I.J. Engineering and Manufacturing **2011, 5, 66-72**
Published Online October 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2011.05.09

Available online at http://www.mecs-press.net/ijem

Prioritization of Candidate Nonsynonymous Single Nucleotide Polymorphisms via Sequence Conservation Features

Jiaxin WU[a], Wangshu ZHANG[a], Rui JIANG[a,*]

[a] *MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China*

Abstract

The Detection of rare variants responsible for human complex diseases has been receiving more and more attentions. However, most existing computational methods for this purpose require the selection of functional variants before statistical analysis. Based on the assumption that nonsynonymous single nucleotide polymorphisms (nsSNPs) associated with specific diseases should be similar in their properties, we propose a method that utilize conservation scores of nsSNPs and the guilt-by-association principle to prioritize the candi-date nsSNPs for specific diseases. Systematic validation demonstrates that our approach is effective in recovering the relationship between nsSNPs and diseases, with the Manhattan distance measure achieving the most precise prediction results.

Index Terms: Prioritization, Nonsynonymous Single Nucleotide Polymorphisms (Nssnps), Guilt-By-Association, Euclidean Distance, Manhattan Distance

1. Introduction

Identification of genetic variants that are responsible for human inherited diseases has achieved remarkable success, represented by the fruits of genome-wide association (GWA) studies, in which hundreds of common variants in diverse complex traits have been reported [1]. One critical assumption held by the GWA studies is the common disease common variant (CDCV) hypothesis, which asserts that common diseases are caused by common variants with small to modest effects [2]. However, recent studies suggest an alternative hypothesis of common disease rare variant (CDVR), stating that the disease etiology is caused collectively by multiple rare variants with moderate to high penetrances, and the effective way to identify these rare variants is through direct sequencing [2,3]. Even so, due to the unaffordable cost of whole-genome sequencing, we should first quantify which variants are potentially functional or neutral, before statistical analysis of the sequence data. The results obtained from bioinformatics tools can be used to determine which variants should be included in the analysis. It has been pointed out that in an ideal situation, all variants that are included in the analysis are functional and no functional variants are excluded [2].

* Corresponding author:
E-mail address: ruijiang@tsinghua.edu.cn

Out of this consideration, we resort to existing functional databases about single nucleotide polymorphisms (SNPs), the most frequent type of human DNA variation [4], and we focus our study on protein-coding nonsynonymous single nucleotide polymorphisms (nsSNPs), whose presence result in amino acid substitutions, which potentially affect protein structures and functions, and further cause human disease [5]. A lot of previous studies have been conducted in identifying the disease nsSNPs against the neutral (non-disease) ones, which is often formulated as a binary classification problem, such as those of PolyPhen [4], SIFT [6] and KBAC [7]. With these studies, the yielding classification results contain only categories for each candidate nsSNP, either disease or neutral, without information about what specific diseases the nsSNP is related to. This raises questions about how to numerically evaluate the importance of the identified SNPs for arbitrary disease and select the top few susceptible ones, which would provide guidance for further prevention, diagnosis and treatment of the disease.

For these purposes, we formulate the identification of disease nsSNPs from candidates as a one-class novelty learning problem. We compute a score from multiple sequence alignment of proteins to quantify the strength of association between a query disease and a candidate nsSNP, and then we prioritize candidate nsSNPs according to their scores to facilitate the selection of susceptibility ones. The scoring approach complies with the guilty-by-association principle [8], on the basis of the assumption that nsSNPs associated with the same disease should have more similarities (such as structure, physicochemical characteristics, conservative level, and etc.) than those that are not associated with this disease. Grounded on the features derived from the conservation scores of nsSNPs, an aggregation similarity score is defined to measure the strength of associations between a certain nsSNP and a query disease in our model. We introduce two popular distance functions to calculate the aggregation similarity score (Euclidean distance and Manhattan distance), as well as four control groups to demonstrate the effectiveness and predictive power of our approach. Systematic validation demonstrates that our proposed approach is effective in identifying the casual relationship between nsSNPs and diseases, with the Manhattan distance achieving the most precise prediction results.

