BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 6 2006, pages 773–774
doi:10.1093/bioinformatics/btk031

Data and text mining

SUSPECTS: enabling fast and effective prioritization of positional candidates

E. A. Adie*, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard

Medical Genetics Section, School of Molecular and Clinical Medicine, University of Edinburgh EH4 2XU, UK

Received on October 26, 2005; revised on December 12, 2005; accepted on December 28, 2005
Advance Access publication January 19, 2006
Associate Editor: Satoru Miyano

ABSTRACT

Summary: SUSPECTS is a web-based server which combines annotation and sequence-based approaches to prioritize disease candidate genes in large regions of interest. It uses multiple lines of evidence to rank genes quickly and effectively while limiting the effect of annotation bias to significantly improve performance.

Availability: SUSPECTS is freely available at http://www.genetics.med.ed.ac.uk/suspects/

Contact: euan.adie@ed.ac.uk

Supplementary information: A quick-start guide in Macromedia Flash format is available at http://www.genetics.med.ed.ac.uk/suspects/help.shtml and Excel spreadsheets detailing the comparative performance of the software are included as Supplementary material.

INTRODUCTION

When searching for the genetic basis of disease the regions of interest identified through complex-trait linkage studies regularly exceed 30 cM in size and can contain hundreds of genes (McCarthy *et al*., 2003). Existing tools to help researchers to prioritize candidates for further study can be separated into two distinct classes; those based on functional annotation (Perez-Iratxeta *et al*., 2002; Freudenberg *et al*., 2002; Van Driel *et al*., 2003; Turner *et al*., 2003; Tiffin *et al*., 2005) and those based on sequence features (Adie *et al*., 2005; Lopez-Bigas *et al*., 2004).

Methods based on functional annotation can suffer from annotation bias as they are unable to deal with genes lacking sufficiently detailed annotation. Sequence-based methods make use of intrinsic characteristics of genes like length, homology to genes in other species and base composition. As these characteristics can be readily computed from sequence they avoid the problem of annotation bias. However, sequence-based methods prioritize genes on the basis of their potential for involvement in disease in general rather than involvement in the specific disease of interest to the user.

SUSPECTS is a novel, consolidated approach that combines the increased precision of annotation-based methods with the better recall of sequence-based methods, avoiding the problems outlined above. Given a set of existing candidate genes for a particular complex or oligogenic disease, it effectively automates further candidate gene selection from large regions on the principle that genes involved in that disease will tend to share the same or similar annotation, reflecting common biological pathways.

*To whom correspondence should be addressed at MMC, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

PRIORITIZING CANDIDATES WITH SUSPECTS

Users of SUSPECTS can enter a region of interest by specifying flanking markers, chromosomal coordinates or bands. Alternatively, the software will examine a region of interest automatically centred on a single marker.

Users then enter the name of the disease to be considered; the software will automatically retrieve genes implicated in that disorder from OMIM (Hamosh *et al*., 2002), HGMD (Cooper *et al*., 1998) and GAD (Becker *et al*., 2004). Alternatively users can manually enter a list of genes thought to be involved in pathogenesis of the disease. These genes are known as the 'training set'.

Each positional candidate gene is then scored automatically (see Methodology). Higher scores represent better candidates. The user is presented with a graphical overview of the region of interest (Fig. 1). The graphical overview is a hyperlinked image map that can be used to obtain more detailed information about each

