Chapter 11
Random Forest for Bioinformatics

Yanjun Qi

11.1 Introduction

Modern biology has experienced an increased use of machine learning techniques for large scale and complex biological data analysis. In the area of Bioinformatics, the Random Forest (RF) [6] technique, which includes an ensemble of decision trees and incorporates feature selection and interactions naturally in the learning process, is a popular choice. It is nonparametric, interpretable, efficient, and has high prediction accuracy for many types of data. Recent work in computational biology has seen an increased use of RF, owing to its unique advantages in dealing with small sample size, high-dimensional feature space, and complex data structures.

The aim of this chapter is twofold. First, to provide a review of notable extensions of RF in bioinformatics, whereby promising direction such as RF-based feature selection is discussed. Second, to briefly introduce the applications of RF and its extensions. RF has been applied in a broad spectrum of biological tasks, including, for example, to classify different types of samples using gene expression of microarrays data, to identify disease associated genes from genome wide association studies, to recognize the important elements in protein sequences, or to identify protein–protein interactions (PPIs).

11.2 Random Forest and Extensions in Bioinformatics

Random forest provides a unique combination of prediction accuracy and model interpretability among popular machine learning methods. The random sampling and ensemble strategies utilized in RF enable it to achieve accurate predictions as

Y. Qi (✉)
Machine Learning Department, NEC Labs America, Princeton, NJ, USA

4 Independence Way, Suite 200, Princeton, NJ, USA
e-mail: yanjun@nec-labs.com; qiyanjun07@gmail.com

C. Zhang and Y. Ma (eds.), *Ensemble Machine Learning: Methods and Applications*,
DOI 10.1007/978-1-4419-9326-7_11,

well as better generalizations. This generalization property comes from the bagging scheme which improves the generalization by decreasing variance, while similar methods like boosting achieve this by decreasing bias [47].

Three features of RF receive the main focus [6]:

- It provides accurate predictions on many types of applications;
- It can measure the importance of each feature with model training;
- Pairwise proximity between samples can be measured by the trained model.

Extending random forest is currently a very active research area in the computational biology community, where most previous efforts focused on extending the features above. Several notable techniques among them are briefly introduced in the sections that follow.

11.2.1 Classification Purpose

Random forest retains many benefits of decision trees while achieving better results through the usage of bagging on samples, random subsets of variables, and a majority voting scheme [6]. It handles missing values, a variety of variables (continuous, binary, categorical), and is well suited to high-dimensional data modeling. Unlike classical decision trees, there is no need to prune trees in RF since the ensemble and bootstrapping schemes help RF overcome overfitting issues. Motivated by the excellent performance of RF, developing RF variants is an active research topic in computational biology [47].

One category of extension tried to revise how to construct trees in RF. For instance, Zhang et al. [48] proposed a deterministic procedure to form a forest of classification trees to maintain scientific interpretability in the structure of the trees. The procedure screens trees by selecting a prespecified number, say 20, of top splits of the root node and another prespecified number, say 3, of the top splits of the two daughter nodes of the root node. This protocol of top nodes gives rise to a total of 180 possible ($20 \times 3 \times 3$) trees (Fig. 11.1), among which, those with perfect or near perfect classification precision are of particular interests. Finally, a fixed number of available trees are selected to form a deterministic forest. Their experiments claimed that the deterministic forest performs similar to RFs, but with better reproducibility and interpretability.

Researchers also tried to extend RF by considering special properties in biological data sets, e.g., too many noisy features in DNA microarray data. Amaratunga et al. [2] designed so-called "enriched random forest" for when the number of features is huge and the percentage of truly informative features is small. To reduce the contribution of trees whose nodes are populated by noninformative features, enriched RF used a simple adjustment to choose the eligible subsets at each node by weighted random sampling instead of simple random sampling. When the feature space is huge and the ratio of noisy features is large, the performance of the base classifiers degrades. This is because, almost all eligible features at each node,