2. Materials and Methods

2.1. Data Sources

We collect from the Swiss-Prot database [9] nsSNPs and corresponding amino acid substitutions. Version 2010_10 (released on Oct. 5th, 2010) of this database collects 62,430 amino acid substitutions in 12,401 human proteins, with each substitution annotated as “Disease,” “Polymorphism,” or “Unclassified.” We refer to amino acid substitutions with the annotation “Disease” as disease nsSNPs and those with the annotation “Polymorphism” as neutral nsSNPs, and we focus only on the disease nsSNPs that have the corresponding OMIM accession numbers. We collect from the Pfam database [10] multiple sequence alignments (MSA) of human proteins. Version 24.0 (released in Oct. 2009) of this database contains curated alignments and models for 11,912 protein families. Focusing on nsSNPs that appearing in multiple sequence alignment of some human proteins, we finally collect 14,511 disease nsSNPs associated with 1,575 diseases and 13,735 neutral nsSNPs.

2.2. Sequence Conservation Features

We use the conservation scores of the original and the substituted amino acids as features to facilitate the prioritization of candidate nsSNPs, because previous studies have shown that these two scores have the most discriminant power in distinguishing disease nsSNPs from polymorphism ones [11,12]. The conservation scores are defined as the frequencies of occurrences of the amino acids (original or substituted) in the corresponding column of the Pfam multiple sequence alignment [13]. Specifically, for the query protein, its homologous proteins are extracted from the Pfam database. With the supposition that the substitution occurs at a position corresponding to the column of the alignment, the conservation scores are then calculated as the rela tive frequency of occurrence for the original (or the substituted) amino acid in the corresponding column of the alignment.

2.3. Guilt-by-association Model

We ground the prioritization of candidate nsSNPs on the guilt-by-association model, which is constructed with the assumption that nsSNPs that are associated with the same disease should share similar features and be relatively close under some distance function. With this assumption, a set nsSNPs that have associations with a specific disease are defined as seeds, and the total similarity scores of a candidate nsSNP to the set of seeds are utilized as the score for prioritization. The aggregate similarity score to measure the proximity between candidate nsSNP i and the disease D is then defined as $S_D(i) = \sum_{j \in V_D} z_{ij}$, where V_D is the set of all seed nsSNPs for disease D, and z_{ij} is the similarity score between nsSNPs i and j. Therefore, measuring the proximity between an nsSNP and a disease is translated to calculating the total similarities between the nsSNP and all the seeds pertained to the disease.

We adopt two distance functions to evaluate the similarity between two nsSNPs in the feature space, and thus obtain the similarity between two nsSNPs. The first function is the Euclidean distance, which is considered as the most traditional and ordinary way to compare two points in the feature space. The second function is the Manhattan distance, which is also known as the rectilinear distance, L_1 distance, city block distance, or taxicab distance. The Manhattan distance is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes [14]. Specifically, the Manhattan distance d between two n-dimensional feature vector x and y is $d(\mathbf{x},\mathbf{y}) = \|x - y\|_1 = \sum_{i=1}^{n} |x_i - y_i|$. According to literature [14], the advantage of the Manhattan over the Euclidean distance is that it weighs differences more heavily.

2.4. 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 nsSNP-disease associations. In each validation run, we select an association between a seed nsSNP and a disease, assume that the association is unknown, and prioritize the nsSNP against a set of control nsSNPs. Performing such validation run for every seed nsSNP 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 nsSNPs (MRR), which is the average rank ratio of all seed nsSNPs for a specific disease. The 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 nsSNPs ranked above the threshold, and specificity the fraction of control nsSNPs 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 four control groups, say, 99 randomly selected polymorphism nsSNPs, 999 randomly selected polymorphism nsSNPs, 9999 randomly selected polymorphism nsSNPs, and all 13735 polymorphism nsSNPs. As the seeds of the same disease should be more similar than the other polymorphism nsSNPs, it is reasonable that all the seeds should rank at the top, and thus we could expect low MRR and high AUC.

3. Results

3.1. Validation of the model

We focus on diseases that have at least 10 seed nsSNPs and obtain 723 diseases. For each of these diseases, we perform the leave-one-out cross-validation experiment, and we calculate MRRs and AUCs under the four control groups with the two distance measures. We summarize the resulting MRRs and AUCs in Fig.1 and 2, respectively. In each figure, we further present three situations: (1) using the conservation score of the original amino acid (feature 1), (2) using the conservation score of the substituted amino acid (feature 2), and (3) using a vector composed of the two conservation scores (feature 1&2).

