ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

A Random Forest proximity matrix as a new measure for gene annotation*

Jose A. Seoane[1], Ian N.M. Day[1], Juan P. Casas[2], Colin Campbell[3] and Tom R. Gaunt[1,4]

1 Bristol Genetic Epidemiology Labs. School of Social and Community Medicine
Oakfield House, Oakfield Grove. University of Bristol. BS8 2BN. Bristol, UK
2 Department of Non-communicable Disease Epidemiology
London School of Hygiene and Tropical Medicine, UK
3 Intelligent Systems Laboratory.
Merchant Venturers Building. University of Bristol. BS8 2BN. Bristol, UK
4 MRC Integrative Epidemiology Unit. School of Social and Community Medicine
Oakfield House, Oakfield Grove. University of Bristol. BS8 2BN. Bristol, UK

Abstract. In this paper we present a new score for gene annotation. This new score is based on the proximity matrix obtained from a trained Random Forest (RF) model. As an example application, we built this model using the association p-values of genotype with blood phenotype as input and the association of genotype data with coronary heart disease as output. This new score has been validated by comparing the Gene Ontology (GO) annotation using this score versus the score obtained from the gene annotation "String" tool. Using the new proximity based measure results in more accurate annotation, especially in the GO categories Molecular Function and Biological Process.

1 Introduction

In large population epidemiological studies, access to the original data is often not straightforward, because of privacy issues related with genetic data [1]. However, summary statistics from the study could be made publicly available and used by third parties in further analyses, without concern about privacy issues. This summary data could include individual association coefficients (betas of the regression) and probability values (p-values) between the genetic variables and phenotypic traits. By comparing the association patterns with multiple traits for one genetic variant with another we could potentially obtain novel information about the functional similarity of those two variants. If two variants present a similar association pattern with a set of phenotypes, it is likely that both genes are functionally related in some way; conversely, if their association patterns are very different they are likely to be functionally unrelated.

There has been considerable effort in the last decade to develop computational methods for gene functional annotation [2]. Using a variety of types of data (amino acid sequences, evolutionary relationships, protein-protein interaction networks, expression or combination of them) many approaches have been developed, using

* This work was funded by the UK Medical Research Council (grant G1000427). The British Women's Heart and Health Study (BWHHS) has been supported by funding from the British Heart Foundation (BHF, PG/07/131/24254) and the UK Department of Health Policy Research Programme. The BWHHS Illumina HumanCVD BeadChip work is funded by the BHF (PG/07/131/24254, PI Tom Gaunt).

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

different methodologies. Some of these methodologies involve machine learning approaches, such as kernel methods, graph-based methods, Markov random field, etc. Some of the state-of-the-art methodologies have been compared in these papers [3] [4].

In this work we propose a new gene-gene interaction measure based on Random Forest to discover new Gene Ontology annotations using patterns of genotype-phenotype association p-values (multi-phenotype association profiles, MPAP).

Because of their high performance in high dimensional analysis, Random Forests [5] have been used in recent years in a number of projects related to genetic analysis [6]. One of the important features of Random Forests is the potential to obtain a set of measures related with the model in addition to the classification model, such as the proximity matrix, the feature importance values or the local importance matrix. Once the Random Forest has been trained, the proximity matrix shows this similarity between the samples in the Out of Bag (OOB) set (internal validation set of the RF algorithm, used to get the performance measures, and also the proximity matrix). The proximity between two samples is calculated by measuring the number of times that these two examples end in the same terminal node of the same tree of the RF, divided by the number of trees in the forest.

2 Methods

2.1 Data

Original association data was obtained by performing a linear regression association analysis of genotype data over a set of 64 blood samples phenotypes using PLINK [7]. Data are from The British Women's Heart and Health Study (BWHHS), a UK-based prospective cohort study of 4286 healthy women aged 60-79 years at baseline (1999-2001) [8]. Genotyping was performed using the Illumina HumanCVD BeadArray (Illumina Inc, San Diego, USA), which comprises nearly 50,000 Single Nucleotide Polymorphisms (SNPs) in over 2,000 genes selected on the basis of cardiovascular candidacy by an international consortium of experts [9]. The different phenotypes used in this study consisted of 64 cardiovascular-disease related blood measures and an indicator of whether a patient has suffered coronary heart disease. All phenotypic features were normalized to zero mean and unit variance.

2.2 Study design

We used a Random Forest (RF) implementation of R (package randomForest) as a regression model using as outcome the association odds ratio between each SNP and the presence or absence of coronary heart diseases CHD, and using as input to the RF the phenotype p-values of the association between each SNP and phenotype as features.

We used cross validation to select the main parameter of the RF, defined by Breiman as the number of features that each new tree selects randomly for classification. The best results were obtained using all the features (64).
We also ran several executions to select the optimum number of trees comprising the RF. We observed graphically that results did not improve significantly when the

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

number of trees was bigger than ~2000 trees, so we selected a conservative number of 2500 trees.

