Prediction of Disease-Associated Single Amino Acid Polymorphisms Based on Physiochemical Features

Jiaxin Wu, Mingxin Gan, Wangshu Zhang, and Rui Jiang

Abstract—**Benefiting from recent advancements of the next generation sequencing technology, it becomes more and more feasible to directly sequence candidate genetic regions and even the whole genome to get the information about rare genetic variants. Although several statistical methods have been developed to identify potential associations between multiple rare variants and a given disease of interest, these methods are quite sensitive to the inclusion of non-functional variants in their statistical analysis. In order to enhance the performance of these statistical methods for uncovering disease-associated rare variants, it is suggested that bioinformatics tools or filters should be adopted to make functional predictions of the variants before statistical analysis. In this paper, we propose to prioritize candidate genetic variants according to the guilt-by-association principle, which depends on the assumption that genetic variants associated with the same disease share some common physiochemical properties. Focusing on a specific type of genetic variants called single amino acid polymorphisms (SAAPs), we take advantages of 8 similarity measures based on physiochemical features of amino acids, sequence information of proteins, and multiple sequence alignment of protein families to illustrate the power of prioritizing candidate SAAPs for specific diseases. Systematic validation experiments demonstrate that our proposed approach is competent for effectively detecting associations between SAAPs and query diseases, while using the Canberra distance to measure the similarity between SAAPs can achieve the highest performance among all the methods compared.**

Index Terms—**Guilt-by-Association,Similarity, Single Amino Acid Polymorphisms (SAAPs), Prioritization.**

I. Introduction

Genome-wide association (GWA) studies have achieved remarkable successes in uncovering the relationship between common genetic variants and human inherited diseases in the last decade [1]. Typically, a GWA study focuses on detecting associations between genetic variants and some observable clinical traits of a specific type of disease by comparing the frequencies of occurrence of genetic variants between a case population and a control population. However, with the advancing next generation sequencing technology, the basic hypothesis of a GWA study, which assumes that the etiology of common diseases is intervened by commonly occurring genetic variants with small to modest effects [2], has been challenge by the fact that both common variants and rare mutations may be involved in the pathogenesis of common diseases. Some studies also point out that multiple rare variants with moderate to high penetrance may work in concert with each other to have stronger phenotypic effects [3]. According to these researches, a common disease-rare variant (CD-VR) hypothesis that indicates that multiple rare variants can also serve as the main factor to influence some common diseases has been proposed.

Even though many existing experimental methods and computational approaches have been proposed for GWA studies and have shown reasonably powerful performances, they may be not competent in uncovering the relationship between multiple rare variants and diseases due to the specific properties of rare variants, such as the low marginal population attributable risk and the wide range of penetrance [4].With the accelerating advancement of the next generation sequencing technology, it becomes more and more feasible to directly sequence candidate genetic regions or the whole genome to obtain a huge number of rare variants. Accordingly, several statistical methods, such as the combined multivariate and collapsing method [2], the cohort allelic sums test method [5], and the weighted-sum statistic method [3] have been developed to deal with such a huge number of rare variants, as well as simultaneously identify multiple rare variants. Even so, it has been suggested that one should first quantify which variants are potentially functional or neutral before statistical analysis of the sequence data [2]. On this scenario, bioinformatics tools or filters are expected to make functional predictions of the variants in study and then choose the functional variants in the successive statistical analysis.

As a typical type of genetic variants, single nucleotide polymorphisms (SNPs) may lead to single amino acid polymorphisms (SAAPs) in proteins, potentially affect structures and functions of proteins, and further cause human diseases [6]. Many existing popular methods, such as PolyPhen [7], SIFT [8], KBAC [9], and MSRV [13], formulate the identification of SAAPs that are associated with diseases as a binary classification problem and give no information about what specific diseases the SAAP is

Manuscript received June 11, 2011; revised June 29, 2011.This work was partly supported by the Natural Science Foundation of China (60805010, 60928007, 60934004), Tsinghua University Initiative Scientific Research Program, Tsinghua National Laboratory for Information Science and Technology (TNLIST) Cross-discipline Foundation.

Jiaxin Wu is with MOE Key laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.

Mingxin Gan is with School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China.

Wangshu Zhang is with MOE Key laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.

Rui Jiang is with MOE Key laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China. (FIT 1-107, Tsinghua University, Beijing 100084, China. e-mail: ruijiang@tsinghua.edu.cn).

