accuracy for RF is low correlation and bias. To keep bias low, trees are grown to maximum depth. To remove correlations between trees, each node in the tree randomly selects several features from all the available features in FS_{m+1} . Growing a tree can be achieved by a traditional decision tree algorithm such as the classification and regression tree (CART) (Breiman *et al*., 1999).

Since most of the unnecessary features are not considered in the classification model, we expect better classification performance. Clearly, we perform classification, instead of regression, as our objective is to classify all the genes into the disease gene class and non-disease gene class. Since we build a classification model, the response vector $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is converted into a vector of two categorical (instead of continuous) factors. Finally, two parameters, namely class weights and cutoff values, can be tuned in order to get the best performance of the random forests classification model. We perform a grid search in order to find the optimal combination for minimizing the test errors. In the R package, the cutoff values are tested from 0 to 1 with step 0.05, while the class weights are tested from 1 to 2 with step 0.2. The parameter setting depends on the data sets. For each disease in Table 3, we built 15 classification models, where different parameter settings were applied. Generally speaking, for the R package, the cutoff values were mainly chosen among (0.5, 0.5), (0.55, 0.45), (0.6, 0.4), (0.45, 0.55), and (0.4, 0.6). The class weights were mainly chosen among (1.0, 1.0), (1.2, 1.0), (1.4, 1.0), (1.6, 1.0), (1.8, 1.0), (2.0, 1.0), (1.0, 1.2), (1,0, 1.4), (1.0, 1.6), (1.0, 1.8), and (1.0, 2.0).

4. RESULTS AND DISCUSSION

We conduct comprehensive experiments to compare our proposed SRF procedure with the existing state-of-the-art techniques. We first introduce our experimental data, settings and evaluation metrics. Then, we present the experimental results.

4.1 Experimental Data, Settings, and Evaluation Metrics

The data formats for phenotype-genotype association from human genome are different from SNPs or microarrays, where n represents the number of subjects and p represents the number of genes. On the other hand, in this paper, n represents the number of genes, whereas p represents the number of the biological features because we consider phenotype-genotype association from human genome. The experimental data include the 3 GO ontologies, protein domains, and protein interaction data. After collecting each gene's experimental data, for each specific disease, we employ the preliminary feature selection approach Yang *et al*. (2012) to perform a preprocessing step to get a total of 4,004 (p) *raw* features. In detail, the first 3,000 features are evenly from *BP*, *MF*, and *CC*, the next 1,000 features are from *D* and the last 4 features corresponded to the topological properties of genes in *PPI*. Also, the number of genes (n) for training data is 230. To comprehensively evaluate various feature selection techniques, we include six disease classes, namely cardiovascular diseases, endocrine diseases, neurological diseases, metabolic diseases, ophthalmological diseases and cancer diseases, which were also used in the recent research papers (Yang *et al*., 2012; Yang *et al*., 2014). Given a disease gene class, disease genes confirmed from OMIM (McKusick, 2007) are treated as a disease gene set *P*, while we randomly sample non-disease genes from Ensembl (Flicek *et al*., 2011), as a non-disease gene set *N* (for 5 times), where the number of the sampled non-disease genes is equal to the number of the disease genes for a specific disease. Disease gene prediction can thus be modeled as a binary classification problem to distinguish the disease genes from the non-disease genes. Basically, the dataset are high-dimensional, where the number of the observations is 230 and the number of the biological features is 4,004. The protein domain features are comprised of zeros and ones. However, the number of the ones is very small so that the protein domain features are not selected by the SRF. The remaining domain features range between 0 and 1. We consider zeros for the missing values.

To perform more comprehensive fair evaluations, we compare two categories of existing techniques with the SRF. The first category of the techniques includes standard filter-based feature selection methods, followed by the random forests classification considering the selected features. These feature selection methods include info gain, gain ratio, Chi-square statistic and relief. The second category of the techniques, on the other hand, focuses on the wrapper classification techniques. These wrapper classification methods include the original random forests classification, the 1-norm SVM, the SCAD SVM, Smalter's method (Smalter *et al*., 2007) which employed the original SVM, Xu and Li's method which employed the KNN classifier (Xu and Li, 2006), the multi-class SVM-based PUDI method which applied the positive unlabeled learning techniques, Fan and Lv's SIS and the gradient boosting trees (Hastie *et al*., 2001).

Note that the parameters of the SVM, KNN, 1-norm

SVM, SCAD SVM, the random forests classification, the SIS and the gradient boosting trees are tuned for minimizing the classification test errors (Hastie *et al*., 2001) where R packages are used to implement them. The PUDI's executable codes are available at http:/www1.i2r.a-star.edu.sg/~xlli/PUDI/PUDI.html. For our proposed SRF, the shrinkage parameter ϵ plays a role in determining the number of the selected features. We set ϵ = 0.0000001 producing minimum test errors for selecting features. Finally, we employ precision, recall and F-measure (Bollmann and Cherniavsky, 1981), widely used measure to evaluate the performance of all the methods mentioned above (Yang *et al*., 2012; Yang *et al*., 2014). The F-measure is the harmonic mean of precision (p) and recall (r), which is represented as $2\times p\times r/(p+r)$. The F-measure evaluates an average effect of both precision and recall. When either of them (p or r) is small, its value will be small. Only when both of them are large, it will be large. This is suitable for our purpose because having either too small a precision (i.e. the percentage of accurately predicted disease genes is small) or too small a recall (the percentage of the identified disease genes is small) for disease gene prediction is unacceptable and would be reflected by a low F-measure.

4.2 Experimental Results

We compared the proposed SRF with the existing methods in terms of the performance metrics and number of the selected features, analyzed those discovered features and identified novel disease genes.