Once the model was trained, we obtained not only the regression model, but also the local importance matrix, which reflects the "importance" of each phenotype to each SNP for the prediction and the proximity matrix, which reflects the similarity between the SNPs.

We decided to use this similarity concept to functionally annotate genes involved in this study. Our hypothesis was that if two SNPs have similar behaviour (similar terminal nodes in random forest training) in a prediction model which relates association values between SNPs/phenotypes and coronary heart diseases, the genes related with these SNPs are likely to have similar functions associated with CHD.

We trained several random forest models to ensure that the proximity matrix is equivalent. We selected the median of 5 proximity matrices. Most of the values in the proximity matrix were zero, or very close to zero. Based on the histogram of these values we selected a threshold of importance to include SNPs in the study of 0.01. We applied a log transformation to these proximity values in order to homogenize the distribution and scale it to (0-1). Each gene is represented by different SNPs, and functions are associated with an individual gene. We assign the proximity value to a gene as the maximum value of all the SNPs associated with this gene.

2.3 **Gene annotation**

In order to assign a function to a gene, formally represented by a Gene Ontology annotation, we selected an approach called guilt-by-association [10]. This approach states that genes that are associated or interact share some function. This concept has been used extensively in computational biology to annotate a gene based on the functions of the genes associated with it, following different types of association, such as co-expression, appearance in the same pathway database, etc. We must note that the guilt-by-association approach has some drawbacks, discussed by [11], such us the importance of critical nodes in the network.

In the first step, for each gene of interest (GoI) in our study, we created a network of genes related with this GoI, using the web tool String [12]. This tool allowed us to retrieve a set of genes that are related with a particular gene, with a relatedness score. This score is calculated based on several concepts, such as co-expression, interaction, occurrence in databases or text mining. This gene network was then used as the baseline of our annotation algorithm. For each gene of interest we selected the 20 most important genes with scores bigger than 100. Later we discarded genes in this set that were not in our study or with a low score in the proximity matrix.

In the second step, we selected all possible Gene Ontology annotations (including ancestors) of genes related with GoI. All these annotations were potential new annotations for our gene of interest.

In order to select which functions are related with the GoI and which not, we used an annotation algorithm proposed in [13]. For each potential GO tag, we annotated the genes that have this tag with a one, the genes without this tag with a minus one and the genes with unknown function (GoI) with a zero. We built a graph where nodes represent genes and the edges represent the relationship between them.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The edges have weights that represent the value of the proximity score between each pair of genes. We then applied the max flow / min cut algorithm [14]. This algorithm minimizes the inconsistent assignments for each GO annotation with a low computational cost using graph theory. The results were a set of candidate GO annotation related with GoI.

2.4 **Validation**

In order to validate the use of a proximity matrix as a new score value for gene annotation, we used a "leave-one-out" approach. We removed all GO annotation for each gene of interest in turn and then calculated the potential annotations for that particular gene using the GO terms of the other related genes. We then compared the obtained gene annotations with the true gene annotation using precision (intersection between number of real GO annotations and retrieved annotations divided by total retrieved annotations), recall (intersection between number of real GO annotations and retrieved annotation divided by total real annotation) and F-score (harmonic mean between precision and recall, 2 x (precision x recall) / (precision + recall)).

We were unable to compare our results directly with the performance of other gene annotation methods, because these results are based on a single dataset.
However, we could compare our proximity score with a well known score, such as the String database score (obtained from several concepts, such as co-expression, interaction, occurrence in databases or text mining). We applied the same methodology described above, using the String score [12] as weights between the nodes instead of proximity score, and compare the precision, recall and F-score between the two scores.

3 Results

As mentioned in the previous section, 5 RF were trained using p-values as input data. The mean of 5 models Out of the Bag (OOB) regression "pseudo R-square" error was 0.0827 (0.0003) and the mean square error was 0.9172 (0.0003). In order to compare the performance of the Gene Annotation, we obtained the precision, recall and F-score value for each gene in the study. In the top panel of figure 1 we use a boxplot to compare the F-score measure of the proximity score approach versus the F-score of the String annotation for each Gene Ontology category: molecular function (MF), biological process (BP) and cellular component (CC). In the bottom panel of figure 1 we show, with more detail, the precision and recall for each GO category. Note that each value corresponds to a different gene. The gray point represents the mean value and black line represents median. In table 1 we show the percentage of genes where precision or recall using the proximity score improves the String database baseline.