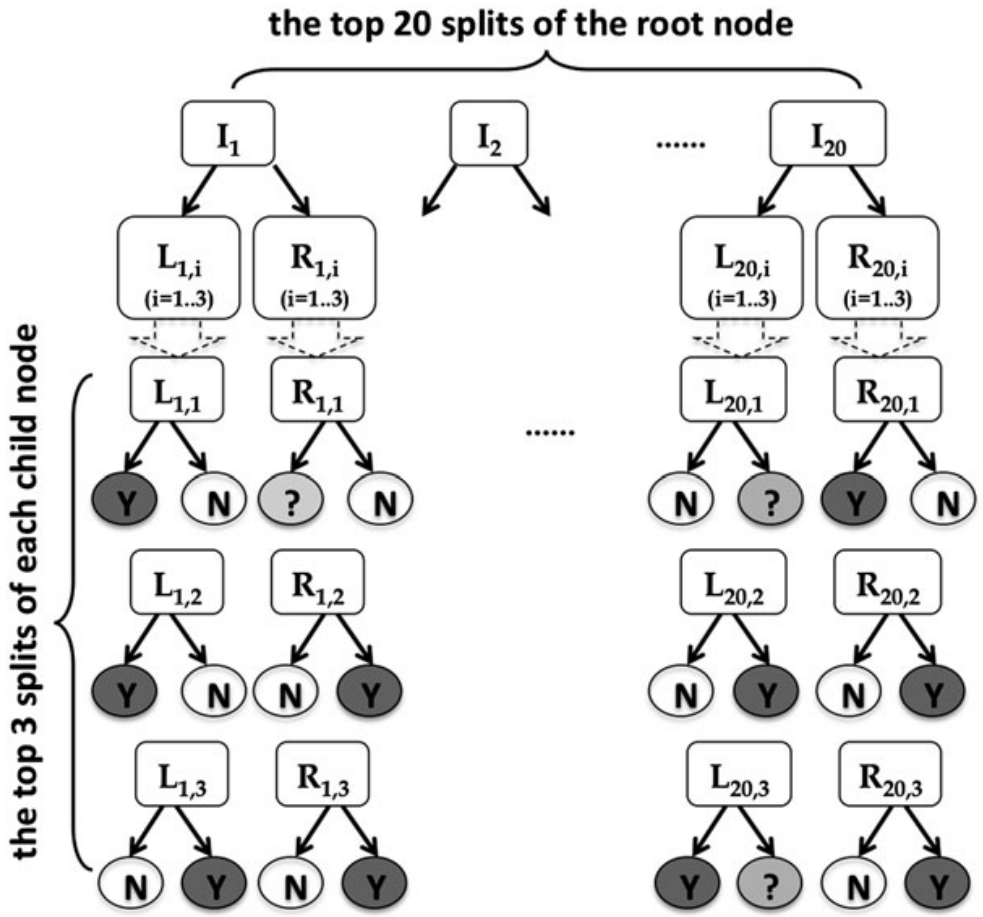

Fig. 11.1 A schematic illustration of the "deterministic forest" method for binary classification (proposed in [48]). $I_1, \ldots, I_{20}$ are the top 20 splits of the root node. Each of these top splits leads to a left (L) and a right (R) child node. The child nodes have their own splits ($L_{j,i}$ and $R_{j,i}$, where $j \in \{1, \ldots, 20\}$ and $i \in \{1, 2, 3\}$). Three top splits are drawn underneath each of them. Based on the combinations of the root splits ($I_{1,\ldots,20}$) and the child splits ($L_{j,i}$ and $R_{j,i}$), the method made multiple trees with different terminal nodes (circles). The terminal nodes are color coded based on the counts of two classes. The more positive examples a terminal node has, the more black it is. Nodes with "?" contain examples from both classes

are predominated by noninformative ones. This issue can be remedied by using weighted, instead of simple, random sampling. By utilizing weights tilted in favor of informative features, the odds of trees containing more informative features being included in the forest increases. Consequently, the resultant enriched RF might contain a higher number of better base classifiers, resulting in a better prediction model.

11.2.2 Measuring Feature Importance

The high-dimensional nature of many tasks in bioinformatics has created urgent needs [37] for feature selection techniques. The goal of feature selection in this field are manifold, where the two most important are: (a) to avoid overfitting and improve

model performance, and (b) to gain a deeper insight into the underlying processes that generated the data. The interpretability of machine learning models is treated as important as the prediction accuracy for most life science problems.

Random forest directly performs feature selection while classification rules are built. In bioinformatics, increased attentions of RF have focused on using it for variable selection, e.g., to select a subset of genetic markers relevant for the prediction of a certain disease. Feature importance is used to rank features and there exist many possible ways [11] to define the measure. The following section discusses several commonly used feature importance based on RF in bioinformatics.

11.2.2.1 Gini Importance

The first commonly used importance measure from RF is the Gini importance. Gini importance is directly derived from the Gini index [6] on the resulting RF trees. The RF classifier uses a splitting function called the Gini index to determine which attribute to split on during the tree learning phase. The Gini index measures the level of impurity/inequality of the samples assigned to a node based on a split at its parent. For instance, under the binary classification case, where there are two classes, let p represent the fraction of positive examples assigned to a certain node k and $1 - p$ as the fraction of negative examples. Then, the Gini index at m is defined as:

$$G_k = 2p(1 - p). \tag{11.1}$$

The purer a node is, the smaller the Gini value is. Every time a split of a node is made using a certain feature attribute, the Gini value for the two descendant nodes is less than the parent node. A feature's Gini importance value in a single tree is then defined as the sum of the Gini index reduction (from parent to children) over all nodes in which the specific feature is used to split. The overall importance in the forest is defined as the sum or the average of its importance value among all trees in the forest.

Learning on biological data is often characterized by a large number of features and few available examples. As a simple estimate of the feature importance for the prediction task, RF Gini feature importance is a popular choice used in biological data mining tasks [37]. However, recent reports [41] pointed out that Gini measures are biased in favor of variables taking more categories if predictors are categorical.

11.2.2.2 Permutation Based Variable Importance

RF permutation importance [11] is another important feature ranking measure when using RF for feature selections. Before introducing this concept, the term of "out-of-bag (OOB) samples" need to be explained. RF does not use all training samples when constructing an individual tree. This leaves a set of OOB samples, which could be used to derive the validated classification accuracy from the tree. RF permutation

importance is measured by randomly permuting the feature variables and computing the increase in OOB estimate of the accuracy loss. Specifically, to measure a feature k's importance in RF trees, the values of this feature is randomly shuffled in the OOB samples. If we use V_k to describe the difference of the classification accuracy between the intact OOB samples and the OOB samples with the particular feature permutated, RF "permutation importance" [6] for feature k is then defined as the average of V_k over all trees in the forest.

RF permutation importance covers the impact of each variable individually while considering multivariate interactions with other features at the same time. It uses an intuitive permutation strategy, and is utilized more frequently than Gini importance in the general "random forest" literature. However, it is time consuming to compute and its magnitude does not have a bounded value range which can be negative. Similar to Gini importance, RF permutation importance was also shown to unreliable when potential variables vary in their scale of measurement or their number of categories [41].

11.2.2.3 Revised RF Feature Importance

The shortcomings mentioned in above two subsections led to several recent variants of RF feature importance from bioinformatics community. Chen et al. [9] proposed the so-called "depth importance" measure to reflect the quality of the node split which is similar to the Gini importance. The major difference is that the depth importance takes into account the position of the node in the trees. It is claimed to be effective in identifying risk genes responsible for complex diseases.

In another notable work, Strobl et al. [41] proposed a revised RF model based on conditional inference trees [21] (pruned trees using stopping criteria based on multiple test procedures). The revised RF provides unbiased variable selection in each individual classification tree. Using subsampling without replacement, the resultant variable importance was claimed to provide reliable variable selection even when the potential variables vary in their scales or vary in the number of categories.

Later, Strobl et al. [40] pointed out another issue of RF variable importance which shows a bias toward correlated predictor variables. The issue of correlated feature variables happens commonly in high-dimensional bioinformatics tasks, e.g., genomics. This paper [40] developed a conditional permutation scheme which used the partition automatically provided by the fitted model as a conditioning grid. The resulting measure was claimed to reflect the true impact of each predictor (variable) better than the original, marginal approach. Simulation results proved that even though the conditional permutation cannot entirely eliminate the preference of correlated predictor variables, it provides a more fair way of comparison that can help to identify the truly relevant feature variables.