4.2.1 Comparison Results

We summarized the experimental results to compare our proposed method with different methods. First, we conducted a comparison with the filter-based methods followed by the random forests classification, where the cancer disease case was considered. We chose a subset of features with the highest correlation measures using Chi-square statistic (Chi_square), information gain (Info_gain), gain ratio (Gain_ratio) and relief (Relief) respectively. We also considered the one-norm SVM, SCAD SVM, the random forests classification without feature selection (RF1) and the random forests classification (RF2) considering only features with the highest mean Gini gain produced by each gene over all trees for completeness of the existing methods. The classification models were built on a training set and the classification performance was evaluated on a test set.

From Table 1, we observed that most of the feature selection methods, including both filter- and wrapper-based approaches performed worse than RF1 which used all the 4,004 features. These results demonstrated that the feature selection methods, albeit useful to find small number of distinguishing features, may not be able to achieve better classification results than simply using all the features. On the other hand, our proposed SRF method, which considered 23 features only, clearly outperformed the other existing methods, indicating that it can capture the most informative distinguishing features to differentiate the disease genes from the non-disease genes. Note that for a fair comparison, we have considered the optimal number of the selected features for all the filter-based methods, Chi-square statistic (Chi_square), information gain (Info_gain), gain ratio (Gain_ratio) and relief (Relief) as well as RF2. Particularly, the SRF approach was able to achieve 3.48%, 4.17% and 7.55% better F-measures than the second best RF1 method, the third best method gain ratio and fourth best method RF2 respectively. Therefore, we can conclude that the feature selection process of the SRF is very effective than the state-of-the-arts. We also found that the RF1, RF2 and gain ratio performed better than the 1-norm SVM considering 322 features and the SCAD SVM considering 3,001 features.

In order to explain why the SRF approach outper-

Table 1. Comparison of performance for cancer diseases

Approach	Precision	Recall	F-measure	Number of features	Feature selection time (min.)
RF1	84.61%	75.86%	80.00%	4004	0
RF2	81.48%	75.86%	78.57%	88	33.35
Chi_square	77.42%	82.76%	80.00%	115	0.39
Info_gain	77.97%	79.31%	78.63%	135	0.45
Gain_ratio	78.69%	82.76%	80.67%	13	0.43
Relief	79.59%	67.24%	72.90%	194	270.93
1-norm SVM	77.36%	70.69%	73.87%	322	0.56
SCAD SVM	63.93%	67.24%	65.55%	3001	4.81
SIS	76.27%	77.59%	76.92%	987	0.42
SRF	84.21%	82.76%	**83.48%**	23	146.24

Table 2. Results of the shrinkage parameter values of the SRF for cancer diseases

Shrinkage parameter	Precision	Recall	F-measure	Number of features
0.0000000001	78.57%	75.86%	77.19%	32
0.000000001	75.86%	77.19%	76.52%	27
0.00000001	80.00%	75.86%	77.88%	25
0.0000001	84.21%	82.76%	**83.48%**	23
0.000001	78.95%	77.59%	78.26%	20
0.00001	81.03%	81.03%	81.03%	18
0.0001	83.33%	77.59%	80.36%	16
0.001	77.19%	75.86%	76.52%	13
0.01	78.18%	74.14%	76.11%	11
0.1	80.77%	72.41%	76.36%	8

formed the other feature selection approaches in Table 1, we additionally conducted another real analysis. For the cancer disease data, we compared the top 10 selected features between the SRF approach and the chi-square approach. We observed that among the top 10 selected features for the SRF approach, there exist low linear correlations, whereas there exist high linear correlations among the top 10 selected features for the chi-square approach. The SRF approach updates the response vector by subtracting the portion explained by the newly added feature from the current response vector. Therefore, a feature highly correlated with the newly added feature at current iteration may not be added to the regression model at next iteration. On the other hand, the chi-square approach only considers correlations between the original response vector and all the features at one time. Therefore, highly correlated features with high chi-square statistics can be simultaneously selected for the model. However, the correlated features can be redundant in explaining the response vector. As a result, we observed that the prediction accuracy 78.45% for the top 10 features selected by the SRF is better than 74.24% for the top 10 features selected by the chi-square approach.

Table 3. Comparison of performance for the six disease cases

Disease class	Approach	Precision	Recall	F-measure
Cardiovascular	PUDI	72.44%	76.67%	74.15%
	SVM	68.71%	76.27%	71.32%
	KNN	69.42%	72.94%	70.99%
	GBM	78.33%	73.53%	**75.12%**
	SRF	80.09%	78.39%	**78.68%**
Endocrine	PUDI	74.64%	84.94%	79.20%
	SVM	76.75%	77.53%	76.32%
	KNN	74.40%	77.04%	75.45%
	GBM	84.61%	81.48%	**83.02%**
	SRF	81.43%	89.63%	**85.16%**
Neurological	PUDI	68.01%	79.36%	73.14%
	SVM	65.90%	75.71%	70.17%
	KNN	66.27%	66.58%	66.25%
	GBM	66.90%	81.28%	**73.93%**
	SRF	75.58%	77.53%	**76.30%**
Metabolic	PUDI	83.77%	86.78%	85.13%
	SVM	86.22%	81.36%	83.49%
	KNN	80.12%	81.89%	80.91%
	GBM	76.53%	85.23%	**80.65%**
	SRF	86.32%	85.38%	**85.78%**
Ophthalmological	PUDI	74.81%	83.81%	78.81%
	SVM	79.31%	75.24%	75.96%
	KNN	68.77%	71.43%	70.32%
	GBM	96.00%	68.57%	**80.00%**
	SRF	85.78%	84.57%	**84.80%**
Cancer	PUDI	74.21%	79.96%	76.89%
	SVM	79.47%	76.55%	77.58%
	KNN	77.89%	78.97%	78.28%
	GBM	68.49%	86.21%	**80.00%**
	SRF	80.88%	82.87%	**81.38%**
Average	PUDI	74.65%	81.92%	77.89%
	SVM	76.06%	77.11%	75.81%
	KNN	72.81%	74.81%	73.70%
	GBM	78.48%	79.38%	78.79%
	SRF	81.68%	83.06%	**82.02%**

