Computational Systems for Selection and Priorization of Candidate Genes that Underlie Human Hereditary Disease

Adášková, Jana
2007

Computational Systems for Selection and Prioritization of Candidate Genes that Underlie Human Hereditary Disease

Post-Graduate Student:
MGR. JANA ADÁŠKOVÁ
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
adaskova@euromise.cz

Supervisor:
PROF. RNDR. JANA ZVÁROVÁ , DRSC.
Department of Medical Informatics
Instutite of Computer Science of the ASCR, v. v. i.
Pod Vodárenskou věží 2
182 07 Prague 8, CZ
zvarova@euromise.cz

Field of Study:
Biomedical Informatics

The work was supported by the grant 1M06014 of the Ministry of Education of the Czech Republic.

Abstract

The aim of this paper is to present an overview of six independent computational methods for the selection and prioritization of candidate genes for human diseases and, rather than selecting a best method, to offer the prospective user a better understanding of the inputs, outputs and functionality of each available method. A survey of these methods also offers the bioinformatics community an opportunity to assess the efficacy of current computational approaches to disease gene identification, and informs future directions for research in this field.

Keywords: candidate gene selection, prioritization, data mining, text mining, human hereditary disease.

1. Introduction

Few areas have moved as fast as human disease gene identification. Before 1980, very few human genes had been identified as disease loci. In the 1980s, advances in recombinant DNA technology allowed a new approach, positional cloning, sometimes given the rather meaningless label "reverse genetics" [11]. The number of disease genes identified started to increase quickly. Now the human and other genome projects have made available a vast range of resources - maps, clones, sequences, expression data and phenotypic data. Identifying novel disease genes has become commonplace and is currently occurring on a weekly basis. Some of the routes that have been followed to identify human disease genes summarizes Figure 1. If the figure seems complicated, that is because there is no standard procedure for gene identification. All pathways converge on mutation testing in a candidate gene, but there is not one single entry point, and there is no unique pathway to the candidate gene. For discussion of the principles, we can divide the methods into those that do not require us to know the chromosomal location of the disease locus and those that depend on this knowledge. Most genes are identified by defining a candidate gene on the basis of both its chromosomal location and its properties [11].

Unlike Mendelian traits, in which a mutation in one gene is causative, or oligogenic traits, where several genes are sufficient but not necessary, complex traits are caused by variation in multiple genetic and environmental factors, none of which are sufficient to cause the trait [8]. The contribution of any given gene to a complex trait is usually modest. In addition, complex traits often encompass a variety of phenotypes and biological mechanisms, making it difficult to determine which genes to study [7].

As a result, traditional methods of genetic discovery, such as linkage analysis and positional cloning, while widely successful in identifying the genes for Mendelian traits, have had more limited success in identifying genes for complex traits. Candidate gene studies have had encouraging success, yet this approach requires an effective method for deciding a priori which genes have the greatest chance of influencing susceptibility to the trait [3]. Recent advances in genotyping technology have provided researchers with the ability to test association in hundreds of genes relatively quickly, and even the entire genome through a genome-wide association study.

