Nucleic Acids Research Advance Access published May 3, 2007

Nucleic Acids Research, 2007, 1–5
doi:10.1093/nar/gkm223

Update of the G2D tool for prioritization of gene candidates to inherited diseases

Carolina Perez-Iratxeta[1,*], Peer Bork[2] and Miguel A. Andrade-Navarro[1,3]

[1]Ontario Genomics Innovation Centre, Ottawa Health Research Institute, 501 Smyth, Ottawa, ON, Canada K1H 8L6, [2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany and [3]Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Canada

Received January 29, 2007; Revised March 20, 2007; Accepted March 28, 2007

ABSTRACT

G2D (genes to diseases) is a web resource for prioritizing genes as candidates for inherited diseases. It uses three algorithms based on different prioritization strategies. The input to the server is the genomic region where the user is looking for the disease-causing mutation, plus an additional piece of information depending on the algorithm used. This information can either be the disease phenotype (described as an online Mendelian inheritance in man (OMIM) identifier), one or several genes known or suspected to be associated with the disease (defined by their Entrez Gene identifiers), or a second genomic region that has been linked as well to the disease. In the latter case, the tool uses known or predicted interactions between genes in the two regions extracted from the STRING database. The output in every case is an ordered list of candidate genes in the region of interest. For the first two of the three methods, the candidate genes are first retrieved through sequence homology search, then scored accordingly to the corresponding method. This means that some of them will correspond to well-known characterized genes, and others will overlap with predicted genes, thus providing a wider analysis. G2D is publicly available at http://www.ogic.ca/projects/g2d_2/

BACKGROUND

In the effort to identify which gene or genes produce what disease phenotype, geneticists are producing an increasing number of genome-wide linkage analyses and gene association studies that are generating too many to test candidate genes. For this reason, during the past five years, the problem of automating the prioritization of candidate genes to inherited diseases has received increasing attention from the bioinformatics community. Computational approaches were made possible due to the availability of the complete human genome sequence and to considerable developments on database annotation and data integration for molecular biology databases. As a result, a number of methods that address this problem have been published (1–12). These methods apply a variety of approaches exploiting known or deduced pieces of information that range from using only the genomic sequence of the target region to data mining analyses that include literature and different annotation systems. For details about these methods see the Introduction sections of (13) and (9). A recent work where many of these methods were applied to the prioritization of gene candidates for obesity and type 2 diabetes, shows that there are many discrepancies between the produced candidate gene lists (13). This study suggests the concurrent use of as many approaches as possible when working with particular diseases, especially as there are large differences in the kind and amount of useful information that it is available for each disease. Following this idea, we are developing and maintaining G2D (genes to diseases), a resource that seeks to encompass an increasing number of methodologies to obtain disease-related genes.

G2D is a web resource for prioritizing genes as candidates to inherited diseases, which started as a public server in 2002. At the time, we provided the users with pre-analyses of hundreds of inherited diseases by applying a literature and data mining algorithm based on fuzzy binary relations (2) using the disease's phenotype as input. In 2005, we made the algorithm available as a web tool (14). Now, we have included two more ways of prioritization that apply other principles. One is the use of genes in a different genomic location already known or suspected to produce a similar variant of the disease of interest. The expectation is that the disease-responsible gene will have a similar function to those. This is a more stringent approximation and requires more information

*To whom correspondence should be addressed: Tel: +1 613 737 8899 Ext 73255; Fax: +1 613 739 6294; Email: cperez-iratxeta@ohri.ca

about the disease, but it is obviously correct in some cases. The SUSPECTS method (7) has already exploited the idea of gene similarity. The second approach added to G2D, uses as input a second locus where the phenotype has also been mapped in addition to the one where the user is looking for the mutation (as in complex diseases). The expectation is that the products of the genes in several of the loci may be functionally associated, e.g. by participating in the same pathway, or even by being part of a protein complex (9,11). The system will prioritize the candidates according to whether functional interactions (both known and predicted) between proteins encoded in genes of both loci are detected. We used the human protein interactions from the STRING database (15) as our source database.

Besides this web implementation, G2D has been applied by us to the prioritization of genes for complex diseases in two different studies so far: the aforementioned collective computational study to suggest candidates for obesity and type 2 diabetes (13), and to the selection of asthma candidates for genotyping in two linkage regions in a French Canadian founder population (manuscript in preparation).