Most RF importance measures reflect the average contributions among all trees in a forest. Recently measures based on extreme statistic in a forest are proposed as well. A good example is the "maximal conditional chi-square importance" from [44]. For a specific feature it is defined as the maximal chi-square statistic among

all nodes' splits in a forest. This score was shown to improve the performance of RF when using top-ranked features to refit RF. It was claimed to be more powerful in identifying feature interactions based on simulation studies [44].

More recently, Altmann et al. [1] introduced a heuristic scheme for normalizing feature importance measures that can correct the feature importance bias. The method normalizes the biased RF measure based on a permutation test and returns significance *P*-values for each feature. The repeated permutations are applied on the response vector to preserve the relations between features. The *P*-value of the observed importance provides a corrected measure that addresses the importance bias issue. An improved RF model was then retrained to use top-ranked significant variables with respect to the proposed new importance and was shown to improve the prediction accuracy.

11.2.3 Random Forest Proximity

RF could provide the measure of pairwise proximity between examples using the trained forest. More specifically, for a given forest f and two samples x_i and x_j, the RF similarity is calculated by the following procedure. First, we propagate the value of each sample down all trees within f. Next, the terminal node position for each sample in each of the trees is recorded. Let $z^{(i)} = (z_1^{(i)}, \ldots, z_K^{(i)})$ be these tree node positions for x_i and similarly define $z^{(j)}$ for sample x_j. Then the similarity between x_i and x_j is set to:

$$S(x_i, x_j) = \frac{\sum_{k=1}^{K} I\left(z_k^{(i)} == z_k^{(j)}\right)}{K}, \tag{11.2}$$

where $I(\cdot)$ is the indicator function. As proposed by [6], the sample proximity from RF could be utilized to remove outlier data samples. The noise issue commonly exists in bioinformatics data sets. This strategy has been proved successful in predicting drug response for cell-line gene expression data by removing outlier cell lines in [36].

RF proximity in bioinformatics can also be used for certain classification tasks where the train set provides no negative examples and exhibits a highly skewed distribution between positive and negative classes. For these prediction tasks, relative ranking among predictions normally matter and the cost associated with various classes are different. In order to overcome the issue of problematic training sets and achieve good relative ranking, Qi et al. [35] converted the classification into a ranking task and handled it with a two-step approach using RF proximity. First, it computes a similarity measure between a pair of samples. Then, this measure is used to rank samples by a weighted *k*-nearest-neighbor (KNN) approach. The proposed method has claimed to work well for the PPI prediction in yeast.

11.3 Bioinformatic Applications of Random Forest and Variants

In the past decade, RFs have been successfully applied to various problems in computational biology. The popularity of RFs in this field arises from the fact that RF can be applied to a wide range of data types, even if the problems are nonlinear or involve complex high-order interaction effects. RF and its variants have been applied on a variety of bioinformatic problems, such as gene expression classification, mass spectrum protein expression analysis, biomarker discovery, sequence annotation, PPI prediction, or statistical genetics. The following survey tries to cover some representative applications.

11.3.1 Analysis of Microarray Gene Expression Data

The advent of DNA microarray technology [37] has enabled researchers to measure the expression levels of large numbers of genes simultaneously. The resultant large-scale data sets have stimulated a large body of research in bioinformatics which also created great challenges for computational techniques. Most microarray gene expression data sets suffer from the commonly known "curse-of-dimensionality" issue where the dimensionality is huge (up to several tens of thousands of genes), and the sample size is small (normally up to hundreds). Moreover, high ratio of noise and variability from microarray experiments raise even more challenges. As shown in Fig. 11.2, computational methods normally treat the microarray data as an $N \times M$ matrix, where N is large, M is small, and $N \ll M$.

One important task in biomedical research is to distinguish disease samples from nondisease samples as well as to classify different disease subtypes [39]. The sample could be a patient, a tissue, or even tissue parts whose features are expressed values of a set of genes or proteins, i.e., the so-called "molecular signature or profile." For using gene expression data to classify disease versus nondisease samples, Lee et al. [26] carried over an extensive study to compare the KNN approach, various versions of linear discriminant analysis (LDA), bagging trees, boosting, and RFs under the same experimental settings. They found that RF was the most successful technique used on the seven microarray data sets they tested.

A closely related popular topic tries to identify a set of biomarkers (normally genes) from gene expression datasets that could maintain high classification accuracy of samples when used alone. Fast and efficient feature selection techniques have attracted lots of attentions since the related data sets are high-dimensional and small. Gene–gene interactions are importance factors to consider when selecting features for disease classification; however, popular univariate selection methods could not take them into account. Thus, researchers have proposed a number of techniques to capture the correlations between genes using RFs based variable importance [2, 15, 40, 46]. Several related methods have been covered in Subsect. 11.2.2. These

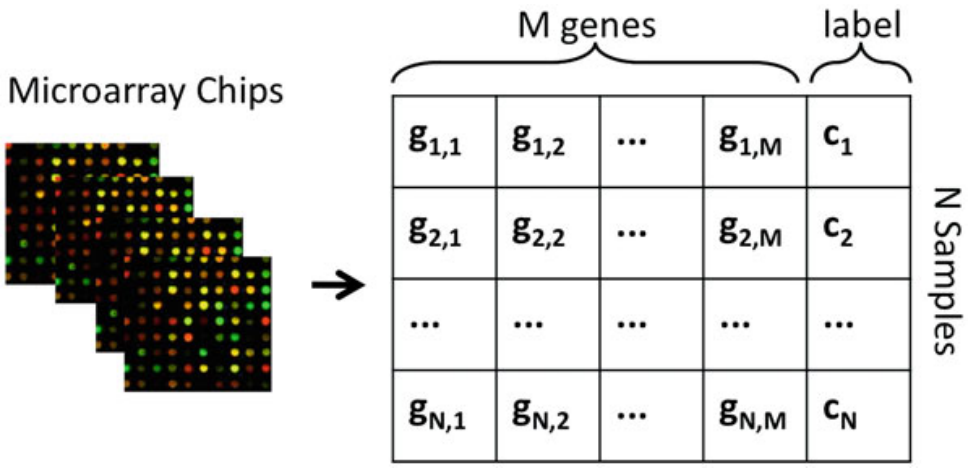

Fig. 11.2 Schematic illustration for gene expression of microarray data. Figure modified from [47]. From the computational perspective, the microarray data is described as an $N \times M$ matrix. Each row describes a sample and each column represents a gene except the last column which means the class label of each sample. $g_{i,j}$ is a numeric value representing the gene expression level of gene j in the ith sample. c_i is the class label of the ith sample [47]