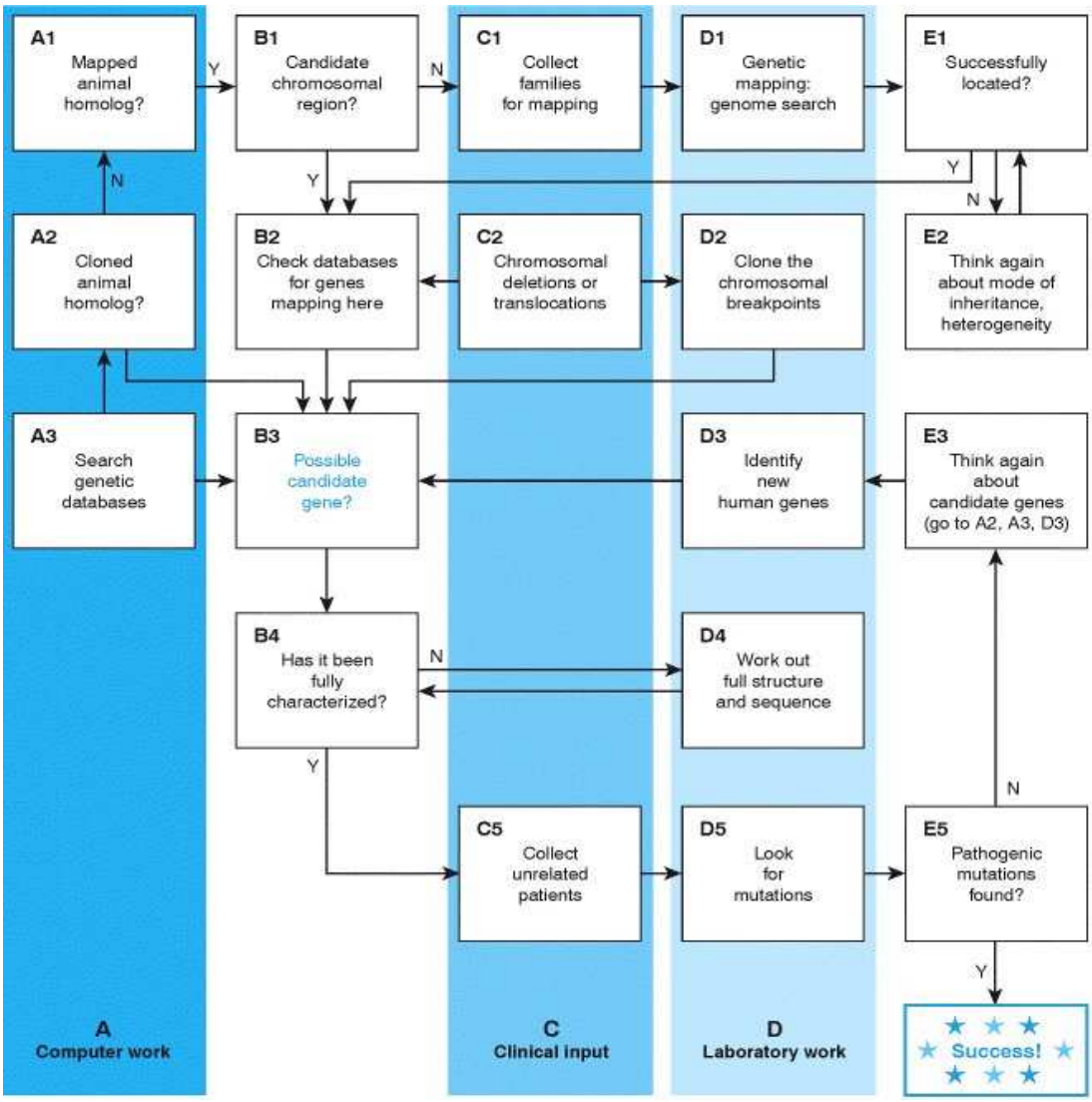

Figure 1: Scheme of the routes to identify human disease genes.

Therefore, one of the greatest challenges in disease association study design remains the intelligent selection of 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 [10]. As a result, a number of methods that address this problem have been published. 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 analysis that include literature and different annotation systems. In this paper we present more details about six independent methods and what we believe to be useful illustration of application of these methods.

2. Existing methods

Note: Detail information about data sources and ontologies used in methods are listed in Table 1 at the end.

2.1. PROSPECTR

It can be shown that genes implicated in disease share certain patterns of sequence based features that can provide a good basis for automatic prioritization of candidate genes by machine learning [2]. PROSPECTR (**PR**i**O**rization by **S**equence & **P**hylogenetic **E**xtent of **C**andida**T**e **R**egions) is an alternating decision tree which has been trained to differentiate between genes

likely to be involved in disease and genes unlikely to be involved in disease. This alternating decision tree with fifteen nodes was produced by training on the training set of genes. PROSPECTR requires only basic sequence information and by using this sequence-based features like gene length, protein length and the percent identity of homologs in other species as input a score (ranging from 0 to 1) can be obtained for any gene of interest. The score itself is a measure of confidence in the classification. Genes with scores over a certain threshold, 0.5, are classified as likely to be involved in some form of human hereditary disease while genes with scores under that threshold are classified as unlikely to be involved in disease. Given this score we can also roughly estimate how much more or less likely it is that a particular gene is involved in human hereditary disease.

Tests on an independent data set of genes taken from the Human Gene Mutation Database suggest that PROSPECTR will, on average, enrich a list of about 200 genes two-fold 74 % of the time, five-fold 33 % of the time and twenty-fold 8 % of the time. 95 % of the time the list was enriched one and a half fold - that is to say that the target gene was in the top three-quarters of the ranked list [2].