OUTLINE OF THE G2D SERVER

Given a chromosomal region where the user is looking for candidate genes, there are three ways of using the G2D server depending on the algorithm that will be applied. The first option takes as input the phenotype of the disease of interest described by means of an online Mendelian inheritance in man (OMIM) disease entry (16). The system will prioritize the genes according to the description of the phenotype as provided by the MeSH disease annotations (http://www.nlm.nih.gov/mesh/) to the linked bibliography in the corresponding OMIM entry. For details of this method see (2). A second option is to input one or several human or mouse genes that are already known or are suspected to be involved in the disease. The system will prioritize the genes in the target region according to their similarity to the known gene(s) as given by their GO annotations and high sequence homology. In order to do this, we compute scores for all GO terms according to their similarity to the GO terms of the known genes (measured in terms of their Resnik similarity). The measure favors rare GO annotations. Next, we apply this GO scoring system to the annotated genes in Entrez Gene. Their corresponding RefSeq proteins are compared through BLASTX to the chromosomal target region, and the scoring is transferred to the hits that are below the selected E-value. The third option can be used when another chromosomal region has been also linked to the disease of interest. The system will look for protein–protein interactions in the STRING database, both known and predicted that may be occurring between a gene in the region of interest and a gene in the second region. Since many interactions in STRING are predicted, every interaction has associated a score (STRING score) that reflects how likely is to be a true interaction. We use the STRING score to sort the candidates within the region of interest. Candidates that are more likely to interact with a gene in the second region are then prioritized. In our benchmark (see Conclusion and Supplementary page in our web site), we observed higher precision when STRING scores are very high. Additionally, it is possible to access from our server a database of pre-calculated results for more than 550 monogenic diseases on published linkage regions using the phenotype method.

In the next sections, we describe the different use options through examples. We have prepared step-by-step web tutorials that contain very detailed information. They can be accessed from our web server page (see Tutorial section at our web site).

USE OF PHENOTYPE

The system will prioritize the genes according to the description of the phenotype and its precomputed associations to gene features as extracted from the literature and the Entrez Gene database. The input consists simply of the region where the user is looking for the mutation, and the phenotype of interest, given as an OMIM identifier. For example, suppose that you are interested in candidate genes to Hirschsprung disease in 22q13.2. You have to enter in the appropriate boxes a MIM number that defines the disease and the target region. In this particular case, you would enter 131224 (the ID of the OMIM entry for Hirschsprung disease) in the PHENOTYPE BOX, and q13.2 in the LOCATION BOX. You would also select the chromosome, 22 and 'Bands' in the LOCATION BOX as you have entered the genomic region in the form of a cytogenetic band. You can refer to our web tutorial for more input and output options. The output is an ordered list of candidate genes, both known and predicted, ordered according to their susceptibility for producing the phenotype. For each candidate, you can explore any overlapping with ESTs and pseudogenes, as well as trace back the reasoning the system followed to associate the candidates to the disease.

USE OF KNOWN GENES

This method can be used when one or more genes are already known or are suspected to be related to the disease. In order to use it, you have to provide your target region in the LOCATION BOX in the same way as for the phenotype method, plus one or several human genes associated to the disease or related mouse genes. Following with our example, Hirschsprung disease is suspected to be a multigenic disease, and it is known that mutations in the endothelin-B receptor (ETRB) gene cause one of the variants of this disease. In order to find similar or functionally related genes to ETRB in 22q13.2, you could input the Entrez Gene IDs of human and mouse ETRBs in the KNOWN GENES BOX. These are 2861 and 13 618, respectively. Like for the phenotype method, the output is an ordered list of candidate genes that you may explore in a similar manner. Candidates that overlap