Rui Jiang: To whom correspondence should be addressed. (FIT 1-107, Tsinghua University, Beijing 100084, China.

E-mail: ruijiang@tsinghua.edu.cn)

associated with. Therefore, limited contribution for practical or clinical applications is provided with the only classification results of these methods [10].

In this paper, we formulate the detection of SAAPs that are associated with a specific type of disease as a one-class novelty learning problem. Specifically, we prioritize candidate SAAPs for a query disease using the guilt-by-association principle, which takes advantages of an association score to quantify the strength of association between a candidate SAAP and the query disease and then rank all candidates according to their scores. The association score between a candidate SAAP and a query disease is quantified as the total similarities between the SAAP and all the seeds known as associated with the query disease. We also take advantage of 8 similarity scores to measure the similarity between two SAAPs under the feature space, which is composed of 44 features extracted from physiochemical properties of amino acids and sequence information of proteins. Systematic validation demonstrates that the proposed model is effective in deciphering the relationship between SAAPs and diseases, with the Canberra distance achieving the most precise prediction results.

II. Materials and Methods

A. Data Sources

We mainly use two databases to collect the related information of SAAPs and the protein sequence data where the SAAPs occur. The Swiss-Prot database [11] is used to provide the information of SAAPs and corresponding single amino acid polymorphisms. Specifically, version 2010_10 (released on Oct. 5th, 2010) of this data database collects 62,430 single amino acid polymorphisms that occur in 12,401 human proteins, with each substitution annotated as "Disease," "Polymorphism," or "Unclassified." The single amino acid polymorphisms annotated with "Disease" are considered as *disease* SAAPs and those annotated with "Polymorphism" are treated as *neutral* SAAPs.

The Pfam database is adopted to extract the multiple sequence alignments (MSA) of human proteins. In version 24.0 of Pfam database, (released in Oct. 2009) [12], there are curated alignments and models for 11,912 protein families. In our study, we focus on SAAPs that have corresponding OMIM accession number in the Swiss-Prot database and appear in multiple sequence alignment of the Pfam database. Finally, we collect 13,735 neutral SAAPs and 14,511 disease SAAPs that are associated with 1,575 human diseases.

B. Physiochemical Features

We take advantage of a set of 44 numeric features extracted only from the sequential information of proteins following the literature [13]. The features are derived based on three physicochemical properties (molecular weight, pI value, and hydrophobicity scale) of amino acids, three relative frequencies of occurrences of amino acids in the secondary structures of proteins, and two evolutionary conservation scores obtained from multiple sequence alignment of proteins.

Given an amino acid polymorphism pair in a certain query protein, all above six properties can be calculated in seven conditions, therefore, we can get 42 physicochemical features. The 7 conditions are the six properties for the original amino acids; the six properties for the substitute amino acids; the six properties in a window-sized situation; the six properties in a column-weighted situation; Relative changes from the six properties for the original amino acid to the six properties for the substituted amino acid; Relative changes from the six properties for the window-sized situation to the six properties for the substituted amino acid; Relative changes from the six properties for the column-weighted circumstance to the six properties for the substituted amino acid. In the window-sized situation, these six properties of the original amino acid and of its neighbors in the query protein sequence are averaged. The column-weighted properties are the average of the corresponding properties of all the amino acids in the same column of the Pfam multiple sequence alignment that the substation occurs. Furthermore, we use the conservation scores of the original and the substituted amino acids as features to facilitate the prioritization of candidate SAAPs. These two conservation scores are defined as the frequencies of occurrences of the amino acids (original or substituted) in the corresponding column of the alignment [14]. Consequently, we can get the 42 physiochemical features (6 properties under 7 situations) and two conservation properties.

C. Guilt-by-association Model

We propose to prioritize candidate SAAPs using the guilt-by-association model based on two assumptions that a disease is associated with a set of SAAPs with similar physiochemical features and conservation properties, and a SAAP is likely to be associated with the query disease if the SAAP shares similar common physiochemical properties and conservation properties with the set of seed SAAPs that are known to be associated with the disease. To mathematically construct the guilt-by-association model, we let ω_{ij} be a type of similarity measure between SAAPs i and j under the feature space, and let $\mathbf{S}_d$ be the set of seed SAAPs that are known to be associated with the query disease d. The association score $A(i \leftrightarrow d)$ between a candidate SAAP i and the disease d is then defined as

$$A(i \leftrightarrow d) = \sum_{j \in \mathbf{S}_d} \omega_{ij}$$

In other words, the strength of association between an SAAP and a disease is quantified as the total similarities between the SAAP and all the seeds known as associated with the query disease.