importance measures could be used to filter the original feature set and then the classification model could be retrained which might be a better fit. For instance, the "enriched random forest" method, proposed by Amaratunga et al. [2], claims to improve the RF performance on ten real gene expression data sets by selecting top-ranked features using a weighted random sampling scheme for biomedical sample classification. Diaz-Uriarte et al. [15] showed that RF is able to preserve predictive accuracy while yielding smaller gene sets selected for the analysis of microarray data when compared to LDA, KNN, and SVM.

In summary as an important subfield in bioinformatics, using gene expression microarray has emerged as popular tools to identify common genetic factors that influence health and disease. Random forest methods and its feature importance measures provide the state-of-art performance for analyzing and identifying patients' molecular profiles from gene expression data sets.

11.3.2 Analysis of Mass Spectrometry-Based Proteomics Data

Modern mass spectrometry technologies allow the determination of proteomic fingerprints (e.g., expression levels of many proteins) of body fluids like serum or urine. Differently from DNA microarrays which only relate to genetic (static) factors of diseases, mass spectrum measurements can be used to diagnose the dynamic status or to predict the evolution of a disease. In modern biology, mass spectrometry technology grows to be an attractive framework for cancer diagnosis and protein-based biomarker detection [5].

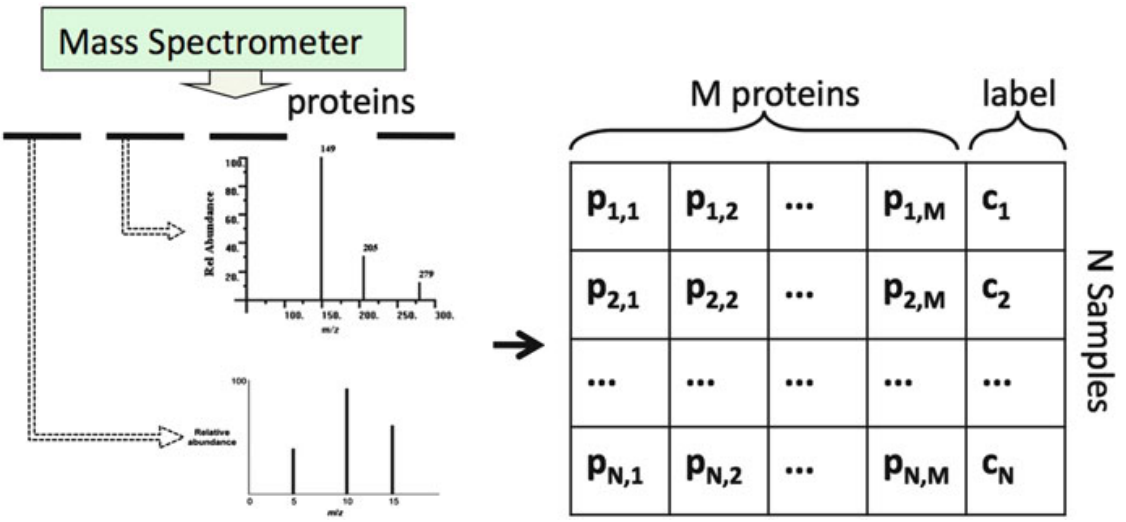

Fig. 11.3 Schematic illustration of mass spectrometry-based proteomics data sets. Figure modified from [47]. The proteomics data generated by mass spectrometer are very similar to gene microarray data in terms of the computational analysis. Differently from microarray data describes the abundance of a protein or peptide in the sample

Figure 11.3 provides a schematic description of mass spectrometry-based proteomics data sets. A typical mass spectrum sample is characterized by thousands of different mass/charge (m/z) ratios on x-axis, and their corresponding signal intensity values are on y-axis. A set of samples' mass spectrum features are treated as a data matrix by computational mining methods. Such mass spectrum data sets are also characterized by a small number of samples and a very high-dimensional feature space. Like DNA microarray data, this "curse-of-dimensionality" issue requires the computational algorithm to select the most relevant features and to make the most use of the limited data samples [47].

Random forest holds a unique position in analyzing mass spectrometry-based proteomics data for clinical classifications [18, 20, 22–24], since it considers feature interactions in learning and is well suited for high-dimensional data samples. For instance, RF has been demonstrated by Izmirlian et al. [22] in classifying SELDI–TOF (surface-enhanced laser desorption/ionization time of flight) proteomic data well with the advantages of robustness to noise and less dependence on tuning parameters. Later, Geurts et al. [18] presented a related tree ensemble approach named "extra trees" [17] which selects at each node the best among K randomly generated splits. Unlike RFs which are grown with multiple sample subsets, the base trees of extra trees are grown from the complete training set and by explicitly randomizing the splits. The approach was successfully validated on two SELDI-TOF data sets for the diagnosis of rheumatoid arthritis and inflammatory bowel diseases.

Recently, Kirchner et al. [24] showed that a RF-based approach is feasible to achieve real-time classification of fractional mass in mass spectrometry experiments. Similarly, Karpievitch et al. [23] proposed a modified RF, named as "RF++" to deal with cluster-correlated data. Many mass spectrometry-based studies produce cluster-correlated data where there exist replicated samples for the same subject.