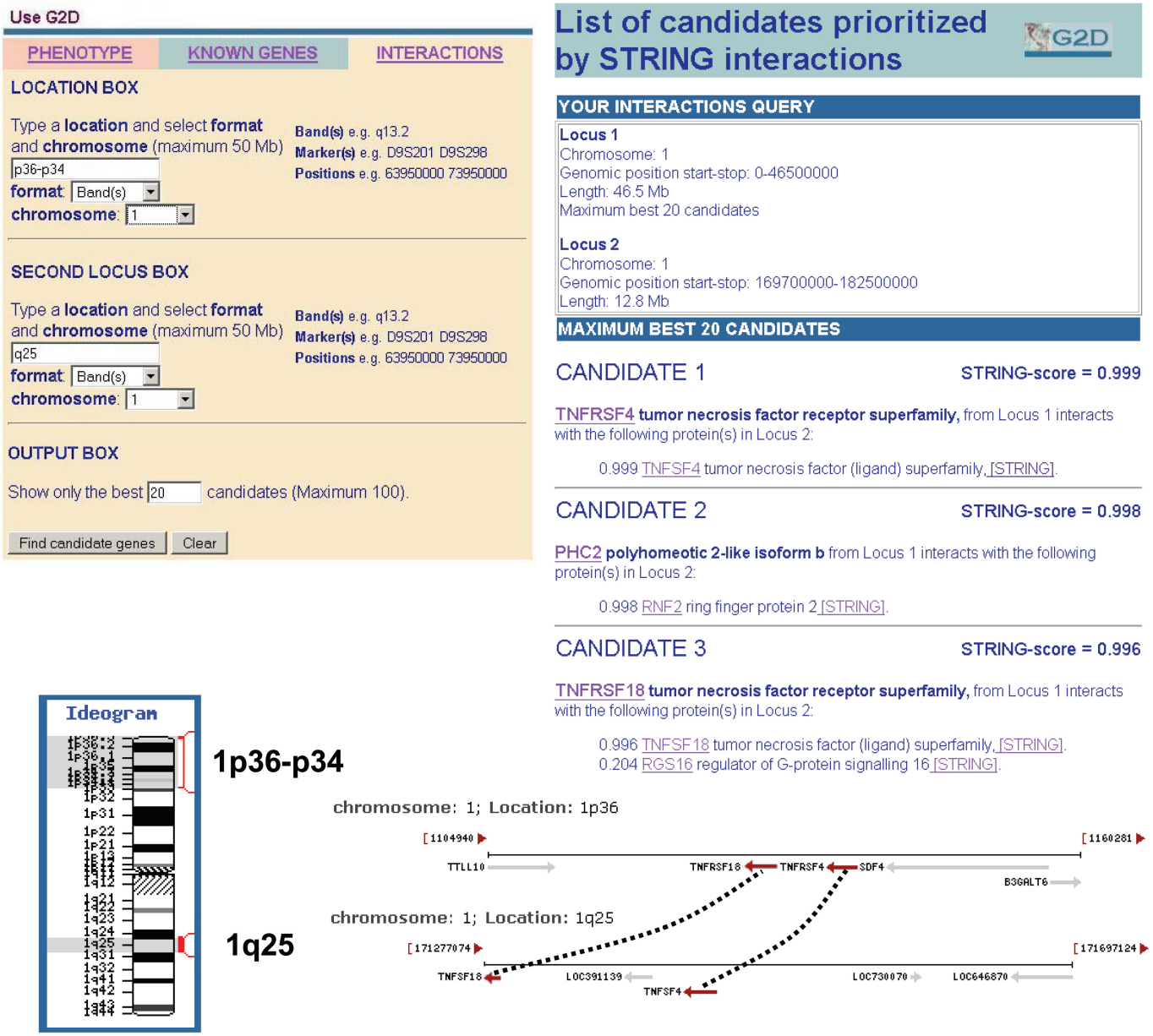

Figure 1. G2D web server interface used to detect genes associated to MIS using the protein-protein method. Top left: the user inputs the coordinates for Locus 1 (1p36-p34) and Locus 2 (1q25) that have been genetically linked to MIS. Top right: G2D displays 20 gene candidates in Locus 1 that code for proteins that interact with proteins encoded by genes found in Locus 2. Bottom: tumor necrosis factor TNFSF4 and its receptor TNFRSF4 are encoded in Locus 2 and Locus 1, respectively, being therefore good candidates; a similar pair of genes (TNFSF8/TNFRSF8) appears as candidate 3.

with genes that have (STRING) interactions with any of the known genes input by the user are flagged.

USE OF STRING PROTEIN–PROTEIN INTERACTIONS

The third method for finding candidate genes for multiple gene phenotypes relies on protein–protein functional interactions, either known or predicted. The rational is that mutations on two proteins that participate on the same pathway, or are directly interacting, will produce the same or very similar phenotypes. The input for this method consists of the target region, entered in the LOCATION BOX like in the previous two methods, and a second region where the phenotype of interest has been also mapped to. The second region is entered in the SECOND LOCUS BOX in the same manner, specifying format and chromosome. The output is a list of genes in the target region that may interact, according to STRING (15), with any genes(s) in the SECOND BOX locus. Candidates are sorted by how likely are their corresponding interactions to be true indicated by their associated STRING scores with 0.99 as maximum value. As an

illustration, suppose that you are interested in candidates for myocardial infarction susceptibility (MIS) on 1p36-p34, one of the loci where the condition has been linked to (17). The OMIM entry for MIS (608446) shows that it is linked as well to regions 1q25, 6q25.1 and 1p22.1. By entering 1q25 as the second locus, the output page displays 20 candidates corresponding to proteins in 1p36-p34 that interact, according to the STRING database, with at least one protein in 1q25. The first candidate (see Figure 1) is the tumor necrosis factor receptor superfamily, member 4 TNFRSF4 (Entrez Gene ID 7293, located on 1p36). This gene comes out as the top candidate because of its direct interaction (STRING score = 0.999) with TNFSF4, the tumor necrosis factor (ligand) superfamily, member 4 (tax-transcriptionally activated glycoprotein 1, 34 kDa) (Entrez Gene ID 7292). The latter has been recently associated to atherosclerosis susceptibility increasing the risk for myocardial infarction (18) making TNFRSF4 a very good candidate. TNFRSF4 is therefore a quite attractive candidate from a biological point of view, not yet tested but within the linkage peak identified by the group that linked MIS to 1p36-p34 (E.J. Topol, personal communication).