As shown in Table 2, we observed the performance results of the shrinkage parameter values in the SRF. The smaller the shrinkage parameter, the more the number of the selected features. The shrinkage parameter value ϵ = 0.0000001 provided the best precision (84.21%), recall (82.76%), and F-value (83.48%), where the number of the selected features was 23 as mentioned. Finally, we compared the SRF method with four existing classification techniques, Smalter's method based on the SVM, Xu and Li's method based on the KNN, the multi-class SVM-based PUDI method based on the positive unlabeled learning techniques and the gradient boosting trees. Note that these four techniques do not focus on feature selection. We have performed 5 times 3 fold cross-validations where each time we selected one fold as a test set and 2 folds as a training set and reported the average results. As shown in Table 3, the proposed SRF approach outperformed the existing approaches for the disease cases in terms of the F-measure consistently. In summary, we observed that on average the SRF is 4.1%, 6.2%, 8.3% and 3.2% better F-measures than the PUDI, Smalter's method, Xu and Li's method and the gradient boosting trees respectively indicating that the SRF's prediction is much more accurate and thus the prediction results are more reliable than the other techniques. For the metabolic disease case, all the approaches performed very well, with the F-measures more than 80%. We performed two-way ANOVA analysis for Table 3. We verified that the mean of the F-measures for the SRF is significantly greater than those of the existing approaches based on pairwise comparisons of means with Tukey contrasts, where the significance level was 0.015. In conclusion, the results indicate that SRF's prediction is much more accurate and thus the prediction results are more reliable than the other techniques with the most informative distinguishing features to differentiate the disease genes from the non-disease genes.

4.2.2 Feature Analysis

Our proposed SRF approach typically only selected

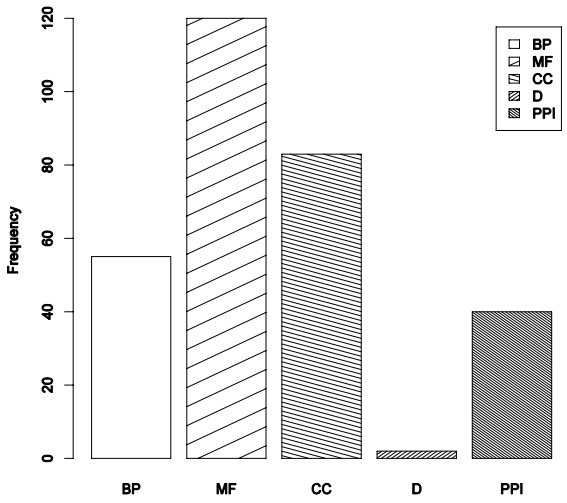

Figure 2. Distribution of first to twentieth selected features for cardiovascular diseases.

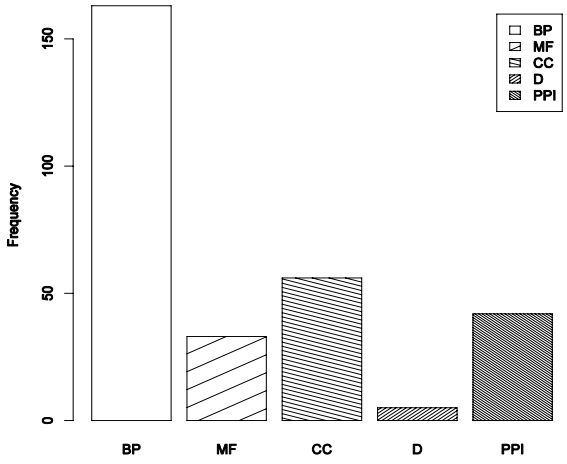

Figure 3. Distribution of first to twentieth selected features for endocrine diseases.

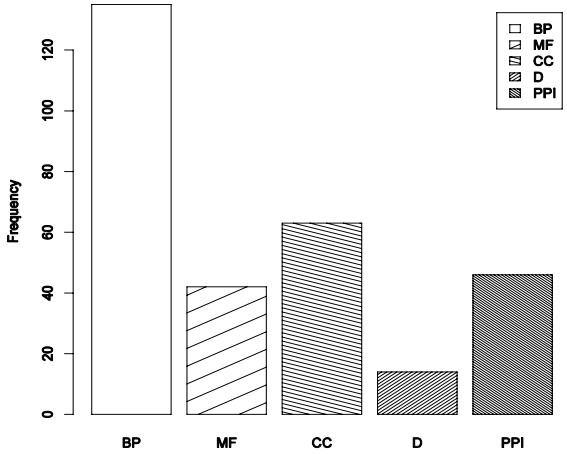

Figure 4. Distribution of first to twentieth selected features for neurological diseases.

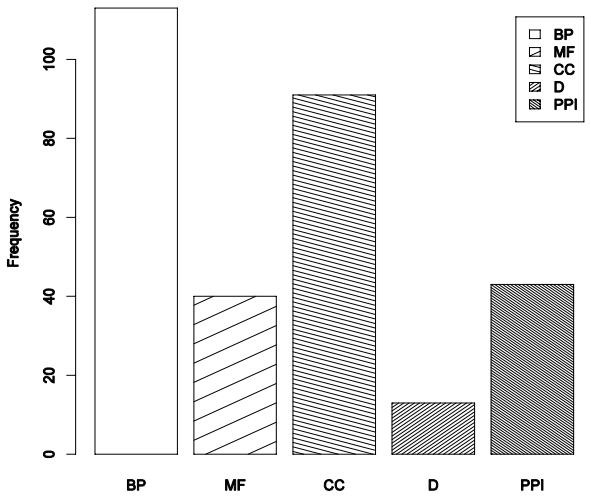

Figure 5. Distribution of first to twentieth selected features for metabolic diseases.