We adopt 8 distance measures to quantify the similarity between two SAAPs in the feature space. Each SAAP can be treated as a point in the high dimensional feature space; therefore, we use some popular distance measures between the two points in the feature space to scale the similarity between two SAAPs.

Firstly, we propose to calculate the reciprocal of the traditional Euclidean distance between two points as the similarity measure between the two corresponding SAAPs $\mathbf{x}$ and $\mathbf{y}$. We have

$$\omega_{\text{Euclidean}} = \|\mathbf{x}-\mathbf{y}\|_2 = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{\frac{1}{2}},$$

where $\|\mathbf{x}\|_2$ is the L_2 norm of a vector and n is the feature dimension of the SAAPs.

Secondly, we propose to use the reciprocal of the Manhattan distance as the similarity measure [15]. We have

$$\omega_{\text{Manhattan}} = \|\mathbf{x}-\mathbf{y}\|_1 = \sum_{i=1}^{n}|x_i - y_i|,$$

where $\|\mathbf{x}\|_1$ the L_1 norm.

Thirdly, we propose to use the cosine value of the angle between the two vectors pointing from the origin to the points as the similarity measure of the corresponding SAAPs. We have

$$\omega_{\text{Cosine}} = 1 - \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}.$$

Fourthly, we propose to adopt the abstract value of the Pearson's correlation coefficient value as the similarity measure. We have

$$\omega_{\text{Correlation}} = 1 - \frac{(\mathbf{x}-\bar{\mathbf{x}})\cdot(\mathbf{y}-\bar{\mathbf{y}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|_2 \|\mathbf{y}-\bar{\mathbf{y}}\|_2}.$$

Fifthly, we propose to adopt Canberra distance [16-18], which is a metric function often used for data scattered around an origin. The Canberra distance is similar to the Manhattan distance and the distinction is that the absolute difference between the variables of the two objects is divided by the sum of the absolute variable values prior to summing. We have

$$\omega_{\text{Canberra}} = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}.$$

Sixthly, we propose to use Chebyshev distance [19] as the similarity measure. Chebyshev distance is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. We have

$$\omega_{\text{Chebyshev}} = \max_i(|x_i - y_i|).$$

Seventhly, we propose to compute the Minkowski distance [20], a metric on Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. We have

$$\omega_{\text{Minkowski}} = \left[\sum_{i=1}^{n}(x_i - y_i)^m\right]^{\frac{1}{m}}.$$

In our research, we arbitrarily define the m=3.

Finally, we propose to compute the χ^2 distance [21], which is weighted Euclidean distance measure between profiles, where each squared difference between profile elements is divided by the corresponding element of the average profile. We have

$$\omega_{\chi^2} = \sum_{i=1}^{n}\frac{(x_i - y_i)^2}{|x_i + y_i|}.$$

D. Validation and Evaluation Methods

We adopt a large-scale leave-one-out cross-validation experiment to validate the performance of our approach in recovering known association between SAAPs and diseases. In each run of the validation, we select an association between a seed SAAP and a disease, assume that the association is unknown, and prioritize the SAAP against a set of control SAAPs. Performing such validation run for every seed SAAP and every disease, we obtain a number of ranking lists. With these lists, we calculate two criteria to measure the performance of the prioritization method. The first criterion is the mean rank ratio of seed SAAPs, which is the average rank ratio of all seed SAAPs for a specific disease. The

