Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-91

Predicting the Pathogenic Impact of Sequence Variation in the Human Genome

Mark F. ROGERS[a,1], Hashem A. SHIHAB[b], Michael FERLAINO[c,d], Tom R. GAUNT[b] and Colin CAMPBELL[a]
[a] *Intelligent Systems Laboratory, University of Bristol, Bristol, BS8 1UB, U.K.*
[b] *MRC Integrative Epidemiology Unit (IEU), University of Bristol, Bristol, BS8 2BN,U.K.*
[c] *Nuffield Department of Obstetrics and Gynaecology, John Radcliffe Hospital,University of Oxford, Oxford, OX3 9DU, U.K.*
[d] *Big Data Institute, University of Oxford, Oxford, U.K.*

Abstract. Sequencing data will become widely available in clinical practice within the near future. Uptake of sequence data is currently being stimulated within the UK through the government-funded 100,000 genomes project (Genomics England), with many similar initiatives being planned and supported internationally. The analysis of the large volumes of data derived from sequencing programmes poses a major challenge for data analysis. In this paper we outline progress we have made in the development of predictors for estimating the pathogenic impact of single nucleotide variants, indels and haploinsufficiency in the human genome. The accuracy of these methods is enhanced through the development of disease-specific predictors, trained on appropriate data, and used within a specific disease context. We outline current research on the development of disease-specific predictors, specifically in the context of cancer research.

Keywords. Prediction, sequence data, variant, annotation, point mutation, indel.

1. Introduction

Substantial improvements in sequencing technologies, and rapidly falling costs, will result in the widespread use of DNA sequence data within clinical practice. This trend is being encouraged within the UK through the Genomics England (100,000 genomes) project. Interpretation of these datasets poses challenges, from the size and complexity of the data through to the necessary linkage of DNA sequence data with other types of data, such as clinical covariates. For the analysis of DNA sequence data, a crucial challenge is the ability to distinguish which genetic variants are functional in disease, against a background of many disease-neutral variants. Accurate understanding of which genetic variants are pathogenic will improve our understanding of the molecular mechanisms underlying human disease and our ability to provide targeted therapies.

In recent research we have developed a variety of methods for predicting the pathogenic impact of genetic variants. In Shihab et al (2015) [1] we proposed an integrative classifier for predicting whether single nucleotide variants (SNVs) are

[1] Corresponding author, Intelligent Systems Laboratory, University of Bristol, Bristol, BS1 1UB, U.K.; E-mail: mark.rogers@bristol.ac.uk

functional in human disease, or neutral (for both coding or non-coding regions of the human genome). A number of sources of data are relevant to predicting if a SNV is functional in disease or is neutral. Consequently we used a variety of feature groups, or sources of data, which could be informative. In the construction of these prediction methods we used sequence conservation across species, histone modification (ChIP-Seq data), transcription factor binding site data, open chromatin data (DNase-Seq peak calls across cell lines from ENCODE), GC content, genome segmentation, and annotations describing DNA footprints across cell types (from ENCODE [3]). Thus, for example, sequence conservation across species proved to be a highly informative source: if a SNV occurs in a genomic region which is highly conserved across species it is much more likely to be functional in disease relative to a SNV which occurs in a region with high variability across species..

In our study, Shihab et al (2015) [1], we therefore used an algorithm-based approach capable of data integration i.e. the algorithm uses and learns to weight these different types of data, according to relative informativeness. In this study we used a specific approach to data integration called multiple kernel learning [2], though other data integration methods can be used. The method was called FATHMM-MKL (see fathmm.biocompute.org.uk for the prediction tool). Aside from giving a predicted label (pathogenic or neutral), the method also assigns a confidence measure to this label. At the default threshold on this confidence measure, FATHMM-MKL has a balanced test accuracy of 89.7%, with a false-positive rate of 3.8%. With a higher cutoff threshold on the confidence, the test accuracy slightly drops to 88.0% but with the false-positive rate dropping to 1.2%. A number of other groups have also proposed predictors for estimating the pathogenic impact of SNVs [4,5,6,7,8].