The web interface of PROSPECTR allows researchers to obtain a ranked list of genes ordered by the scores for regions of the genome or individual gene of interest. The software is now freely accessible together with training and test sets of genes at URL: www.genetics.med.ed.ac.uk/prospectr/ .

2.2. SUSPECTS

SUSPECTS is a consolidated candidate gene approach that combines the increased precision of annotation-based methods with the better recall of sequence-based methods. 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 [1]. In principle SUSPECTS is built on top of the PROSPECTR candidate prioritization system by incorporating annotation data from Gene Ontology (GO), InterPro and expression libraries.

The server takes two inputs - firstly, the coordinates of the genomic region that you are interested in. You can specify this using markers, bands, chromosomal coordinates or genes. The second input is a list of genes thought to be involved in pathogenesis of the same complex disease as the one you are interested in (as a shortcut, you may simply enter the name of the disease; The software will automatically retrieve genes implicated in that disorder from databases OMIM, the HGMD and GAD). This list is known as the "training set".

Each gene in the region of interest is then scored automatically on its suitability as a candidate for further study based on four lines of evidence: first by PROSPECTR (*see above*) on the basis of its sequence features, second by the extent of coexpression with the training set based on GNF (Genomics Institute of the Novartis Research Foundation) expression data (scores depend on how well correlated any matching profiles are), 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 that the GO (Gene Ontology) terms assigned to it share with the GO terms assigned to genes in the training set [1]. 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. Higher scores represent better candidates. The list of candidate genes ranked by score is presented as the graphical overview of region of interest which is a hyperlinked image map that can be used to obtain more detailed information about each candidate gene and the reasoning behind its score.

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

2.3. Disease Gene Prediction (DGP)

DGP (**D**isease **G**ene **P**rediction) is a database of human genes with their probability of being involved in a hereditary disease. The genes that are already known to be involved in monogenic hereditary disease have been shown to follow specific sequence property patterns that would make them more likely to suffer pathogenic mutations. Based on these patterns, DGP is able to assign probabilities to all the genes that indicate their likelihood to mutate solely based on their sequence properties. This probability has been assigned with a data mining algorithm using parameters that have been shown to follow specific trends in the already known disease genes. In particular, the properties analysed by DGP are

protein length, degree of conservation, phylogenetic extent and paralogy pattern [6].

The performance of this method has been assessed previously on a test dataset by building a model with a part of the data (learning set: 75 %) and testing with the rest (test set: 25 %). On average 70 % of the disease genes in the test set were predicted correctly with 67% precision [6]. Genes involved in complex diseases, similarly to monogenic disease genes, need to have mutations or variations in the gene sequence that impair or modify the function or expression of the protein they encode, leading to a disease phenotype. Thus, we believe that, although DGP has been designed for the prediction of Mendelian diseases, it can also be useful for the identification of complex-disease genes as it will identify those genes with higher likelihood of suffering mutations. DGP is freely available on the World Wide Web at http://cgg.ebi.ac.uk/services/dgp/ .

2.4. GeneSeeker

GeneSeeker is a web-based data mining tool that filters positional candidate disease genes based on expression and phenotypic data from both human and mouse. It queries nine different databases through the web, guaranteeing that the most recent data are used at all times and removing the need for local repositories, and then combines this information using Boolean operators. This results in a quick overview of candidate genes in the genetic region of interest. The GeneSeeker system is built in a modular fashion, making it easy to maintain and expand [4]. The GeneSeeker is freely available via the web interface at www.cmbi.ru.nl/geneseeker/ .

The input for GeneSeeker is the genetic mapping information. This can be a chromosome, a chromosome arm, or a range and if necessary, a combination of genetic localization can be also entered. Second input is the tissue names where either direct RNA expression or phenotypic expression of the candidate gene is expected. The query entered by the user is pre-processed for Human and Mouse databases and subsequently reformulated into the format appropriate for each database. GeneSeeker uses the Genome Database (GDB) and the Online Mendelian Inheritance in Man (OMIM) to obtain human mapping data. Genes searched in specified chromosome location in humans are also translated with the aid of an "Oxford-grid", to search the appropriate Mouse databases (e.g. Mouse Genome Database (MGD)). The key tissues affected by the genetic disorder are used to query phenotypic or expression related databases, including the OMIM phenotype fields, Swissprot, and Medline for data on human phenotypes and the Gene Expression Database (GXD), the Transgenic/Targeted Mutation Database (TBASE), and the Mouse Locus Catalog (MLC) for gene expression patterns and phenotypes in mice [4]. The output of the analysis is presented in four tables: (1) A list of human genes in the correct genetic region and matching the specified expression profile, (2) a list of mouse genes matching the syntenic region as well as the expression profile, but with no matching human gene name, (3) a list of mouse genes found in the syntenic region in mouse, for which the homologous human gene is found to map outside the critical interval, and (4) a list of all the remaining human genes that are present in the genetic interval, but which do not match the expression profile.