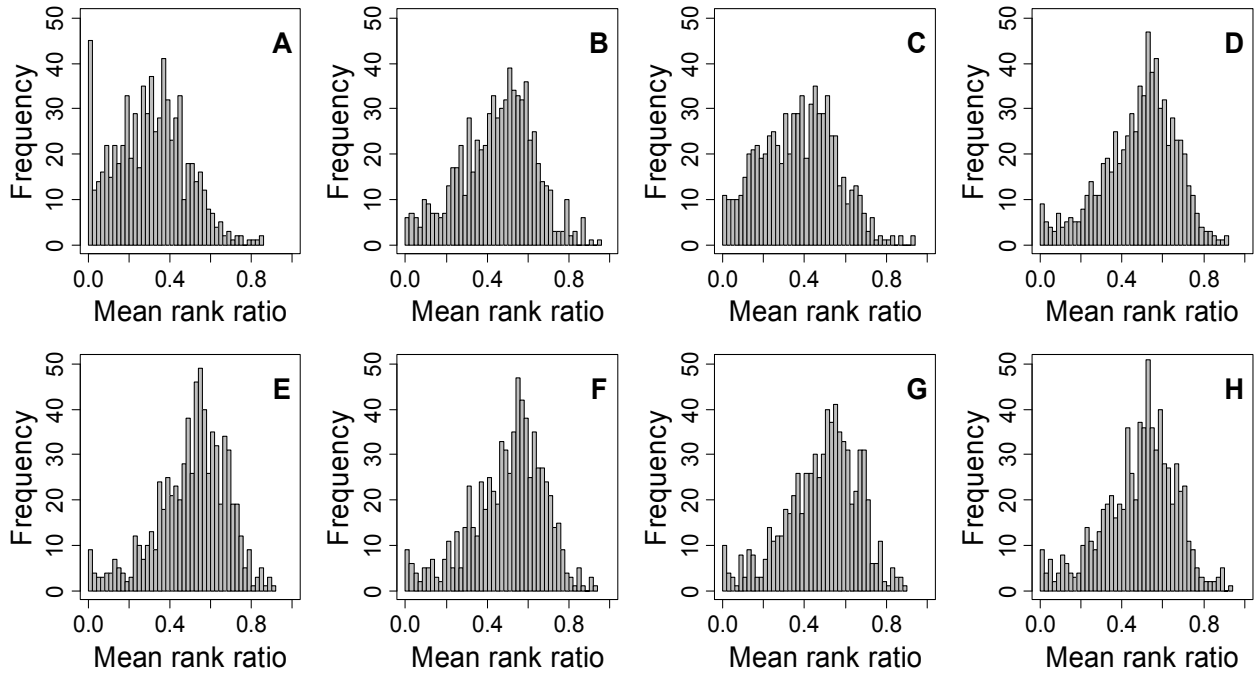

Figure 1. Distributions of mean rank ratios. A: using the Canberra distance measure. B: using the Chebyshev distance measure. C: using the Pearson's correlation coefficient measure. D: using the cosine measure. E: using the Euclidean distance measure. F: using the χ2 distance measure. G: using the Manhattan distance measure. H: using the Minkowski distance measure.