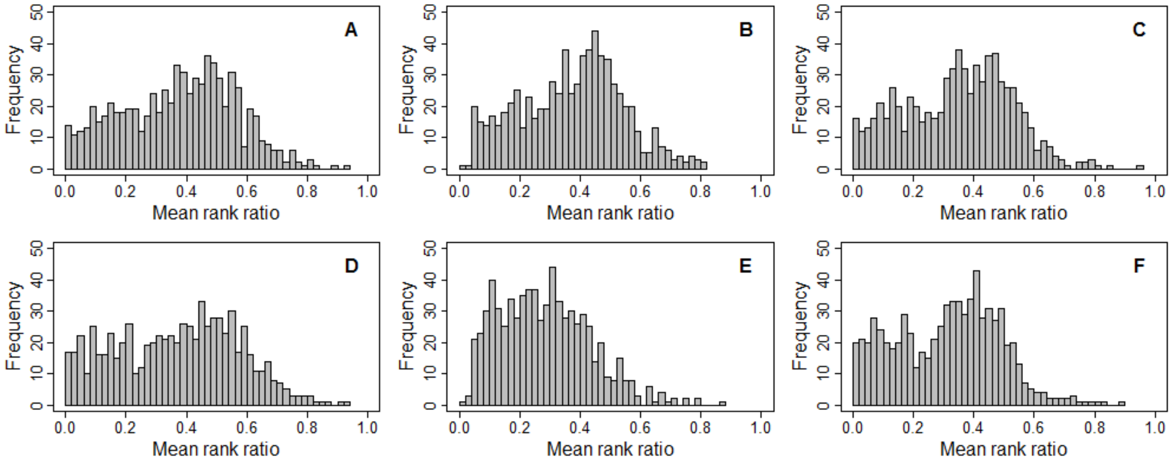

Fig. 1. Distribution of mean rank ratios of all 723 diseases, against all 13735 polymorphism nsSNPs. A-C: Euclidean distance. D-F: Manhattan distance. A, D: feature 1. B, E: feature 2. C, F: feature 1&2.

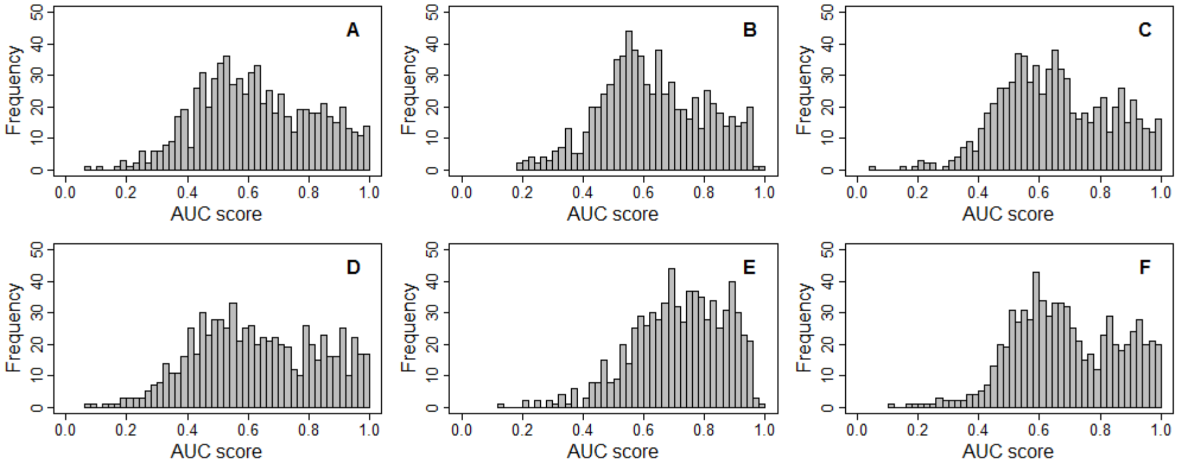

Fig. 2. Distribution of AUC scores of all 723 diseases, against all 13735 polymorphism nsSNPs. A-C: Euclidean distance. D-F: Manhattan distance. A, D: feature 1. B, E: feature 2. C, F: feature 1&2.