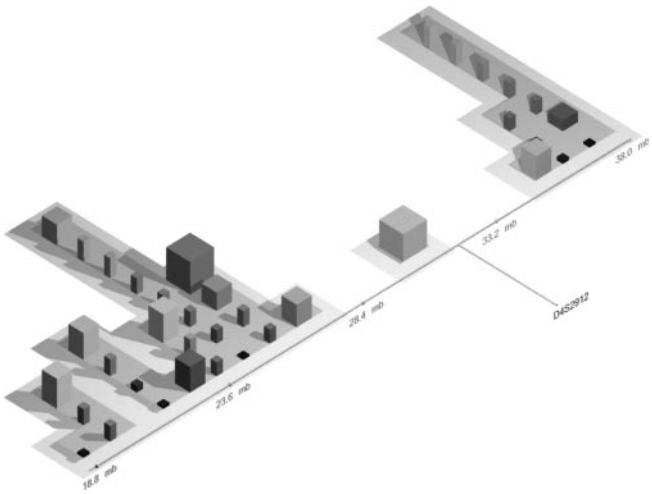

Fig. 1. Graphical overview of results produced by SUSPECTS. SUSPECTS presents the user with a graphical overview of the region of interest. Each gene in the region is represented as a coloured 3D shape. The height, width and colour of these shapes represent the score, number of lines of evidence contributing to that score and the literature-based relevance of the gene, respectively. Literature-based relevance is determined by searching PubMed (shapes are blue if the gene that they represent are mentioned in abstracts containing the name of the disease under study and orange otherwise). Each shape is a hyperlink to more detailed information about that gene further down the results page.

candidate gene and the reasoning behind its score. The list of candidate genes ranked by score is presented as a table underneath the graphical overview.

METHODOLOGY

Each gene in the region of interest is scored on its suitability as a candidate for further study based on four lines of evidence; first by Prospectr (Adie *et al*., 2005) on the basis of its sequence features, second by the extent of coexpression with the training set based on GNF expression data (Su *et al*., 2002), third by the number of rare (found in <5% of all proteins) Interpro domains shared with the training set and finally by the level of semantic similarity (Lord *et al*., 2003) that the GO terms assigned to it share with the GO terms assigned to genes in the training set.

The four scores are then combined. Each score is weighted depending on the amount of information available for each line of evidence. If little or no information is available then the importance of that score is decreased accordingly. This ensures that the scores of genes which lack sufficiently detailed GO terms or expression profiles do not suffer from annotation bias. The final score ranges from 0 to 100 where 100 represents a perfect match between the candidate gene and all genes in the training set.

COMPARATIVE PERFORMANCE

Approaches based on functional annotation rely on good quality information being available for each possible candidate gene. Conversely, SUSPECTS is able to prioritize all genes including those which lack detailed GO, domain or expression data, although when available those lines of evidence contribute favourably to overall performance.

The performance of SUSPECTS was tested with a set of oligogenic and complex disorders including Alzheimer's disease, hypertension, autism and systemic lupus erythematosus. The set is derived from that used by Turner *et al*. to test POCUS, an annotation-based classifier (Turner *et al*., 2003).

At least three implicated genes for each disease were available. For each implicated gene, a region of interest was created containing the implicated gene itself (the 'target gene') and every gene within 7.5 Mb on either side. On an average each region of interest contained 155 genes. An associated training set was then created containing the remaining implicated genes for each disorder.

We first ranked each region of interest using a classifier based on sequence features alone (Prospectr). On average the target gene was in the top 31.23% of the resulting ranked lists of candidates and in the top 5% of those lists 20 times out of 155 (13%).

In comparison, on average the target gene was in the top 12.93% of the ranked list from SUSPECTS, which took both the region of interest and the training set as input in each case. The target gene was in the top 5% of the ranked list 87 times out of 155 (56%). The test results for both the sequence features classifier and SUSPECTS have been made available as Supplementary information.

In conclusion, SUSPECTS significantly improves on the performance of candidate prioritization methods which use annotation or sequence data alone and is of value to researchers faced with large regions of interest. It is fast, easy to use and freely available on the World Wide Web at http://www.genetics.med.ed.ac.uk/suspects/

Conflict of Interest: none declared.

REFERENCES

Adie,E.A. *et al*. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.

Becker,K.G. *et al*. (2004) The Genetic Association Database. *Nat. Genet*., **36**, 431–432.

Cooper,D.N. *et al*. (1998) The human gene mutation database. *Nucleic Acids Res*., **26**, 285–287.

Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18**, S110–S115.

Hamosh,A. *et al*. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res*., **30**, 52–55.

Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*., **32**, 3108–3114.

Lord,P.W. *et al*. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

McCarthy,M. *et al*. (2003) New methods for finding disease-susceptibility genes: impact and potential. *Genome Biol*., **4**, 119.

Perez-Iratxeta,C. *et al*. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet*., **31**, 316–319.

Su,A. *et al*. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.

Tiffin,N. *et al*. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*., **33**, 1544–1552.

Turner,F. *et al*. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*., **4**, R75.

Van Driel,M.A. *et al*. (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. J. Hum. Genet*., **11**, 57–63.