In a test using 10 syndromes, GeneSeeker reduced the candidate gene lists from an average of 163 position-based candidate genes to an average of 22 candidates based on position and expression or phenotype [4]. Though particularly well suited for syndromes in which the disease gene shows altered expression patterns in the affected tissues, it can also be applied to more complex diseases.

2.5. Genes to Disease (G2D)

G2D (**G**enes to **D**iseases) is a web resource for prioritizing genes as candidates for inherited diseases using a combination of data mining on biomedical databases and gene sequence analysis. 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 (Search Tool for the Retrieval of Interacting Proteins) database [9].

The G2D system scores all terms in GO (Gene Ontology) according to their relevance to each disease starting from MEDLINE queries featuring the name of the disease. This is done by relating symptoms to GO terms through chemical compounds, combining fuzzy binary relations between them previously inferred from the whole MEDLINE and RefSeq databases. Then, to identify candidate genes in a given a chromosomal region, G2D (Genes to Diseases) performs BLASTX (search protein databases using a translated nucleotide query) searches of the region against all the (GO annotated) ge-

nes in RefSeq. All hits in the region with an E-value $< 10e^{-10}$ are registered and sorted according to the GO-score of the RefSeq gene they hit (the average of the scores of their GO annotations) [10].

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 [9]. G2D is publicly available at http://coot.embl.de/g2d// . Additionally, it is possible to access from this server a database of precalculated results for more than 550 monogenic diseases on published linkage regions using the phenotype method.

In a test with 100 diseases chosen at random from OMIM (Online Mendelian Inheritance in Man), using bands of 30 Mb [the average size of linkage regions], G2D detected the disease gene in 87 cases. In 39 % of these it was among the best three candidates, and in 47 % among the best 8 candidates [9].

2.6. CAESAR

CAESAR (**C**andid**A**t**E** **S**earch **A**nd **R**ank) represents a novel selection strategy in that it combines text and data mining to associate genetic information with extracted trait knowledge in order to prioritize candidate genes. CAESAR exploits the knowledge of complex traits in literature by using ontologies to semantically map the trait information to gene and protein-centric information from several different public data sources, including tissue-specific gene expression, conserved protein domains, protein-protein interactions, metabolic pathways and the mutant phenotypes of homologous genes [5]. CAESAR uses four possible methods of integration to combine the results of data searches into a prioritized candidate gene list. In contrast to PROSPECTR, SUSPECTS, DGP and GENESEEKER, gene selection is not limited to one or more genomic regions, as all genes annotated in one of the databases are potential candidates.

CAESAR is comprised of three main steps: text mining, data mining and data integration. It requires a body of text (referred to as corpus) describing the biology of a trait as its only input. Recommended forms of input text include published trait review articles and trait OMIM records. First, genes mentioned in the input text are identified and ontology terms are ranked based on their similarity to an input text. Second, genes are rankd for each data source independently based on the relevance of the ontology terms with which they are annotated. Third, the individual gene lists are integrated to provide a single ranked list of candidate genes that combines evidence from all data sources [5].

CAESAR can be used to prioritize a smaller number of candidates within a region of linkage, or to prioritize among polymorphisms annotated with ranked genes that show significant association in a genome-wide study. However this method is particularly valuable for complex traits, which may be affected by a wider array of biological processes, some of which may not have been directly implicated by previous studies. CAESAR also reports the evidence supporting the prioritization rank of each gene, allowing an investigator to trace the line of reasoning and to exercise his or her own judgment as to its validity. Thus, it can be seen as a very sophisticated aid to prioritization [5]. Currently, CAESAR can only be accessed by downloading and running locally. Test data can be downloaded from http://visionlab.bio.unc.edu/caesar/ .

In a test of its effectivness, CAESAR successfully selected 7 out of 18 (39 %) complex human trait susceptibility genes within the top 2 % of ranked candidates genome-wide, a subset that represents roughly 1 % of genes in the human genome and provides sufficient enrichment for an association study of several hundred human genes [5].