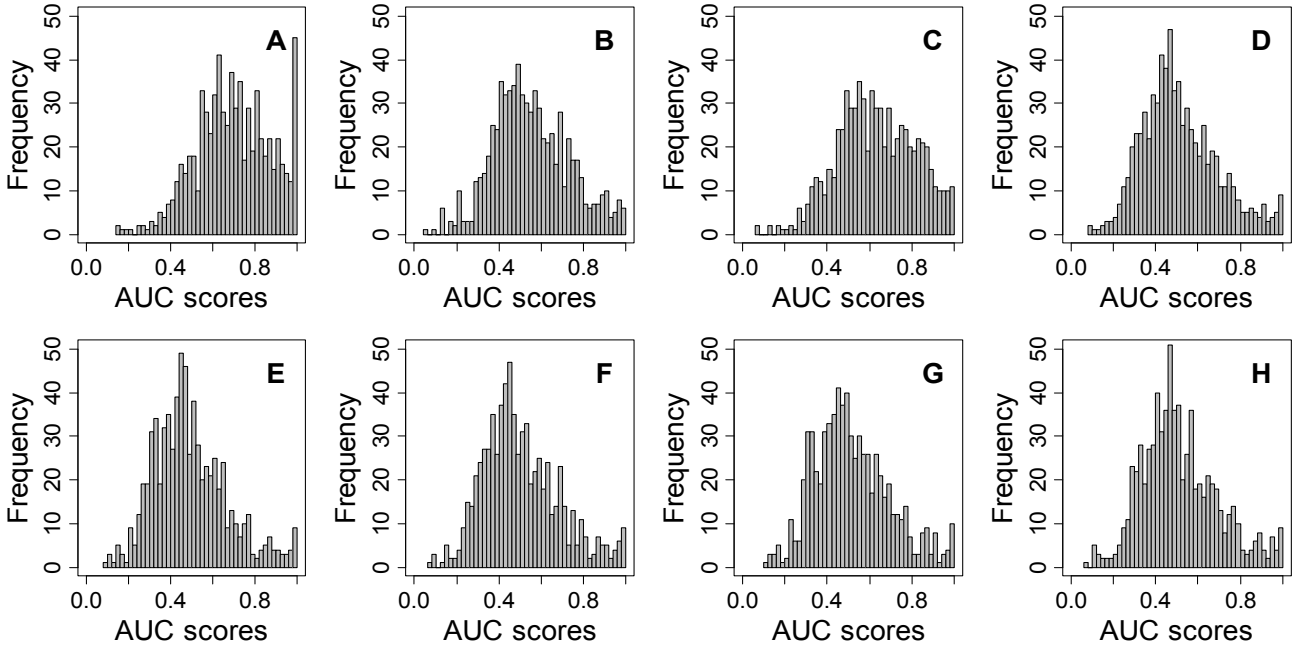

Figure 2. Distributions of AUC scores. A: using the Canberra distance measure. B: using the Chebyshev distance measure. C: using the Pearson's correlation coefficient measure. D: using the cosine measure. E: using the Euclidean distance measure. F: using the $\chi 2$ distance measure. G: using the Manhattan distance measure. H: using the Minkowski distance measure.

second criterion is the area under the receiver operating characteristic (ROC) curve (AUC). At a certain rank threshold, we define the sensitivity as the fraction of seed SAAPs ranked above the threshold, and specificity the fraction of control SAAPs ranked below the threshold. Varying the threshold, we are able to obtain a ROC curve. The area under this curve is then defined as the AUC score.

We choose all 13735 polymorphism SAAPs as the control group. As the seeds of the same disease should be more similar than the polymorphism SAAPs, it is expected that all the seeds should rank at the top, and thus we could expect low mean rank ratios and high AUC scores.

III. Results

A. Validation of the Model

We focus on diseases that have at least 4 seed SAAPs and obtain a total of 723 diseases. For each of these diseases, we perform a leave-one-out cross-validation experiment based on each of the eight similarity measures, and we present the resulting mean rank ratios in Fig. 1 and AUC scores in Fig. 2, from which we can see the effectiveness of the proposed method. In order to quantitatively demonstrate the performance of our methods under the eight similarity measures, we compute the percentage of the distribution of their mean rank ratios (shown in Table 1). For example, when using the Canberra distance measure, we can see that most of seeds can be ranked at top 50% among the control groups. In other words, we can recover the relationship between a large number of seeds and the corresponding diseases. Moreover, we calculate that for 86.31% (624) diseases, the mean rank ratios are less than 50%; for 70.68% (511) diseases, the mean rank ratios are less than 40%; for 48.13% (348) diseases, the mean rank ratios are less than 30%; for 30.29% (219) diseases, the mean rank ratios are less than 20%; for 15.08% (109) diseases, the mean rank ratios are less than 10%. We further run a Wilcoxon signed rank test against the alternative hypothesis that the median of the mean rank ratios is less than 50% (random situation), and we find that the p-value is less than 2.2×10^{-16}. In other words, it is statistically significant that our method can effectively

TABLE 1. Mean Rank Ratios of the Validation Experiment.

Cutoff	Canberra (%)	Chebyshev (%)	Correlation (%)	Cosine (%)	Euclidean (%)	Kai (%)	Manhattan (%)	Minkowski (%)
10%	15.08	4.56	7.19	3.87	3.18	3.60	3.60	3.60
20%	30.29	9.54	20.61	7.33	6.22	6.92	7.19	7.05
30%	48.13	20.61	35.68	14.94	12.03	13.55	14.94	14.94
40%	70.68	35.82	55.05	28.22	24.34	26.14	28.63	27.80
50%	86.31	56.85	74.83	46.75	42.32	44.12	47.58	46.75