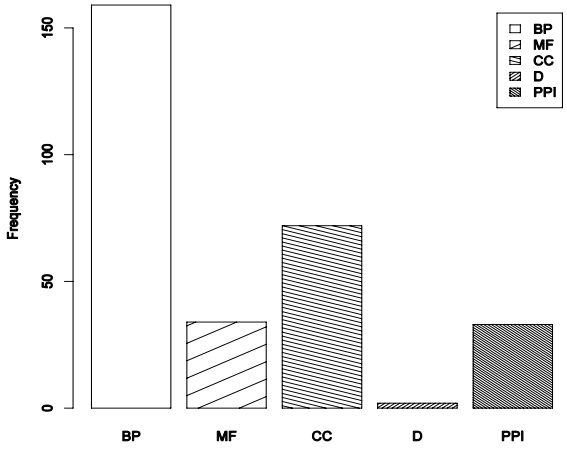

Figure 6. Distribution of first to twentieth selected features for ophthalmological diseases.

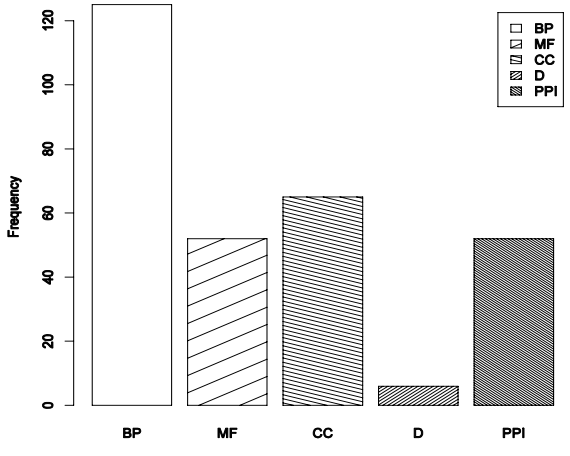

Figure 7. Distribution of first to twentieth selected features for cancer diseases.

less than 30 features in Table 3, depending on the disease classes that we studied. Based on the SRF results of Table 3, we have performed feature analysis, focusing on feature distributions and feature-disease association. Firstly, we studied the distribution of top 20 selected features for each disease class to better understand which feature groups these selected features come from. Figure 2 illustrates the distribution results for cardiovascular diseases where the *x*-axis represents the five feature groups and *y*-axis denotes their corresponding frequencies. We observed from Figure 2 that more features were selected from *MF* and cellular components *CC*. As shown in Figure 3, our approach mainly selected features from *BP* (the number of features from the other feature groups are relatively small) for endocrine diseases. Similar to Figure 3, Figure 4 shows that features from *BP* were substantially selected for neurological diseases. On the other hand, features from *CC* and *BP* were prominent for metabolic diseases, as shown in Figure 5. In Figure 6, features from *BP* were clearly much more than the other feature groups for ophthalmological diseases. For cancer diseases, features from *BP* are the most remarkable, while those from *CC*, *MF*, and *PPI* are similarly substantial as shown in Figure 7.

Table 4. The top 5 highly ranked selected features

Disease	Feature groups	Attributes
Cardiovascular	*PPI*	2N-index
	PPI	Clustering coefficient
	MF	GO0008092
	MF	GO0016922
	BP	GO0001906
Endocrine	*PPI*	2N-index
	BP	GO0001906
	BP	GO0015975
	BP	GO0043067
	BP	GO0031589
Neurological	*PPI*	2N-index
	BP	GO:0018262
	BP	GO:0018941
	BP	GO:0043067
	BP	GO:0015975
Metabolic	*PPI*	2N-index
	PPI	Clustering coefficient
	BP	GO0015975
	BP	GO:0018262
	BP	GO:0042772
Ophthalmological	*PPI*	2N-index
	BP	GO:0007059
	BP	GO:0018298
	BP	GO:0044343
	BP	GO:0018262
Cancer	*PPI*	2N-index
	PPI	Clustering coefficient
	BP	GO0015975
	BP	GO:0018262
	BP	GO:0042772

In summary, throughout Figures 2-7, we observed that features from *BP* are particularly striking in the distribution results, indicating that biological processes (e.g., cell division) are the most important to the rest of the diseases except for cardiovascular diseases - the diseases could be caused when various biological processes slow down, halt or even reversed. We can also see that features from *PPI*, albeit consisting of only 4 features, were consistently chosen for the diseases. This clearly shows that proteins either directly or indirectly interacting with different disease proteins (1N-index and 2N-index), are likely to be disease proteins too. In addition, those protein hubs or coordinators can affect the diseases in terms of the degree and highly connected subgraphs (clustering coefficient). Their mutations could maximally disrupt the operations of the modules which impact cell fitness (Tew *et al*., 2007).

Next, we investigated specific features that have been chosen by our proposed technique. Table 4 shows those top 5 highly ranked selected features for each disease class. Most significantly, the feature 2N-index was selected for all the six disease classes. As such, it is the most important disease-independent feature. Clustering coefficient is the second frequently selected disease-independent feature, which was chosen for the three diseases, namely cardiovascular diseases, metabolic diseases and cancer diseases.

On the other hand, there are some disease-specific features. For example, GO 0008092 (cytoskeletal protein binding), GO 0016922 (ligand-dependent nuclear receptor binding) and GO 0001906 (cell killing) were chosen for cardiovascular diseases. Our literature search has found that heart failure is in fact associated with the cytoskeleton as it leads to cardiac muscle contraction (Katz, 2000). GO 0016922 (ligand-dependent nuclear receptor binding) interacts with some nuclear receptor proteins. The orphan nuclear receptors, ERRalpha and gamma affect cardiac functions (Dufour *et al*., 2007). Regarding GO 0001906 (cell killing), both gradual and acute cell death is hallmarks of cardiac pathology, including heart failure, myocardial infarction and ischemia/reperfusion (Bollmann and Cherniavsky, 1981).