A common practice for dealing with replicated data is to average each subject's replicate sample set, which will reduce the data set size and might incur loss of information. However, failure to account for correlation among samples may result in overfitting of the training data and producing over optimistic error estimations. Two strategies were utilized in RF++ to tackle this issue [23]: (1) a modified RF grown using subject-level averages, and (2) a modified RF using subject-level bootstrapping to substitute the original resampling step. The second scheme was shown to be effective for classifying clustered mass-spectrum proteomics data.

11.3.3 Genome-Wide Association Study

Like gene expressions from microarray experiments and protein expressions from mass-spectrum based technologies, comparing the genomes (whole DNA sequences) of different samples can also give critical information of different diseases [47]. More importantly, such studies, termed as "genome-wide association study" (GWAS), can help to determine the susceptibility of each different individual to complex diseases, as well as the response to different drugs based on individuals' genetic variations [45].

With the revolutionary advancements of next-generation sequencing technologies, huge volumes of high-throughput sequence data have become easily obtained and extremely cheap. This information has largely enhanced biologists' knowledge of many organisms and also expanded the impact of the genomes on biomedical research. Genomewide association study is becoming increasingly important for clinical decision support with respect to the diagnosis of complex diseases [45].

GWAS computational task involves scanning markers across the complete sets of DNA sequences, or genomes, from many people to find genetic variations associated with a particular disease or a biological symptom. One important concept in GWAS is the so-called "SNPs" (single nucleotide polymorphisms), which is generated from the following procedure. GWAS studies normally compare two groups of samples, (people with or without the disease) by extracting DNA from each person's sample of cells. DNA is then spread on gene chips which could read millions of DNA sequences. Rather than reading the entire DNA sequence, GWAS usually reads the SNPs which are markers indicating the DNA sequence variation at a single nucleotide position. It is estimated that the human genome has approximately seven million of SNPs [25].

To fully understand the basis of complex disease, it is critical to identify the important genetic factors involved, and the complex relationships between these factors. Many complex diseases such as diabetes, asthma, or cancer arise from a combination of multiple genes which often regulate and interact with each other to produce the disease. Therefore, the goal of studying GWASs for these diseases is to identify the complex interactions among multiple SNPs and together with environmental factors which may substantially increase the risk of developing these diseases [45]. This difficult task is commonly formulated into simpler tasks which

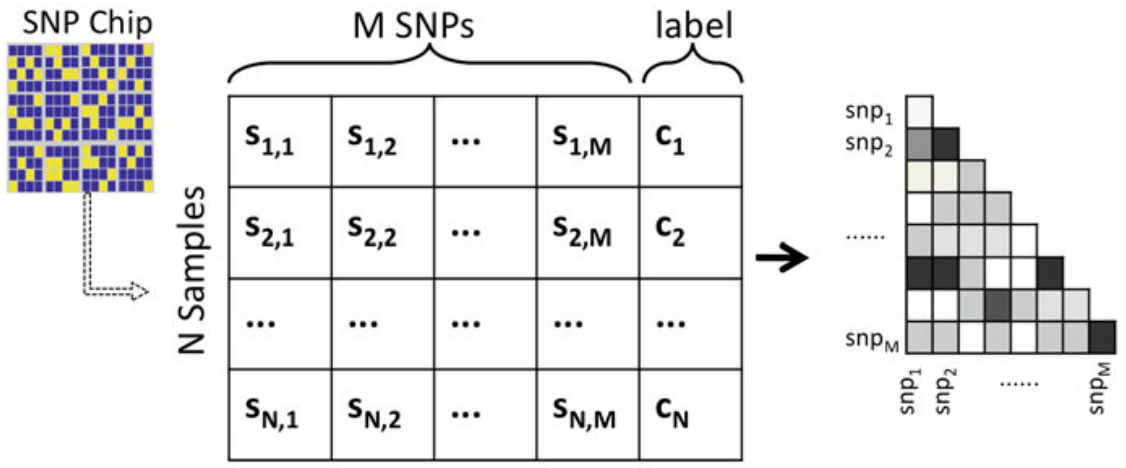

Fig. 11.4 Schematic illustration of pairwise SNP–SNP interaction effects on sample classification. The data matrix obtained from the SNP chip is similar to DNA microarray studies except that each column describes a SNP variable. The pairwise SNP–SNP interactions are schematically illustrated as the gray boxes in the right heat map where darker colors indicating stronger interactions and associations with the disease of interest. Figure modified from [47]

try to identify pairwise SNP–SNP interactions or SNP-environment interactions. Figure 11.4 provides a schematic illustration of pairwise interaction relationship between multiple SNPs. Again, the set of samples (N) and their SNP features (M) could be treated as a data matrix from computational perspective (see Fig. 11.4).

Owing to the intrinsic ability to consider multiple SNPs jointly in a nonlinear fashion [32], RF [6] has become a popular choice of many recent GWAS studies for SNP–SNP interaction identification [3, 4, 9, 30, 45]. Using the feature importance estimated from RF, it is possible to identify important SNP subsets that are associated with the outcome of the disease.

RF is especially useful to identify features that show small marginal contributions individually, but gives a larger effect when combined together. For example, the initial attempt from [28] utilized RF permutation importance (Subsect. 11.2.2.2) as a screening procedure to identify small numbers of risk-associated SNPs among large numbers of unassociated SNPs using 16 complex disease models. RF was concluded to outperform Fisher's exact test when interactions between SNPs exist. Later, a similar study from Bureau et al. [7] used a similar RF importance measure and extended the concept on pairs of predictors, in order to capture joint effects. These early studies normally limited the number of SNPs under analysis to a relatively small range (30).

Recent studies developed feature importance variants from RF to a much larger dimensional range, e.g., several hundred thousands of candidate SNPs. Besides, the issue of correlated variables are also taken into account which commonly exist in GWAS data. Cheng et al. [9] investigated the power of random forests in identifying SNP interaction pairs by proposing the "depth importance" measure (Subsect. 11.2.2.3) from RF trees. It was applied to analyze the complex disease of age-related macular degeneration. Later, Wang et al. [44] proposed an alternative

importance measure, "maximal conditional chi-square" (MCC in Subsect. 11.2.2.3), for feature selection in GWASs. MCC measures the association between a SNP and the outcome where the association is conditional on other SNPs. The method estimated empirical *P*-values of SNPs by revising the RF permutation importance. Compared with the existing importance measures, the MCC importance showed more sensitivity to complex effects of risky SNPs.