CONCLUSION

The G2D web server allows users to prioritize candidate genes for practically any disease phenotype in OMIM through three different methods. The results of benchmarking the methods on 227 diseases that have been shown to be caused by more than one gene, are given in detail in our web server (see Supplementary section at our web site), and they can serve as a guidance on the expected performance of each method. The phenotype and the known-genes methods show a similar performance prioritizing the responsible gene in average among the top 25 and 27% of all genes in the band, respectively, when considering target bands of 30 MB (averaging a content of 300 genes). It must be taken into account, however, that more input information about the molecular cause of the disease is required by the known-genes method. The protein–protein interactions method has a very low recall but shows high precision when the STRING scores associated to the candidates are 'high' and 'very high' (0.9 score value and higher). This method could be applied to a proportion of one in four cases, and in such situations the responsible gene was among the top 10 candidates, 70% of the times for bands of size 30 MB.

Running times for the algorithms vary from very few seconds to almost immediate yielding of results. Users can examine, for all three procedures, the rational used by the system to make the prioritization, keeping the process transparent. We support this by including extensive hyperlinks to related resources such as NCBI sequence databases, Gene Ontology and the UCSC Genome Browser.

SUPPLEMENTARY DATA

Supplementary Data are available at G2D web site.

ACKNOWLEDEGMENTS

We thank Christian von Mering for his assistance with STRING data. This work was supported by funding from the Canadian Institutes of Health Research, the Ontario Research Development Challenge Fund, the Canadian Foundation for Innovation, and the Canada Research Chair Program. Funding to pay the Open Acess publication charges for this article was provided by CIHR grant MOP-77640.

Conflict of interest statement: None declared.

REFERENCES

1. Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18**(Suppl. 2), S110–S115.
2. Perez-Iratxeta,C., Bork,P. and Andrade,M. A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.
3. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
4. Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucle Acids Res.*, **32**, 3108–3114.
5. Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, **33**, 1544–1552.
6. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P., Leunissen,J.A., Brunner,H.G. and Vriend,G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, **33**, W758–W761.
7. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.
8. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
9. George,R.A., Liu,J.Y., Feng,L.L., Bryson-Richardson,R.J., Fatkin,D. and Wouters,M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, **34**, e130.
10. Ma,X., Lee,H., Wang,L. and Sun,F. (2006) CGI: a new approach for prioritizing genes by Combining Gene expression and protein-protein Interaction data. *Bioinformatics*, **23**, 215–21.
11. Oti,M., Snel,B., Huynen,M.A. and Brunner,H.G. (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.
12. Rossi,S., Masotti,D., Nardini,C., Bonora,E., Romeo,G., Macii,E., Benini,L. and Volinia,S. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleis Acids Res.*, **34**, W285–W292.
13. Tiffin,N., Adie,E., Turner,F., Brunner,H.G., van Driel,M.A., Oti,M., Lopez-Bigas,N., Ouzounis,C. *et al.* (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.*, **34**, 3067–3081.
14. Perez-Iratxeta,C., Wjst,M., Bork,P. and Andrade,M.A. (2005) G2D: a tool for mining genes associated with disease. *BMC Genet.*, **6**, 45.
15. von Mering,C., Jensen,J.L., Kuhn,M., Chaffron,S., Doerks,T., Kruger,B., Snel,B. and Bork,P. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.

17. Wang,Q., Rao,S., Shen,G.Q., Li,L., Moliterno,D.J., Newby,L.K., Rogers,W.J., Cannata,R., Zirzow,E. *et al.* (2004) Premature myocardial infarction novel susceptibility locus on chromosome 1P34-36 identified by genomewide linkage analysis. *Am. J. Hum. Genet.*, **74**, 262–271.

18. Wang,X., Ria,M., Kelmenson,P.M., Eriksson,P., Higgins,D.C., Samnegard,A., Petros,C., Rollins,J., Bennet,A.M. *et al.* (2005) Positional identification of TNFSF4, encoding OX40 ligand, as a gene that influences atherosclerosis susceptibility. *Nat. Genet.*, **37**, 365–372.