For endocrine diseases, GO 0001906 (cell killing) was selected. Type I diabetes are affected by β-cell death (Tew *et al*., 2007). For GO 0015975 (energy deriva-

tion by oxidation of reduced inorganic compounds), endocrine diseases are related to oxidation of compounds (Olmez-Hanci *et al*., 2009). GO 0043067 (regulation of programmed cell death) regulates programmed cell death which affects endocrine-dependent tissues (Martimbeau and Tilly, 1997). GO 0031589 (cell-substrate adhesion) attaches a cell to the underlying substrate via adhesion molecules. Men1 (multiple endocrine neoplasia 1) is associated with negative regulation of cell-substrate adhesion (http://www.informatics.jax.org/searches/GOannot_report.cgi?id=GO:0031589).

For neurological diseases, GO 0018262 (isopeptide cross-linking) forms a covalent cross-link between or within peptide chains. Parkin is an ubiquitin-protein isopeptide ligase. It has been suggested that loss of function in parkin causes accumulation and aggregation of its substrates, leading to death of dopaminergic neurons in Parkinson's disease (Zhong *et al*., 2005). Also, Suk *et al.* (2001) showed that isopeptide is formed in neuronal cell death (*Suk et al*., 2001). GO 0018941 (organomercury metabolic process) is associated with any organic compound containing a mercury atom. Mahaffey *et al.* (2004) discussed that increased risk of adverse neurodevelopmental effects is related to methyl mercury exposure (Mahaffey *et al*., 2004). For GO 0043067 (regulation of programmed cell death), neuron cell death was discussed in the literature (Mahaffey *et al*., 2004). Also, Johnson *et al.* (2005) defined the form of cell death regarding neuro-2a cells (Johnson *et al*., 2005). Regarding GO 0015975 (energy derivation by oxidation of reduced inorganic compounds), oxidative stress (OS) leading to free radical attack on neural cells contributes to neuro-degeneration, which can cause a range of disorders such as Alzheimer's disease, Parkinson's disease, aging and many other neural disorders (Uttara *et al*., 2009).

We observed that GO 0015975 (energy derivation by oxidation of reduced inorganic compounds) is also associated with metabolic diseases. Metabolism is usually divided into two categories: catabolism and anabolism, where catabolism breaks down organic matter and harvests energy by way of cellular respiration (http://en.wikipedia.org/wiki/Metabolism). For GO 0018262 (isopeptide cross-linking), a significant different biochemical behavior of the isopeptides was observed in terms of vitro stability, vivo metabolism as well as biodistribution (Hultsch *et al*., 2005). For GO 0042772 (DNA damage response, signal transduction resulting in transcription), a strong correlation was established between the systemic DNA damage response to inhibit ongoing malignant transformation and metabolic syndrome characteristics (Erol, 2010).

Our proposed method has chosen GO 0007059 (chromosome segregation) for ophthalmological diseases. This makes sense as recent studies suggested that the DNase domain-containing protein TATDN1 plays an important role in both chromosome segregation and eye development in zebrafish (Johnson *et al*., 2005). For GO 0018298 (protein-chromophore linkage), protein-chromophore interacts with metarhodopsins (Renk and Crouch, 1989). Rhodopsin known as visual purple (a light-sensitive receptor protein) is a biological pigment in photoreceptor cells of the retina (http://en.wikipedia.org/wiki/Rhodopsin). Regarding GO 0044343 (canonical Wnt signaling pathway involved in regulation of type B pancreatic cell proliferation), it was investigated that pancreatic cardinoma is associated with the optic nerve (Ring, 1967). For GO 0018262 (isopeptide cross-linking), Tong *et al.* (2011) discussed transglutaminase in ocular health and pathological processes. Transglutaminase (TG) is a big class of intra- and extra-cellular enzymes with 9 members, all of which catalyze the formation of epsilon-(γ-glutamyl) lysine isopeptide linkages between peptide substrates, except for a catalytically inactive member Band 4.1 (Tong *et al*., 2011).

For the cancer disease case, regarding GO 0015975 (energy derivation by oxidation of reduced inorganic compounds), in humans, oxidative stress is thought to be involved in the development of cancer, Parkinson's disease, Alzheimer's disease, atherosclerosis, heart failure, myocardial infarction, fragile X syndrome, sickle cell disease, lichen planus, vitiligo, autism, infection and chronic fatigue syndrome (http://en.wikipedia.org/wiki/Oxidative_stress). For GO: 0018262 (isopeptide cross-linking), it was proven that lysine-isopeptides are associated with cancer (Szende *et al*., 2002). For GO: 0042772 (DNA damage response, signal transduction resulting in transcription), Karanika *et al*. (2014) pointed out that DNA damage response is associated with prostate cancer (Karanika *et al*., 2014). In conclusion, as the selected biological features match very well with the existing biological knowledge, other selected biological features could be putative features for biologists and clinicians to validate.

4.2.3 Novel Disease Gene Identification

We tested if our proposed SRF method can identify novel disease genes. Particularly, those genes from the non-disease gene test set (do not belong to known disease genes) with prediction probabilities more than 0.9 actually belonging to the disease gene class, are considered as novel disease genes. This is reasonable as we only have limited known disease genes for each specific disease. In fact, those unknown genes, currently being regarded as non-disease genes, could be potential disease genes, especially when our classification model classifies them into the disease gene class or we believe that they are more similar to existing disease genes.

Based on the SRF results of Table 3, we discovered novel disease genes for endocrine diseases. Our experimental results predicted six novel disease genes, namely *SDC1*, *NCOA2*, *HES1*, *PCSK5*, *GRIN1* and *MARKAPK3*. Our literature search has found that *SDC1* has been asso-

ciated with endocrine diseases (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly). In addition, Jeong *et al.* (2007) found that *NCOA2* regulates murine endometrial function, progesterone independent, and dependent gene expression. Furthermore, Johansson *et al*. (2008) has shown that HES1 is related to pancreatic endocrine tumors. It was discussed that *PCSK5* mediates posttranslational endoproteolytic processing for several integrin alpha subunits (Cao *et al*., 2001). Also, Xin *et al*. (2009) implied that *GRIN1* is associated with endocrine diseases. Finally, it was pointed out that *MARKAPK3* is associated with thyroid (Endocrine System) (http://www.genecards.org/cgi-bin/carddisp.pl?gene=MAPKAPK3).