Both GWASs and biomarker discovery involve feature selection technology and therefore they are closely related to each other [47]. However, they have different goals with respect to feature selection. The objective of biomarker discovery is to find a small set of biomarkers (e.g., genes or proteins) to achieve good prediction accuracies. This allows the development of cheaper and more efficient diagnostic tests. Instead, the goal in GWASs is to find important genetic factors that are associated with the outcome symptoms and to estimate the significance level of the association.

11.3.4 Protein–Protein Interaction Prediction

Protein–protein interactions are critical for virtually every biological function in the cell. However, experimental determination of pairwise PPIs is a labor-intensive and expensive process. Therefore, predicting PPIs from indirect information is an active field in computational biology. Recently, researchers suggested supervised learning for the task of classifying pairs of proteins as interacting or not. Three independent studies [10, 27, 33] compared the performance of multiple classifiers in predicting protein interactions. In all three studies, RF achieved the best performance on this task when integrating various biological features such as gene expression, gene ontology features, and sequence data. Figure 11.5 shows a schematic illustration of how a RF performs information integrations for the task of classifying pairs of proteins as interacting or not in yeast.

Most of the early studies have been carried out in yeast or in human [34], which aimed to predict protein interactions within a single organism (called "intraspecies PPI prediction"). More recently, researchers extended RF to predicting PPIs between organisms (called "interspecies PPI prediction"), especially between host and pathogens. For instance, Tastan et al. [43] applied the supervised RF classification framework to predict PPIs between HIV-1 viruses and human proteins. By integrating multiple biological information sources, RF defined the state-of-art performance for this task. Figure 11.6 shows a schematic illustration of protein interactions between HIV-1 and human proteins.

11.3.5 Biological Sequence Analysis

Computational analysis of biological sequences is a classic and still expanding subfield in bioinformatics. Biological sequence describes continuous chains of nucleotide acids (DNA) or amino acids (protein) which can be categorized based

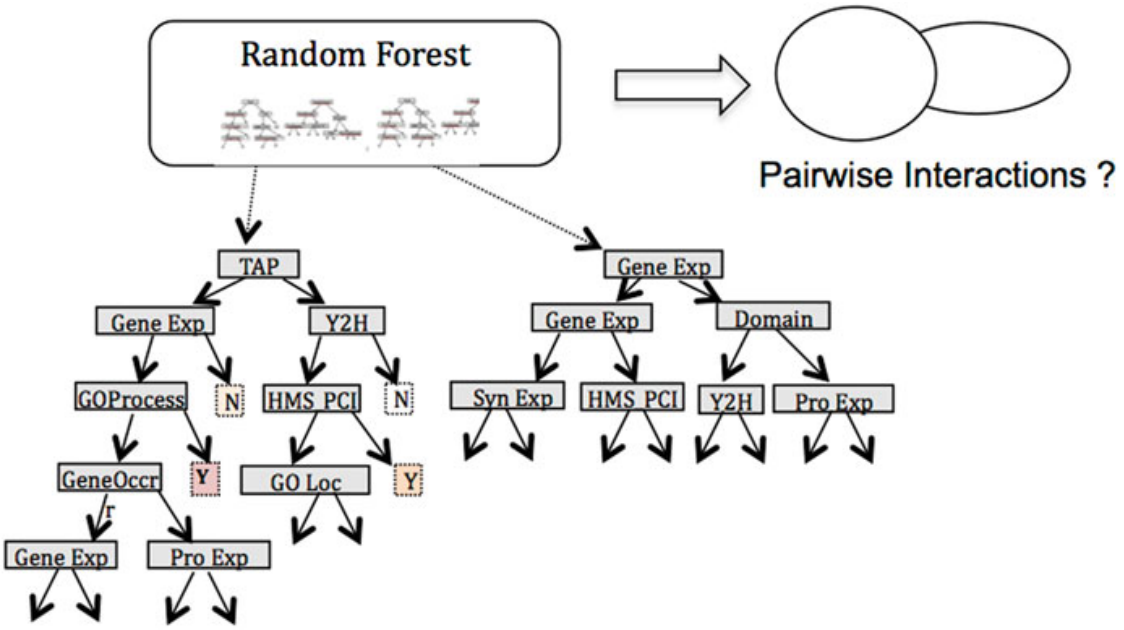

Fig. 11.5 Evidence was integrated using a random forest classifier for protein–protein interaction prediction. Figure modified from [35]

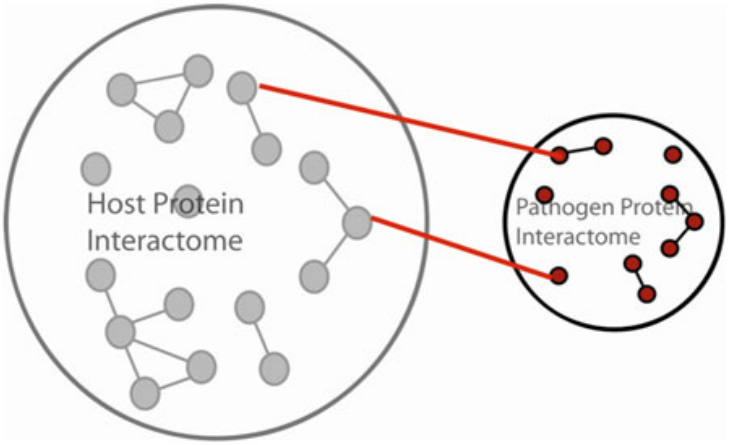

Fig. 11.6 Schematic illustration of protein–protein interactions between HIV-1 (rightside) and human proteins (leftside). Figure modified from [43]

on the underlying molecule type: DNA, RNA, or protein sequence. Since more and more species genomes have been sequenced, this area remains one of the most important in bioinformatics. With biological mutations and evolution, sequence data sets are usually enormous and complex, where efficient and accurate learning models become critical factors [8].