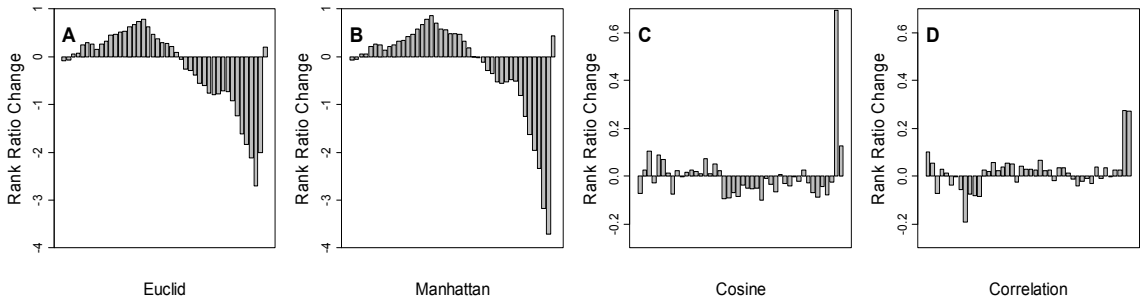

Figure 3. Relative importance of individual features. A: using the Euclidean distance measure. B: using the Manhattan distance measure. C: using the cosine measure. D: using the Pearson's correlation coefficient measure.

prioritize seed SAAPs among the top of candidate SAAPs.

B. Comparison of the Similarity Measures

The above results (Fig. 1, Fig. 2 and Table 1) of the leave-one-out cross-validation experiments also allow us to compare the performance of the eight similarity measures. From Table 1, we see that our model with the Canberra distance measure can give us the most accurate prediction result, the model with the Pearson correlation coefficient measure may also provides the distinguishing power for uncovering the relationships between the candidate SAAPs and query disease; while the model with the Chebyshev distance measure, the cosine measure, the Euclid distance measure, the χ^2 distance measure, the Manhattan distance measure, and the Minkowski distance measure seem not competent in this prioritization problem. To further elucidate this observation, we run seven Wilcoxon rank sum tests against the alternative hypothesis that mean rank ratios obtained using the Canberra distance measure have a negative location shift over those using the other measures. The results show that all the p-value are smaller than 2.2×10^{-16}. It is therefore clear that the Canberra distance measure is more suitable in measuring the similarity between two SAAPs.

C. Relative Importance of Individual Features

We try to understand the contributions of each individual feature in the model and find the most discriminative properties for the disease SAAPs. We adopt a permutation method to measure the importance of each feature [6]. Specifically, by shuffling the values of a feature in the samples, the information contained in the feature is broken. When the permuted feature is used with the remaining un-permuted features, the performance of the prioritization model may be impaired accordingly. Then, we compute the change in mean rank ratios for all SAAPs (the mean rank ratio calculated by pool all SAAPs) in validation before and after the permutation procedure. Consequently, we can use the change in mean rank ratios for all SAAPs to give a reasonable measure for the relative importance of a feature.

The results of the change in mean rank ratios for all SAAPs based on four most frequently used similarity measures (the Pearson correlation coefficient measure, the cosine measure, the Euclid distance measure, and the Manhattan distance measure) are presented in Fig. 3. Although the model with different similarity measure exhibit different contributions for the first 42 features, the model using the effective Pearson's correlation coefficient measures gives positive approval towards the discriminant power of the two

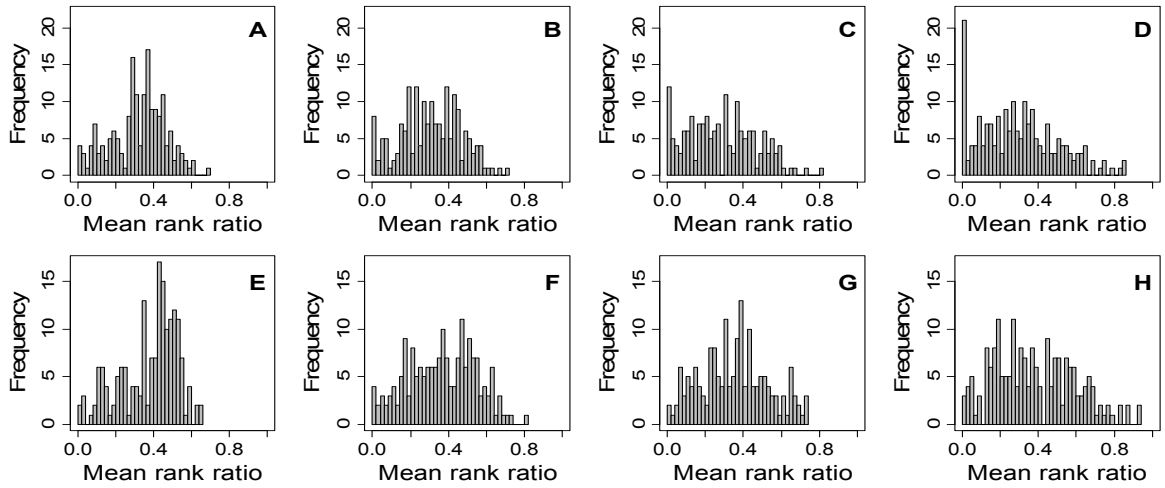

Figure 4. Seed effect. A-D: using the Canberra distance measure. E-H: using the Pearson's correlation coefficient measure.