On the other hand, we applied our proposed SRF method to discover novel disease genes for cancer diseases. As a result, five novel disease genes were predicted: *RXRA*, *NEK1*, *MTOR*, *EPHA3* and *MAP3K7*. Tsujie *et al*. (2003) suggested that activation of *RXRA* pathway might affect growth inhibition of pancreatic cancer cells. Additionally, mutation of NEK1 leads to chromosome instability and ultimately to increased mutation rates and acquisition of the multiple mutations that result in cancer (Chen *et al*., 2011). We found that *MTOR* has emerged as a critical effector in cell-signaling pathways commonly deregulated in human cancers (Guertin and Sabatini, 2007). *EPHA3*, moreover, maintains tumorigenicity and is highly expressed on the tumor-initiating cell population in glioma (Day *et al*., 2013). We also observed that *MAP3K7* is associated with lung and breast cancer (Kondo *et al*., 1998; Neil and, Schiemann, 2008). In conclusion, as our predicted disease genes match very well with the existing biological knowledge, other novel predicted disease genes could be putative disease genes for biologists and clinicians to validate.

5. CONCLUSION

The existing classification methods utilized the multiple biological data sources for disease gene identification. However, they did not focus on how to automatically select a small subset of useful features to distinguish the disease genes from the non-disease genes. Because the effectively selected features could improve the accuracy of disease gene identification as well as provide biologists and clinicians with more biological insights, in this paper, we proposed the SRF for feature selection and disease gene identification. The SRF consists of the two stages. In the first stage of the SRF, biological features are iteratively selected one by one in a forward selection manner. At each iteration step, a one-dimensional random forests regression is applied to select a feature which can best reduce the residual or unexplained portion. After we get a best feature for the current iteration, we update the multiple random forests regression model using all the features selected so far as well as the residual vector. As a result, we can effectively determine a subset of good features in the first stage. In the second stage, the random forests classification is implemented for disease gene identification.

Our extensive experimental results demonstrated that our approach performs significantly better than the existing standard feature selection and classification approaches, and state-of-the-art classification methods in terms of feature selection and disease gene identification. Particularly, for the cancer disease case, although the SRF selected only 23 features, it remarkably outperformed the existing approaches in terms of the F-measure. Furthermore, we showed that the selected features are related to the diseases via literature search. Finally, we identified novel disease genes based on the SRF, which would be useful for disease diagnostics and treatments.

ACKNOWLEDGEMENTS

The authors thank the editors and referees for reviewing our paper. This work was supported by the Dong-A University research fund.

REFERENCES

Adie, E., Adams, R., Evans, K., Porteous, D., and Pickard, B. (2005), Speeding disease gene discovery by sequence based candidate prioritization, *BMC bioinformatics*, 6-55.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L., De Moor, B., Marynen, P., and Hassan, B. (2006), Gene prioritization through genomic data fusion, *Nat Biotechnol*, **24**, 537-544.

Blum, A. and Langley, P. (1997), Selection of relevant features and examples in machine learning, *Artificial Intelligence*, **97**(1-2), 245-271.

Bollmann, P. and Cherniavsky, V. S. (1981), *Restricted Evaluation in Information Retrieval*, ACM SIGIR.

Botstein, D. and Risch, N. (2013), Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease, *Nature Genetics*, **33**, 228-237.

Botta, V., Louppe, G., Geurts, P., and Wehenkel, L. (2014), *Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies*, PLOS ONE, http://dx.doi.org/10.1371/journal.pone.0093379.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1991), *Classification and Regression Trees*, CRC Press, New York.

Breiman, L. (2001), Random forests, *Machine Learning*,

45, 5-32.

Brown, K. and Jurisica, I. (2005), Online predicted human interaction database, *Bioinformatics*, **21**(9), 2076-2082.

Brunner, H., Van, D. M (2004), From syndrome families to functional genomics, *Nature Reviews Genetics*, **5**, 545-551.

Cao, H., Mok, A., Miskie, B., and Hegele, R. A. (2001), Single-nucleotide polymorphisms of the proprotein convertase subtilisin/kexin type 5 (PCSK5) gene, *J Hum Genet*, **46**, 730-732.

Chen, Y., Chen, C. F., Chiang, H. C., Pena, M., Polci, R., Wei, R. L., Edwards, R. A., Hansel, D. E., Chen, P. L., and Riley, D. J. (2011), Mutation of NIMA-related kinase 1 (NEK1) leads to chromosome instability, *Molecular Cancer*, doi:10.1186/1476-4598-10-5.

Chiong, M., Wang, Z., Pedrozo, Z., Caom D., Troncoso, R., and Ibacache, M. (2011), Cardiomyocyte death: mechanisms and translational implications, *Cell Death & Disease*, **2**, e244.

Day, B. W., Stringer, B. W., Al-Ejeh, F., Ting, M. J., Wilson, J., Ensbey, K. S., Jamieson, P. R., Bruce, Z. C., Lim, Y. C., Offenhäuser, C., Charmsaz, S., Cooper, L. T., Ellacott, J. K., Harding, A., Leveque, L., Inglis, P., Allan, S., Walker, D. G., Lackmann, M., Osborne, G., Khanna, K. K., Reynolds, B. A., Lickliter, J. D., and Boyd, A. W. (2013), EphA3 maintains tumorigenicity and is a therapeutic target in glioblastoma multiforme, *Cancer Cell*, **23**(2), 238-248.