Though there exist enormous biological sequence mining tasks, this section covers only four typical ones where RF achieved good results. All these tasks try to computationally identify the functional properties of subregions (sites) of DNA or protein sequences.

The first type of task is to predict the phenotypes (symptoms) based on protein sequence or DNA sequence. Segal et al. [38] utilized RFs to predict the replication capacity of viruses, such as HIV-1, based on amino acid sequence from reverse transcriptase and protease. Similarly, Cummings et al. [13] used RFs to model the

relationships between the amino acid sequence of gene "rpoB" and the rifampin resistance ("rifampin" is a bactericidal antibiotic drug). Gene "rpoB" is the gene encoding the beta subunit of DNA-dependent RNA polymerase.

The second related task tries to cope with RNA editing. RNA editing represents the process whereby RNA is modified from the sequence of the corresponding DNA template. For instance, cytidine-to-uridine conversion (abbreviated as C-to-U conversion) is common in plant mitochondria. The mechanisms of this conversion remain largely unknown, although the role of neighboring nucleotides is emphasized. Cummings et al. [12] suggested to use information from subregions' flanking sites of interest to predict if C-to-U editing happens on mitochondrial RNA sequences. Random forest was applied for this prediction task in three plant species: "*Arabidopsis thaliana*", "*Brassica napus*", and "*Oryza sativa* [12]". Recently, Strobl et al. [41] proposed to work on the same C-to-U editing task by employing a revised RF method based on learning conditional inference trees.

The third typical biosequence task RF has been applied to the identification of "Post translational modifications (PTMs)." PTMs occur in a vast majority of proteins and are essential for certain protein functions. Prediction of the sequence location of PTMs is an important step in understanding the functional characterization of proteins [19]. Among many possible PTMs, glycosylation site and phosphorylation site are the two critical kinds of functional sites in protein sequences. Their accurate localization can elucidate many important biological process such as protein folding, subcellular localization, and protein transportation. Hamby et al. [19] utilized the random forest algorithm for glycosylation sites prediction and prediction rule extraction for yeast. Their work made use of the pairwise patterns surrounding glycosylation sites for better predictions. The authors claimed to observe a significant increase of prediction accuracy in the prediction of "Thr" and "Asn" glycosylation sites.

The last task to cover in this section is associated with HIV-1 viruses. Human Immunodeficiency Virus (HIV) is the pathogen causing the disease AIDS. The invasion of HIV-1 Virus into human cells relies on the contact of its glycoprotein "gp120" with two human cellular proteins, a receptor, and a coreceptor. The type of coreceptor is crucial for the aggressiveness of the virus and the available treatment options. Hence, Dybowski et al. [16] proposed to predict coreceptor usage based on the viral genome sequences. A random forest-based method is developed to predict coreceptor usage for new sequences using structures and sequences of "gp120." The good accuracy achieved in [16] made random forest a strong candidate for computational diagnosis of viral diseases.

11.3.6 Some Other Related Applications

Moreover, RF has been tried on many other biomedical domains. For instance, RF [14] shows to be a powerful statistical classifier in computational ecology. Cutler et al. [14] compared the accuracies of RF and four other commonly used

statistical classifiers on three different ecological data sets describing: (1) invasive plant species' presence in US California, (2) the rare lichen species' presence in the US Pacific Northwest, and (3) the nest sites for cavity nesting birds in Utah. RF showed high classification accuracy in all three applications.

Another interesting application is for computational drug screening [29, 36], where panels of cell lines are used to test drug candidates for their ability to inhibit proliferation. Riddick et al. [29] built regression models using RF to predict drug response for 19 Breast Cancer and 7 Glioma cell lines. RF was used in three specific ways: (1) feature selection of drug gene expression signatures based on RF permutation importance, (2) removing outlier cell lines based on RF proximity, and (3) RF multivariate regression model for predicting continuous drug response.

More applications of RFs can be found in other different fields like quantitative structure-activity relationship modeling [42], nuclear magnetic resonance spectroscopy [31], or clinical decision supports in medicine in general [11].

11.4 Summary

With the data explosion in modern biology, machine learning algorithms are becoming increasingly popular. Since the data complexity is always rising, as a nonparametric model, RF provides a unique combination of prediction accuracy and model interpretability. This chapter mainly focused on explaining the notable extensions and applications of RF in bioinformatics. The covered references are by no means an exhaustive list, but are topics which have received much attention. We therefore sincerely apologize to related papers that are not covered in this chapter.

References

1. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340 (2010)
2. Amaratunga, D., Cabrera, J., Lee, Y.: Enriched random forests. Bioinformatics **24**(18), 2010 (2008)
3. Bao, L., Zhou, M., Cui, Y.: nssnpanalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Research **33**(suppl 2), W480 (2005)
4. Barenboim, M., Masso, M., Vaisman, I., Jamison, D.: Statistical geometry based prediction of nonsynonymous snp functional effects using random forest and neuro-fuzzy classifiers. Proteins: Structure, Function, and Bioinformatics **71**(4), 1930–1939 (2008)
5. Barrett, J., Cairns, D.: Application of the random forest classification method to peaks detected from mass spectrometric proteomic profiles of cancer patients and controls. Statistical Applications in Genetics and Molecular Biology **7**(2), 4 (2008)
6. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). DOI 10.1023/A:1010933404324
7. Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., Van Eerdewegh, P.: Identifying snps predictive of phenotype using random forests. Genet Epidemiol **28**(2), 171–82 (2005). DOI 10.1002/gepi.20041