conservation scores. We also run two Wilcoxon rank sum tests against the alternative hypothesis that mean rank ratios obtained using permuted feature 43 (or permuted feature 44) with other un-permuted features have a positive location shift over those using the original dataset, and we obtain small p-value (less than 2.2×10^{-16}) for both features. This result is consistent with the analysis of relative importance of the features in the literature [6], which points out the conservation scores have the most powerful discriminative ability to identify the disease-associated SAAPs against the neutral ones.

D. Effect of the Number of Seed SAAPs

We try to learn the influence of the number of seed SAAPs for our proposed model and see whether the model will give unstable results according to the change of the number of seeds. The amount of seed SAAPs extracted from Swiss-Prot database for each disease is quite different, ranging from 4 to 260. We thus divide our test dataset into four groups according to the amount of seed SAAPs: the first group has 177 diseases, the number of seed SAAPs is ranging from 20 to 260; the second group has 182 diseases, the number of seed SAAPs is ranging from 10 to 19; the third group has 168 diseases, the number of seed SAAPs is ranging from 6 to 9; the fourth group has 196 diseases, the number of seed SAAPs is ranging from 4 to 5. Here, we use the two powerful measures (the Canberra distance measure and the Pearson correlation coefficient) as examples. The histograms result of each group using the Canberra distance measure is shown in Fig. 4(A-D), and histograms result of each group using the Pearson correlation coefficient measure is shown in Fig. 4(E-H). Form the figure we can see that the number of seed SAAPs has little influence for the mean rank ratios. Therefore, the proposed method is stable to the number of seed SAAPs known to be associated with query diseases.

IV. Conclusions and Discussion

In this paper, we formulate the problem of identifying disease single amino acid polymorphisms against neutral ones for specific types of diseases as a one-class novelty learning problem. Comply with the guilt-by-association principle that a single amino acid polymorphism is considered as having association with a disease if the single amino acid polymorphism shares some common properties (such as physiochemical features, conservative level, and etc.) with a set of known seed SAAPs of the disease, we solve this problem using a guilt-by-association model. We implement our method using eight similarity measures with a set of physiochemical features and two conservation scores that are drawn only from protein sequence information. We demonstrate that the method is effective in ranking single amino acid polymorphisms that are responsible for specific diseases among the top of candidates. We also study the effects of different features and distance measures.

We can further carry out our study from the following aspects. First, we focus on the single amino acid polymorphisms occurring in known protein domains and collect the conserved protein domains for the query protein sequence based on the Pfam database. This limitation can be improved by using some other multiple-sequence alignment methods, such as PSI-BLAST [22], PANTHER [23] and so on. Second, we currently use a feature set including 42 physiochemical features and two conservation properties to construct our prediction model. From the analysis of feature importance, we can see that some features have negative contribution to our prediction result, and some features may be high correlated. In our future studies, we will combine our prioritization method with some feature selection mechanism to find out more effective features for our model. Finally, our approach is currently limited to single amino acid polymorphisms occurring in protein coding regions. However, mutations in other genome regions such as the transcriptional-factor binding sites and promoter regions may also be associated with human diseases. Further studies are needed for these mutations.

Acknowledgment

This work was partly supported by the Natural Science Foundation of China (60805010, 60928007, 60934004), Tsinghua University Initiative Scientific Research Program, Tsinghua National Laboratory for Information Science and Technology (TNLIST) Cross-discipline Foundation.