Deshmukh, M., Li, Y., Yokota, T., Gama, V., Yoshida, T., Gomez, J. A., Ishikawa, K., Sasaguri, H., Cohen, H. Y., Sinclair, D. A., Mizusawa, H., and Matsuyama, S. (2007), Bax-inhibiting peptide protects cells from polyglutamine toxicity caused by Ku70 acetylation, *Cell Death and Differentiation*, **14**, 2058-2067.

Dufour, C. R., Wilson, B. J., Huss, J. M., Kelly, D. P., Alaynick, W. A., Downes, M., Evans, R. M., Blanchette, M., and Giguère, V. (2007), Genome-wide orchestration of cardiac functions by the orphan nuclear receptors ERR alpha and Gamma, *Cell Metab*, **5**(5), 345-356.

Erol, A. (2010), Systemic DNA damage response and metabolic syndrome as a premalignant state, *Current Molecular Medicine*, **10**(3), 321-334.

Fan, J. and Lv, J. (2008), Sure independence screening for ultrahigh dimensional feature space, *JRSS. B.*, **70**, 849-911.

Finn, R., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2010), The pfam protein families database, *Nucl Acids Res*, **38**, 211-222.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J. P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. J. (2011), Ensembl, *Nucl Acids Res*, **39**(1), 800-806.

Gene Ontology Consortium (2004), The gene ontology database and informatics resource, *Nucleic Acid Res.*, **32**(1), 258-261.

Giallourakis, C., Henson, C., Reich, M., Xie, X., and-Mootha, V. (2005), Disease gene discovery though integrative genomics, *Annu Rev Genomics Hum Genet*, **6**, 381-406.

Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabási, A. (2007), The human disease network, *Proc Natl Acad Sci USA*, **104**, 8685-8690.

Greenwood, P., and Nikulin, M. (1996), *A Guide to Chi-squared Testing*, John Wiley & Sons.

Guertin, D. A. and Sabatini, D. M. (2007), Defining the role of mTOR in cancer, *Cancer Cell*, **12**(1), 9-22.

Guyon, I. and Elisseeff, A. (2003), An introduction to variable and feature selection, *Journal of Machine Learning Research*, **3**, 1157-1182.

Hall, M. (1999), *Correlation-based Feature Selection for Machine Learning*, Ph.D. thesis.

Hastie, T., Tibsharani, R., and Friedman, J. H. (2001), The elements of statistical learning, *Springer*, New York.

Hultsch, C., Bergmannm R., Pawelke, B., Pietzsch, J., Wuest, F., Johannsen, B., and Henle, T. (2005), Biodistribution and catabolism of 18F-labelled isopeptide N(epsilon)-(gamma-glutamyl)-L-lysine, *Amino Acids*, **29**(4), 405-413.

Ideker, T. and Sharan, R. (2008), Protein networks in disease, *Genome Research*, **18**, 644-652.

Jeong, J. W., Lee, K. Y., Han, S. J., Aronow, B. J., Lydon, J. P., O'Malley, B. W., and DeMayo, F. J. (2007), The P160 steroid receptor coactivator 2, SRC-2, regulates murine endometrial function and regulates progesterone-independent and -dependent gene expression, *Endocrinology*, **148**, 4238-4250.

Jiang, R., Tang, W., Wu, X., and Fu, W. (2009), A random forest approach to the detection of epistatic interactions in case-control studies, *BMC Bioinformatics*, DOI:10.1186/1471-2105-10-S1-S65.

Johansson, T., Lejonklou, M. H., Ekeblad, S., Stålberg, P., and Skogseid, B. (2008), Lack of nuclear expression

of hairy and enhancer of split-1 (HES1) in pancreatic endocrine tumors, *Horm Metab Res*, **40**(5), 354-359.

Johnson, V. J., Kim, S., and Sharma, R. P. (2005), Aluminum-maltolate induces apoptosis and necrosis in neuro-2a cells: Potential role for p53 signaling. *Toxicol Sci*., **83**(2), 329-339.

Karanika, S., Karantanos, T., Li, L., Corn, P. G., and-Thompson, T. C. (2014), DNA damage response and prostate cancer: Defects, regulation and therapeutic implications. *Oncogene*, doi:10.1038/onc.2014.238.

Katz, A. M. (2000), Cytoskeletal abnormalities in the failing heart out on a LIM?, *Circulation*, **101**(23), 2672-2673.

Kenji, K. and Rendell, L. (1992), The feature selection problem: traditional methods and a new algorithm, *Proceeding AAAI'92 Proceedings of the Tenth National Conference on Artificial Intelligence*, 129-134.

Kohavi, R. and John, G. (1997), Wrappers for feature selection, *Artificial Intelligence*, **97**(1-2), 273-324.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. (2008), Walking the interactome for prioritization of candidate disease genes, *The American Journal of Human Genetics*, **82**, 949-958.

Kondo, M., Osada, H., Uchida, K., Yanagisawa, K., Masuda, A., Takagim K., Takahashim T., and Takahashi. T. (1998), Molecular cloning of human TAK1 and its mutational analysis in human lung cancer, *Int. J. Cancer*, **75**(4), 559-563.

Liu, Y. and Wu, Y. (2007), Variable selection via a combination of the L0 and L1 penalties, *J. Comp. Graph. Statist.*, **16**, 782-798.

Mahaffey, K. R., Clickner, R. P., and Bodurow, C. C. (2004), Blood organic mercury and dietary mercury intake: National health and nutrition examination survey, *Environ Health Perspect*, **112**(5), 562-570.

Martimbeau, S. and Tilly, J. L. (1997), Physiological cell death in endocrine-dependent tissues: An ovarian perspective, *Clinical Endocrinology*, **46**(3), 241-254.

McKusick, V. (2007), Mendelian inheritance in man and its online version, OMIM. *Am J Hum Genet*, **80**, 588-604.

Mitchell, M. (1997), *Machine learning*, WCB.