We have extended this line of investigation in a variety of directions. Small insertions and deletions (indels) can also have a significant influence in human genetic disease. In terms of relative frequency, indels are second only to SNVs as mutations. To date, classifiers for predicting the functional impact of indels have been restricted to their effect in the human exome (e.g. [9,10,11,12]). However, non-coding regions also contain many functional elements. Indeed, the vast majority of catalogued SNV-trait associations fall within non-coding regions of the human genome [13]. We have proposed an integrative predictor for estimating the pathogenic impact of indels in non-coding regions of the human genome [14]: the method is called FATHMM-indel and is available via the Web (indels.biocompute.org.uk). Using nested cross validation, this classifier achieves a balanced accuracy of 86%. In other work [15] we have proposed a Genome Tolerance Browser to visualise the possible pathogenic impact of SNVs in the genome (this tool is available at gtb.biocompute.org.uk). A further project has been to develop a state-of-the-art predictor, called HiPred, for estimating the effect of haploinsufficiency [16]. Cells in the human body are diploid, they contain two complete sets of chromosomes, one from each parent. Haploinsufficiency occurs if there is only one functional copy of a gene, and this single copy does not produce a sufficient amount of a gene product, resulting in a disease trait.

2. Disease-specific prediction

The predictors for SNVs and indels have a high accuracy in many simple disease contexts but are still not sufficiently accurate when applied to more complex multifactorial diseases. For a complex disease, such as cancer, oncogenesis is typically driven by a

combination of disease-enabling genetic variants. For construction of a prediction tool, this creates a label-dependency problem during classifier training: a single point mutation could be labelled pathogenic or neutral, depending on the labels at other locations in the cancer genome. In any case, training a classifier with domain-specific data, such as sequence data exclusively from a particular type of cancer, would likely offer improved accuracy. Indeed, our previous studies have suggested that disease-specific predictors are more accurate than generic predictors [18].

With this motivation we are devising cancer-specific predictors, for predicting the oncogenic impact of single point mutations. As for FATHMM-MKL, these predictors are trained using a variety of data sources falling into three main categories: genomic (genomic features include GC content, sequence spectra, repeat regions and measures of region uniqueness), evolutionary (as for FATHMM-MKL, evolutionary features include a comprehensive set of conservation-based measures) and consequences (for coding regions only: data derived from the Variant Effect Predictor [19] and other sources). To train the classifier for handling single point somatic mutations, we used high recurrence rate SNVs from the COSMIC cancer database [20] as the positives, with negatives derived from the 1000 Genomes project [21]. We achieved state-of-the-art performance with a substantial gain over competitors (Figure 1, left). This predictor is called CScape and is available via the Web (cscape.biocompute.org.uk) [22]. Evaluated via leave-one-chromosome-out cross-validation (LOCO-CV), the approximately balanced test accuracy is 72.3% in coding regions and 62.9% in non-coding regions. As with FATHMM-MKL we also devised a confidence measure associated with the predicted class label.

Though promising, the test accuracy of the resultant classifiers remains inadequate for use by cancer researchers. However, if we restrict prediction to the highest confidence instances then it is possible to achieve 91.7% test accuracy (with LOCO-CV, coding regions only). Given a positive predictive value (PPV) of 0.78, and a large number of true positives, this test accuracy is not achieved by predominant accurate prediction of negatives (non-oncogenic single point mutations). This strong performance comes at the expense of yielding predictions for just 17.7% of coding region nucleotide positions (Figure 1, right). Nevertheless, this becomes an experimentally usable level of accuracy.

However, this classifier (cscape.biocompute.org.uk) still remains generic, in that it is trained on COSMIC data [20], derived from a variety of cancer types. Thus further improvement can be achieved by developing predictors trained on, and specific to, individual types of cancer. As an example, using data from the Cancer Genome Atlas [22] and the International Cancer Genome Consortium [23] we have derived specialist predictors for particular types of cancer. Thus for a specialised breast cancer predictor (CScape-brca), we can achieve a baseline predictor (coding regions, all nucleotide positions) with approximate 80% accuracy and capable of a greater test accuracy, if restricted to higher confidence predictions.

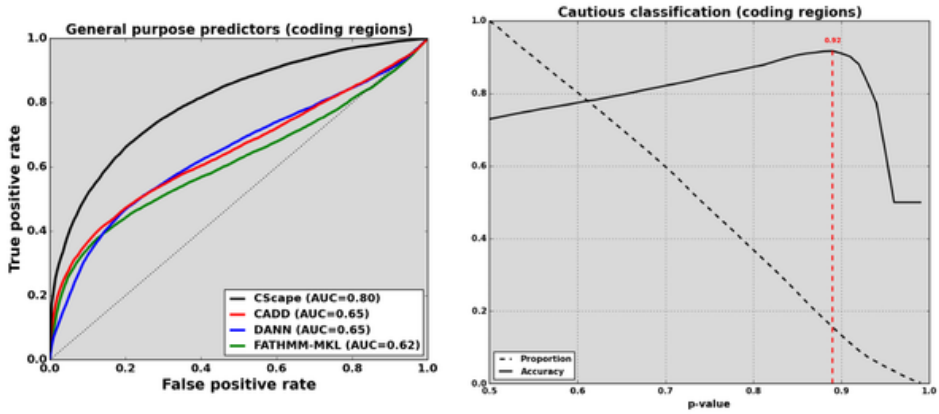

Figure 1. Left: ROC curves for a comparison of the proposed classifier (*CScape*) for predictions in the coding regions of the cancer genome, against alternative methods. Right: the solid curve gives the test accuracy (approximately balanced), the dashed curve gives the proportion of nucleotide positions with high enough confidence for prediction at the given level of test accuracy: this dashed curve is derived from test data, the 12.1% figure quoted in the text is for whole-genome prediction (coding regions).