References

[1] R. Robinson, "Common disease, multiple rare (and distant) variants," *PLoS Biol,* 2010, 8: e1000293.
[2] B. Li, S. M. Leal , " Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data,"*Am J Hum Genet*, 2008, 83: 311-321.
[3] B.E. Madsen and S.R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS Genet*, 2009, 5(2):e1000384.
[4] D. J. Liu, S.M. Leal, "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions," *PLoS Genet*, 2010, 6(10):e1001156.
[5] S. Morgenthaler, W.G. Thilly, "A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST) ," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 2007, 615: 28–56.
[6] J. Wu, W. Zhang, R. Jiang, "Comparative study of ensemble learning approaches in the identification of disease mutations," *BMEI 2010*.
[7] V. Ramensky, P. Bork, S. Sunyaev , "Human non-synonymous SNPs: server and survey," *Nucleic Acids Res*, 2002, 30: 3894-3900.
[8] P.C. Ng, .S. Henikoff , "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res*, 2003, 31: 3812-3814.
[9] D.J. Liu, S.M. Leal, "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions," *PLoS Genet* 6, 2010, e1001156.
[10] D. Altshuler, M. Daly, L. Kruglyak, "Guilt by association," *Nat Genet*, 2000, 26: 135-137.
[11] T.U. Consortium , "The Universal Protein Resource (UniProt) in 2010," *Nucleic Acids Res*, 2010, 38: D142-148.
[12] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, et al. , "Pfam: clans, web tools and services," *Nucleic Acids Res*, 2006, 34: D247-251.
[13] R. Jiang, H. Yang, L. Zhou, C.C. Kuo, F. Sun, et al. , " Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations," *Am J Hum Genet*, 2007, 81: 346-360.
[14] R. Jiang, H. Yang, F. Sun, T. Chen, "Searching for interpretable rules for disease mutations: a simulated annealing bump hunting strategy," *BMC Bioinformatics*, 2006, 7: 417.
[15] P. Stenström, "High performance embedded architectures and compilers : third international conference," *HiPEAC 2008*, Göteborg, Sweden, January 27-29, 2008 : proceedings. Berlin ; New York: Springer. xiii, 400 p. p.
[16] S.M.Emran, N.Ye , " Robustness of Canberra metric in computer intrusion detection," *Proceedings of the 2001 IEEE*, Workshop on

Information Assurance and Security, United States Military Academy, West Point, New York, 5-6 June, 2001.
[17] G.N. Lance, W.T. Williams, " Computer programs for hierarchical polythetic classification ("similarity analysis") ," *Computer Journal*, 1966, 9:60-64.
[18] G.N. Lance, W.T. Williams , " Mixed-data classificatory programs in Agglomerative Systems," *Australian Computer Journal*, 1967, 1:15-20.
[19] J. M. Abello, P. M. Pardalos, and Mauricio G. C. Resende (editors) , " Handbook of Massive Data Sets," *Springer*, 2002, ISBN 1402004893.
[20] J.B. Kruskal, " Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis," *Psychometrika*, 1964, 29(1):1-27
[21] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," *Proc. iFlP 2nd Working Conf Visual Database systems*, 1992, pages 502-505.
[22] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman , " Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucl Acids Res*, 1997, 25:3389–3402.
[23] P.D. Thomas, A. Kejariwal, M.J .Campbell, H.Y. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, et al , " PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification," *Nucl Acids Res*, 2003, 31: 334–341.

Jiaxin Wu received her B.Sc. degree in Communication Engineering in 2005 from Beijing Jiaotong University, Beijing, China. She is now a M.S. candidate in the Department of Automation, Tsinghua University, Beijing, China. Her research interest includes pattern recognition, machine learning, data mining, and bioinformatics.

Mingxin Gan received her Ph.D. degree in Management Science and Engineering in 2006 from Beijing Institute of Technology, Beijing, China. She is now a lecture in the School of Economics and Management, University of Science and Technology Beijing, Beijing, China. Her research interest includes pattern recognition, machine learning, data mining, complex networks, natural language processing, and information retrieval.

Wangshu Zhang received her B.E. degree in Control Science and Engineering in 2008 from Harbin Engineering University, Harbin, China. She is now a M.S. candidate in the Department of Automation, Tsinghua University, Beijing, China. Her research interest includes pattern recognition, machine learning, data mining, and bioinformatics.

Rui Jiang received his Ph.D degree in Control Science and Engineering in 2002 from Tsinghua University, Beijing, China. He is now an associate professor in the Department of Automation, Tsinghua University, Beijing, China. His research interest includes bioinformatics, systems biology, pattern recognition, and machine learning.