8. Chen, X., Jeong, J.: Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics **25**(5), 585 (2009)
9. Chen, X., Liu, C.T., Zhang, M., Zhang, H.: A forest-based approach to identifying gene and gene–gene interactions. Proc Natl Acad Sci USA **104**(49), 19,199–203 (2007). DOI 10.1073/pnas.0709868104
10. Chen, X., Liu, M.: Prediction of protein–protein interactions using random decision forest framework. Bioinformatics **21**(24), 4394 (2005)
11. Chen, X., Wang, M., Zhang, H.: The use of classification trees for bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(1), 55–63 (2011)
12. Cummings, M., Myers, D.: Simple statistical models predict c-to-u edited sites in plant mitochondrial rna. BMC Bioinformatics **5**(1), 132 (2004)
13. Cummings, M., Segal, M.: Few amino acid positions in rpob are associated with most of the rifampin resistance in mycobacterium tuberculosis. BMC Bioinformatics **5**(1), 137 (2004)
14. Cutler, D., Edwards Jr, T., Beard, K., Cutler, A., Hess, K., Gibson, J., Lawler, J.: Random forests for classification in ecology. Ecology **88**(11), 2783–2792 (2007)
15. Diaz-Uriarte, R., de Andrés, S.: Variable selection from random forests: application to gene expression data. Arxiv preprint q-bio/0503025 (2005)
16. Dybowski, J.N., Heider, D., Hoffmann, D.: Prediction of co-receptor usage of hiv-1 from genotype. PLoS Comput Biol **6**(4), e1000,743 (2010). DOI 10.1371/journal.pcbi.1000743
17. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**, 3–42 (2006)
18. Geurts, P., Fillet, M., De Seny, D., Meuwis, M., Malaise, M., Merville, M., Wehenkel, L.: Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics **21**(14), 3138 (2005)
19. Hamby, S., Hirst, J.: Prediction of glycosylation sites using random forests. BMC Bioinformatics **9**(1), 500 (2008)
20. Hanselmann, M., Ko the, U., Kirchner, M., Renard, B., Amstalden, E., Glunde, K., Heeren, R., Hamprecht, F.: Toward digital staining using imaging mass spectrometry and random forests. Journal of Proteome Research **8**(7), 3558–3567 (2009)
21. Hothorn, T., Hornik, K., Zeileis, A., Wien, W., Wien, W.: Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics **15**(3), 651–674 (2006)
22. Izmirlian, G.: Application of the random forest classification algorithm to a seldi-tof proteomics study in the setting of a cancer prevention trial. Annals of the New York Academy of Sciences **1020**(1), 154–174 (2004)
23. Karpievitch, Y., Hill, E., Leclerc, A., Dabney, A., Almeida, J.: An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of rf++. PloS one **4**(9), e7087 (2009)
24. Kirchner, M., Timm, W., Fong, P., Wangemann, P., Steen, H.: Non-linear classification for on-the-fly fractional mass filtering and targeted precursor fragmentation in mass spectrometry experiments. Bioinformatics **26**(6), 791 (2010)
25. Kruglyak, L., Nickerson, D.A.: Variation is the spice of life. Nat Genet **27**(3), 234–6 (2001). DOI 10.1038/85776
26. Lee, J., Lee, J., Park, M., Song, S.: An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis **48**(4), 869–885 (2005)
27. Lin, N., Wu, B., Jansen, R., Gerstein, M., Zhao, H.: Information assessment on predicting protein–protein interactions. BMC Bioinformatics **5**(1), 154 (2004)
28. Lunetta, K., Hayward, L., Segal, J., Van Eerdewegh, P.: Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics **5**(1), 32 (2004)
29. Ma, Y., Ding, Z., Qian, Y., Shi, X., Castranova, V., Harner, E., Guo, L.: Predicting cancer drug response by proteomic profiling. Clinical Cancer Research **12**(15), 4583 (2006)
30. Meng, Y., Yu, Y., Cupples, L., Farrer, L., Lunetta, K.: Performance of random forest when snps are in linkage disequilibrium. BMC Bioinformatics **10**(1), 78 (2009)

31. Menze, B., Kelm, B., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.: A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics **10**(1), 213 (2009)
32. Moore, J., Asselbergs, F., Williams, S.: Bioinformatics challenges for genome-wide association studies. Bioinformatics **26**(4), 445 (2010)
33. Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J.: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics **63**(3), 490–500 (2006)
34. Qi, Y., Dhiman, H., Bhola, N., Budyak, I., Kar, S., Man, D., Dutta, A., Tirupula, K., Carr, B., Grandis, J., et al.: Systematic prediction of human membrane receptor interactions. Proteomics **9**(23), 5243–5255 (2009)
35. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random forest similarity for protein–protein interaction prediction from multiple sources. In: Proceedings of the Pacific Symposium on Biocomputing (2005)
36. Riddick, G., Song, H., Ahn, S., Walling, J., Borges-Rivera, D., Zhang, W., Fine, H.: Predicting in vitro drug sensitivity using random forests. Bioinformatics **27**(2), 220 (2011)
37. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507 (2007)
38. Segal, M.R.: Machine learning benchmarks and random forest regression. Technical Report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco (2004)
39. Statnikov, A., Wang, L., Aliferis, C.: A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics **9**(1), 319 (2008)
40. Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinformatics **9**(1), 307 (2008)
41. Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics **8**(1), 25 (2007)
42. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P.: Random forest: a classification and regression tool for compound classification and qsar modeling. J Chem Inf Comput Sci **43**(6), 1947–58 (2003). DOI 10.1021/ci034160g
43. Tastan, O., Qi, Y., Carbonell, J., Klein-Seetharaman, J.: Prediction of interactions between HIV-1 and human proteins by information integration. In: Pac Symp Biocomput, vol. 516 (2009)
44. Wang, M., Chen, X., Zhang, H.: Maximal conditional chi-square importance in random forests. Bioinformatics **26**(6), 831 (2010)
45. Wang, W.Y.S., Barratt, B.J., Clayton, D.G., Todd, J.A.: Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet **6**(2), 109–18 (2005). DOI 10.1038/nrg1522
46. Wu, X., Wu, Z., Li, K.: Identification of differential gene expression for microarray data using recursive random forest. Chin Med J **121**(24), 2492–2496 (2008)
47. Yang, P., Hwa Yang, Y., Zhou, B., Zomaya, Y., et al.: A review of ensemble methods in bioinformatics. Current Bioinformatics **5**(4), 296–308 (2010)
48. Zhang, H., Yu, C., Singer, B.: Cell and tumor classification using gene expression data: construction of forests. Proceedings of the National Academy of Sciences **100**(7), 4168 (2003)