3. Discussion

These new methods indicate that usable levels of accuracy can be achieved for predicting the pathogenic impact of genetic variants. Aside from predicting possible new drug targets, the refined insights from these tools could assist in establishing subtypes of disease, hence improving personalised approaches to therapy and predicting an individual's response to a drug. There is the prospect that these methods can be further enhanced through the incorporation of additional sources of data. Aside from disease-specific prediction, another avenue for investigation would be region-specific prediction, for example, dedicated predictors for non-coding variants residing at or near splicing regions. We will report on these developments in later work.

References

[1] H. Shihab, M. Rogers, J. Gough, M. Mort, D. Cooper, I. Day, T. Gaunt and C. Campbell, An integrative approach to predicting the functional effects of non-coding and coding sequence variation, *Bioinformatics*, **31** (2015), 1536-1543.
[2] C. Campbell, C. and Y. Ying, *Learning with Support Vector Machines*, Morgan and Claypool, 2011.
[3] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature*, **489** (2012), 57–74.
[4] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov and S.R. Sunyaev, S.R., A method and server for predicting damaging missense mutations, *Nature Methods*, **7** (2010), 248-249.
[5] P. Kumar, S. Henikoff and P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, *Nature Protocols*, **4** (2009), 1073–81.
[6] B. Reva, Y. Antipin and C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, *Nucleic Acids Research*, **39** (2011), e118.
[7] M. Kircher, D. Witten, P. Jain, B. O'roak, G. Cooper, G and J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, **46**, (2014) 310-315.
[8] D. Quang, D., Y. Chen X. and Xie. DANN: a deep learning approach for annotating the pathogenicity of genetic variants, *Bioinformatics*, **31** (2014), 761–763.

[9] Y. Choi and A.P. Chan, PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels, *Bioinformatics*, **31** (2015), 2745–2747.
[10] C. Douville, D.L. Masica, P.D. Stenson, D.N. Cooper, D.G. Gygax, R. Kim, M. Ryan and R. Karchin, Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-indel), *Human Mutation*, **37** (2016), 28–35.
[11] Folkman, Y. Yang, Z. Li, B. Stantic, A. Sattar, M. Mort, D.N. Cooper, Y. Liu and Y. Zhou, DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels, *Bioinformatics*, **31** (2015), 1599–1606.
[12] H. Zhao, Y. Yang, H. Lin, X. Zhang, M. Mort, D.N. Cooper, Y. Liu and Y. Zhou, DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels, *Genome Biology*, **14** (2013), R23.
[13] L. Hindorff, P. Sethupthay, H. Junkins, E. Ramos, J. Mehta, F. Collins and T. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *PNAS*. U.S.A., **106** (2009), 9362–9367.
[14] M. Ferlaino, M. Rogers, H. Shihab, T. Gaunt, M. Mort, D. Cooper and C. Campbell. An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. *Journal submission* (2017).
[15] H. Shihab, M. Rogers, C. Campbell and T. Gaunt. GTB - An Online Genome Tolerance Browser. *BMC Bioinformatics* **18** (2017), 20.
[16] H. Shihab, M. Rogers, C. Campbell and T. Gaunt, HiPred: an integrative approach for predicting haploinsufficiency in the human genome. *Bioinformatics*, doi.10.1093/bioinformatics (2017).
[17] M. Rogers, H. Shihab, T. Gaunt and C. Campbell, CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Journal submission* (2017).
[18] H. Shihab, J. Gough, M. Mort, D. Cooper, I. Day and T. Gaunt, Ranking Non-Synonymous Single Nucleotide Polymorphisms based on Disease Concepts, *Human Genomics*, **8** (2013), 11.
[19] http://www.ensemb.org/info/docs/tools/vep/index.html
[20] http://cancer.sanger.ac.uk/cosmic/help/gene/analysis
[21] The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, *Nature*, **491** (2012), 56–65.
[22] https://cancergenome.nih.gov/
[23] http://icgc.org/