The results show that nearly all 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. In addition, the four control groups provide similar prediction performances under a certain distance functions (data not shown). Taking Fig. 1 (B) as an example, we calculated that for 91.56% (662) diseases, the MRRs are less than 50%; for 32.64% (236) diseases, the MRRs are less than 20%; for 10.79% (78) diseases, the MRRs are less than 10%. We further run a Wilcoxon signed rank test against the alternative hypothesis that the median of the MRRs is less than 50% (random situation), and we find that no matter which features are used, the p-value is less than 2.2e-16. In other words, it is statistically significant that our method can effectively prioritize seed nsSNPs among the top of candidate nsSNPs.

3.2. Comparison between the Similarity Measures

We also observe that MRRs tend to be smaller in the leave-one-out cross-validation when using the Manhattan distance. To further elucidate this observation, we run a Wilcoxon rank sum test against the alternative hypothesis that MRRs obtained using the Euclidean distance have a positive location shift over those using the Manhattan distance. The results show that the p-value is 0.3629 for feature 1, less than 2.2E-16 for feature 2, and 3.51e-05 for feature 1&2. It is therefore clearly to see that the Manhattan distance measure is more suitable in measuring the similarity between two nsSNPs.

Table 1. Prediction performances for Disease (MIM:143890)

Condition			**99 Polymorphism nsSNPs**	**999 Polymorphism nsSNPs**	**9999 Polymorphism nsSNPs**	**All Polymorphism nsSNPs**
Mean Rank Ratio	Euclidean	Feature 1	0.1951 ±0.0272	0.1915 ±0.0086	0.1930 ±0.0010	0.1926
		Feature 2	0.4585 ±0.0471	0.4484 ±0.0113	0.4482 ±0.0003	0.4481
		Feature 1&2	0.1752 ±0.0254	0.1899 ±0.0086	0.18752 ±0.0009	0.1871
	Manhattan	Feature 1	0.1388 ±0.0176	0.1485 ±0.0068	0.1479 ±0.0009	0.1470
		Feature 2	0.2042 ±0.0201	0.1937 ±0.0077	0.1946 ±0.0010	0.1955
		Feature 1&2	0.1350 ±0.0129	0.1406 ±0.0046	0.1419 ±0.0006	0.1416
AUC score	Euclidean	Feature 1	0.8104 ±0.0277	0.8090 ±0.0086	0.8070 ±0.0010	0.8073
		Feature 2	0.5409 ±0.0480	0.5515 ±0.0113	0.5517 ±0.0003	0.5519
		Feature 1&2	0.8311 ±0.0258	0.8107 ±0.0086	0.8125 ±0.0009	0.8129
	Manhattan	Feature 1	0.8678 ±0.0174	0.8521 ±0.0068	0.8521 ±0.0009	0.8522
		Feature 2	0.8003 ±0.0203	0.8067 ±0.0078	0.8054 ±0.0010	0.8045
		Feature 1&2	0.8715 ±0.0127	0.8601±0.0045	0.8581 ±0.0006	0.8585

3.3. Comparison between Features

From Fig. 1, we roughly see that MRRs tend to be smaller in the leave-one-out cross-validation when feature 2 is used. To further elucidate this observation, we run a Wilcoxon rank sum test against the alternative hypothesis that MRRs obtained using feature 1 have a positive location shift over those using feature 2, and we obtain a small p-value (2.2E-16). That is to see, feature 2 has higher discriminant power than feature 1 in this prioritization problem. Similarly, we conclude that feature 2 has higher discriminant power than 1&2 (p-value = 0.0001), and feature 1&2 has higher discriminant power than 1 (p-value = 4.541E-9). This result is consistent with the analysis of relative importance of the features in literature [12], which points out the conservation score for the substituted amino acid has the most powerful discriminative ability to identify the disease-associated nsSNPs against the neutral ones.