	BP	MF	CC
Precision	64.34%	59.61%	61.98%
Recall	45.13%	50.28%	2.29%

Table 1: Percentage of improvement.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

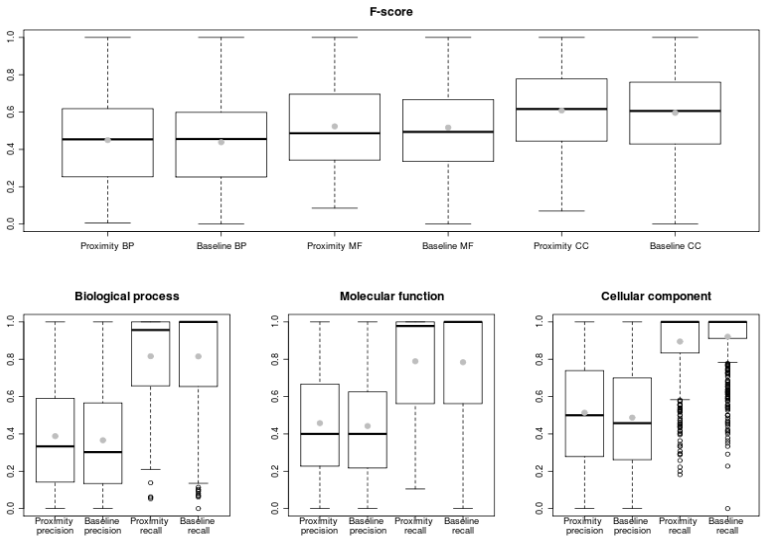

Fig. 1: Results comparison boxplot: Top panel shows the comparison of F-score values between proximity and baseline scores in BP, MF and CC. Bottom panels show detail of BP, MF and CC proximity and recall comparisons. Gray point shows the mean of values and black line represents median

4 Discussion

In a machine learning context, the obtained RF performance measures are not particularly good (r-square 0.08). However, in the genetic association context, these results could be significant. Also, the objective of this work was not the prediction of association between a gene and cardiovascular disease, but to obtain a new useful score to group genes related with cardiovascular disease into sets of related genes, and use this score to discover and annotate gene functions.

Even with the low prediction performance of Random Forest, the score derived from the proximity matrix improves the gene annotation F-score in each of the three GO categories: cellular component (CC), molecular function (MF), and biological process (BP), as shown in the top panel of figure 1.

More detailed analysis of precision and recall values (bottom panel in figure 1), shows that the proximity score offers the greatest benefit to precision values. In BP terms, both the mean and median of precision in the proximity score improves over the String score approach. However, in recall, mean remains similar but median is lower (median proximity recall 0.95 vs. median String recall 1)

In MF GO annotation, the median of precision is similar, but the mean of the proximity score improves the mean of the String score. The mean of recall values is similar and the median of recall in the proximity is lower than the median of recall in the String score.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

In CC, both median and mean of precision with the proximity score improve the String score and the median of recall in both is the same (1). However, the mean of the proximity scores is much lower (0.89 vs 0.92) than the String score. This difference is reflected in the percentage of improvement, where 61% of terms improve the String score, and only 21% of them improve the String score in recall, which could indicate some biased result in CC, such as a lower number of selected annotations. These results indicate that the improvement of the proximity matrix appears mainly in MF and BP, which are more related with functional variations that potentially lead to a disease than CC.

In conclusion, this work proposes a new measure to infer gene-gene relationship, based on the proximity matrix of a Random Forest. Our measure obtains some improvements especially in terms of precision on molecular function and biological process, and similar results in recall. Even when the results show only a small improvement, they illustrate that our proposed score based in genotype-phenotype association provides some new information, discovering new true GO terms, so it could improve the gene annotation process, by including it as a new measure in gene annotation process, in combination with other scores obtained by experimentation (similar expression, same pathway membership, etc.). The main advantage of this method is that raw genetic data is not needed, only association p-values between SNPs and phenotypes. This is important because p-values and betas can potentially be obtained from very large consortia without any privacy issues.

References

[1] Homer, N., Szelinger, S., Redman, M., Duggan, D., et al.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4, e1000167 (2008)

[2] Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Mol Syst Biol 3, 88 (2007)

[3] Pena-Castillo, L., Tasan, M., Myers, C.L., Lee, H., et al.: A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol 9 Suppl 1, S2 (2008)

[4] Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., et al.: A large-scale evaluation of computational protein function prediction. Nat Methods 10, 221-227 (2013)

[5] Breiman, L.: Random forests. Machine learning 45, 5-32 (2001)

[6] Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., et al.: Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform 14, 315-326 (2013)

[7] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575 (2007)

[8] Lawlor, D.A., Bedford, C., Taylor, M., Ebrahim, S.: Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. J Epidemiol Community Health 57, 134-140 (2003)

[9] Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., et al.: Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS One 3, e3583 (2008)

[10] Oliver, S.: Guilt-by-association goes global. Nature 403, 601-603 (2000)

[11] Gillis, J., Pavlidis, P.: "Guilt by association" is the exception rather than the rule in gene networks. PLoS Comput Biol 8, e1002444 (2012)

[12] Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., et al.: STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37, D412-416 (2009)

[13] Murali, T.M., Wu, C.J., Kasif, S.: The art of gene function prediction. Nat Biotechnol 24, 1474-1475; author reply 1475-1476 (2006)

[14] Goldberg, A.V., Tarjan, R.E.: A new approach to the maximum-flow problem. Journal of the ACM (JACM) 35, 921-940 (1988)