Neil, J. R. and Schiemann, W. P. (2008), Altered TAB1: I KappaB kinase interaction promotes transforming growth factor beta-mediated nuclear factor-kappaB activation during breast cancer progression, *Cancer Res*, **68**(5), 1462-1470.

Olmez-Hanci, T., Imren, C., Arslan-Alaton, I., Kabdaşli, I., and Tünay, O. (2009), H2O2/UV-C oxidation of potential endocrine disrupting compounds: A case study with dimethyl phthalate, *Photochem Photobiol Sci*., **8**(5), 620-627.

Oti, M. and Brunner, H. (2007), The modular nature of genetic diseases, *Clin Genet*, **71**, 1-11.

Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., and Pandey, A. (2009), Human protein reference database, *Nucleic Acids Res*, **37**, 767-772.

Qiu, Y., Zhang, S., Zhang, X., and Chen, L. (2010), Detecting disease associated modules and prioritizing active genes based on high throughput data, *BMC Bioinformatics*, 11-26.

Radivojac, P., Peng, K., Clark, W., Peters, B., Mohan, A., Boyle, S., and Mooney, S. (2008), An integrated approach to inferring gene-disease associations in humans, *Proteins*, **72**(3), 1030-1037.

Renk, G. and Crouch, R. K. (1989), Analogue pigment studies of chromophore-protein interactions in metarhodopsins, *Biochemistry*, **28**(2), 907-912.

Ring, H. G. (1967), Pancreatic carcinoma with metastasis to the optic nerve, *Arch Ophthalmol*, **77**(6), 798-800.

Smalter, A., Lei, S., and Chen, X. (2007), Human disease-gene classification with integrative sequence-based and topological features of protein-protein interaction networks, *BIBM*.

Suk, S., Kim, Y., and Lee, S. (2001), Formation of nuclear isopeptide in the process of neuronal cell death following interstitial hyperthermia in normal rat brain, *Journal of Korean Neurol Association*, **19**(6), 633-640.

Szende, B., Szökán, G., Tyihá, E., Pál, K., Gáborjányi, R., Almás, M., and Khlafulla, A. R. (2002), Antitumor effect of lysine-isopeptides, *Cancer Cell International*, doi:10.1186/1475-2867-2-4.

Tew, K., Li, X., and Tan, S. (2007), Functional centrality: Detecting lethality of proteins in protein interaction networks, *Proceedings of 18th International Conference on Genome Informatics*.

Tong, L., Png, E., Lan, W., and Petznick, A. (2011), Recent advances: Transglutaminase in ocular health and pathological processes, *J Clinic Experiment Ophthalmol*, doi:10.4172/2155-9570.S2-002.

Tsujie, M., Nakamori, S., Okami, J., Takahashi, Y., Hayashi, N., Nagano, H., Dono, K., Umeshita, K., Sakon, M., and Monden, M. (2003), Growth inhibition of pancreatic cancer cells through activation of peroxisome proliferator-activated receptor Gamma/Retinoid X Receptor Alpha pathway, *Int J Oncol*, **23**(2), 325-331.

Usmani-Brown, S., Lebastchi, J., Steck, A. K., Beam, C., Herold, K. C., and Ledizet, M. (2014), Analysis of β-cell death in type 1 diabetes by droplet digital PCR, *Endocrinology*, **155**(9), 3694-3698.

Uttara, B., Singh, A. V., Zamboni, P., and Mahajan, R. T. (2009), Oxidative stress and neurodegenerative diseases: A review of upstream and downstream antioxidant therapeutic options, *Curr Neuropharmacol*, **7**(1), 65-74.

Wang, G., Fu, G., and Corcoran, C. (2015), A forest-based feature screening approach for large-scale genome data with complex structures, *BMC Genetics*, DOI: 10.1186/s12863-015-0294-9.

Wang, Z. D., Payattakool, R., Philip, S., and Chen, C. (2007), A new method to measure the semantic similarity of GO terms, *Bioinformatics*, **23**(10), 1274-1281.

Xin, G., Qiu, Y., Loh, H. H., and Law, P. Y. (2009), GRIN1 regulates μ-opioid receptor activities by tethering the receptor and G protein in the lipid raft, *Journal of Biological Chemistry*, **284**(52), 36521-36534.

Xu, J. and Li, Y. (2006), Discovering disease-genes by topological features in human protein-protein interaction network, *Bioinformatics*, **22**(22), 2800-2805.

Yang, H., Liu, C., Jamsen, J., Wu, Z., Wang, Y., Chen, J., Zheng, L., and Shen, B. (2012), The DNase domain-containing protein TATDN1 plays an important role in chromosomal segregation and cell cycle progression during zebrafish eye development, *Cell Cycle*, **11**(24), 4626-4632.

Yang, P., Li, X., Chua, H., Kwoh, C., and Ng, S. (2014), Ensemble positive unlabeled learning for disease gene identification, *PloS one*, **9**(5).

Yang, P., Li, X., Mei, J., Kwoh, C., and Ng, S. (2012), Positive-unlabeled learning for disease gene identification, *Bioinformatics*, **28**(20), 2640-2647.

Yang, P., Li, X., Wu, M., Kwoh, C., and Ng, S. (2011), Inferring gene-phenotype associations via global protein complex network propagation, *PloS one*, **6**(7), e21502.

Zhang, H., Ahn, J., Lin, X., and Park, C. (2006), Gene selection using support vector machines with non-convex penalty, *Bioinformatics*, **22**, 88-95.

Zhong, T., Tan, Y., Zhou, A., Yu, Q., and Zhou, J. (2005), RING finger ubiquitin-protein isopeptide ligase Nrdp1/FLRF regulates parkin stability and activity, *Journal of Biological Chemistry*, **280**(10), 9425-9430.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004), 1-norm support vector machines, *The Annual Conference on Neural Information Processing Systems*.

Zou, H. (2007), An improved 1-norm support vector machine for simultaneous classification and variable selection, *J. Machine Learn. Res., Proceedings Track*, **2**, 675-681.