3.4. Case studies

By prioritizing candidate nsSNPs, we aim at finding nsSNPs that are most relevant to the disease of interest, thereby promoting the detection of potential functional rare variants in successive association studies. Taking Familial hypercholesterolemia (FH) [MIM:143890] as an example, we apply the proposed method with the use of feature 1&2 and the Manhattan distance measure, and we obtain overall MRR=14.16% and AUC = 85.85%. From the literature, we know that FH results from defective low-density lipoprotein receptor (LDLR) activity, mainly due to LDLR gene defects [15,16,17]. According to the ranking results, we can thus get the top five significant disease-related nsSNPs, which are D579Y, P608S, D221Y, D224V, and D221G in the gene LDLR and their relative ranks are all less than 1.00% (rank top 140 among 13736 nsSNPs).

We also study some common complex diseases, such as Breast cancer (BC) [MIM: 114480]. It was found in the middle of 1990s that mutated variants in BRCA1 or BRCA2 gene significantly raised a person's odds of developing breast cancer [18]. In our study, the top 5 variants selected statistically significant association with breast cancer from our prediction results are T826K in BRCA1, T2515I in BRCA2, S2072C in BRCA2, H888Y in BRCA1, and G960D in BRCA1, and their relative ranks are all <6.74% (rank top 1000 among 13736 nsSNPs).

4. Conclusion

In this paper, we model the problem of identifying nsSNPs underlying diseases against neutral ones for specific types of diseases as a one-class novelty learning problem, and we solve this problem from the viewpoint of guilt-by-association principle. We implement our method using two distance measures with four control groups on the basis of two features. We demonstrate that the method is effective in ranking nsSNPs that are responsible for specific diseases among the top of candidates. We also analyze the effects of different features and distance measures.

Certainly, there are several limitations of the proposed approach. First, we use the Pfam multiple sequence alignment to extract conserved protein domains for the query protein sequence. As a result, we are limited to the mutations occurring in known protein domains. This limitation can be overcome by using some other multiple-sequence alignment methods, such as BLAST. Second, we currently use only the two conservation scores to construct our prediction model. In our future studies, we will combine some useful information such as the physicochemical characteristics of amino acids, or the structure information of proteins to form a more comprehensive feature set. Finally, our approach is limited to nsSNPs found in protein coding regions. However, mutations in other genome regions such as the transcriptional-factor binding sites and promoter regions are also known to cause diseases. Further studies are needed for these mutations.

Acknowledgements

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] Robinson R (2010) Common disease, multiple rare (and distant) variants. PLoS Biol 8: e1000293.

[2] Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.

[3] Xiong M, Zhao J, Boerwinkle E (2002) Generalized T2 test for genome association studies. Am J Hum Genet 70: 1257-1268.

[4] Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-3900.

[5] Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037-2048.

[6] Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.

[7] Liu DJ, Leal SM (2010) 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: e1001156.

[8] Altshuler D, Daly M, Kruglyak L (2000) Guilt by association. Nat Genet 26: 135-137.

[9] Consortium TU (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142-148.

[10] Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247-251.

[11] Jiang R, Yang H, Zhou L, Kuo CC, Sun F, et al. (2007) Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. Am J Hum Genet 81: 346-360.

[12] Wu J, Zhang W, Jiang R (2010) Comparative study of ensemble learning approaches in the identification of disease mutations. BMEI 2010.

[13] Jiang R, Yang H, Sun F, Chen T (2006) Searching for interpretable rules for disease mutations: a simulated annealing bump hunting strategy. BMC Bioinformatics 7: 417.

[14] Stenström P (2008) 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.

[15] Bourbon M, Duarte MA, Alves AC, Medeiros AM, Marques L, et al. (2009) Genetic diagnosis of familial hypercholesterolaemia: the importance of functional analysis of potential splice-site mutations. J Med Genet 46: 352-357.

[16] Taylor A, Tabrah S, Wang D, Sozen M, Duxbury N, et al. (2007) Multiplex ARMS analysis to detect 13 common mutations in familial hypercholesterolaemia. Clin Genet 71: 561-568.

[17] Humphries SE, Neely RD, Whittall RA, Troutt JS, Konrad RJ, et al. (2009) Healthy individuals carrying the PCSK9 p.R46L variant and familial hypercholesterolemia patients carrying PCSK9 p.D374Y exhibit lower plasma concentrations of PCSK9. Clin Chem 55: 2153-2161.

[18] DNA Mutation Diseases. http://wwwexplorednacouk/dna-mutation-diseaseshtml.