3. Conclusion and future work

This short overview of six independent computational methods for identifying candidate disease genes was given together with references to available literature and web tools. As shown here, computational prediction of disease relevant genes must be regarded as an extremely hard problem, with probably no biomedical optimal solution attainable at all. No computational system can select candidate genes with certainty. More then ever, one cannot expect to predict these genes with high confidence by one single method. Instead, information about candidate genes gained by different independent methods has to be combined. Candidate genes selected by more methods with very diverse data inputs may carry more weight than a candidate genes selected only by using one single method.

The presented paper should be seen as a small step of our ongoing work, using computational methods to select a subset of the most likely candidate genes in cardiovascular disease for their next empirical validation.

Source		URL	Records	Content
Ontology				
MP	Mammalian phenotype ontology	www.informatics.jax.org	3 850	Phenotype
eVOC	eVOC anatomical ontology	www.evocontology.org/	394	Anatomy
GO bp	Gene ontology biological process	www.geneontology.org/	9 687	Function
GO mf	Gene ontology molecular function	www.geneontology.org/	7 055	Function
Database				
OMIM	Online Mendelian Inheritance in Man	www.ncbi.nih.gov/	16 564	Disease
Gene	Entrez Gene	www.ncbi.nih.gov/	32 859	Gene
Ensemble		www.ensembl.org/	20 134	Gene
SwissProt		www.ebi.ac.uk/uniprot/	13 434	Expression
TrEMBL	Nucleotide sequence database	www.ebi.ac.uk/uniprot/	57 551	Expression
InterPro	Protein domen database	www.ebi.ac.uk/intepro/	12 542	Domain
BIND	Biomolecular interaction	www.bind.ca/	35 661	Interaction
HPRD	Human protein reference database network database	www.hprd.org/	33 710	Interaction
KEGG	Kyoto encyclopedia of genes genomes pathway database	www.genome.jp/kegg/	209	Pathway
MGD	Mouse genome database	www.informatics.jax.org/	7 705	Phenotype
GAD	Genetic association database	http://hpcio.cit.nih.gov/gad.html	8 176	Association
GOA	Gene ontology annotation database	www.ebi.ac.uk/goa/	27 768	Function
RefSeq	Reference sequence	www.ncbi.nlm.nih.gov/RefSeq/	10 329	Gene
HGMD	Human gene mutation database	www.hgmd.cf.ac.uk/ac/index.php		Gene Mutation
GNF	Genomics Institute of the Novartis Research Foudation database	www.hgmd.cf.ac.uk/ac/index.php		Expression
MEDLINE		http://medline.cos.com/	10 752 796	References

Table 1: Information about data sources and ontologies used in methods.

References

[1] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteos, B. S. Pickard, "SUSPECTS: Enabling Fast and Effective Prioritization of Positional Candidates", *Bioinformatics*, vol. 22 (6), pp. 773–774, 2006.

[2] E. A. Adie, R. R. Adams, K. L. Evans , D. J. Porteos, B. S. Pickard, "Speeding Disease Gene Discovery by Sequence Based Candidate Prioritization", *BMC Bioinformatics*, vol. 6, pp. 55, 2005.

[3] M. Dean, "Approaches to Identify Genes for Complex Human Diseases: Lessons from Mendelian Disorders", *Hum. Mutat.*, vol. 22, pp. 261–274, 2003.

[4] M. A. van Driel, K. Cuelenaere, P. Kemmeren, J. Leunissen, H. Brunner, "A New Web-based Data Mining Tool for the Identification of Candidate Genes for Human Genetic Disorders", *European Journal of Human Genetics*, vol. 11, pp. 57–63, 2003.

[5] K. J. Gaulton, K. L. Mohlke, T. J. Vision, "Computational System to Select Candidate Genes for Complex Human Traits", *Bioinformatics*, vol. 23 (9), pp. 1132–1140, 2007.

[6] N. López-Bigas, Ch. A. Ouzonis, "Genome-wide Identification of Genes Likely to be Involved in Human Genetic Disease", *Nucleic Acids Research*, vol. 32, pp. 3108–3114, 2004.

[7] C. Newton-Cheh, J. Hirschhorn, "Genetic Association Studies of Complex Traits: Design and Analysis Issues", *Mutation Research*, vol. 573, pp. 54–69, 2005.

[8] L. Peltonen, V. McKusick, "Genomics and Medicine: Dissecting Human Disease in the Postgenomic Era", *Science*, vol. 291, pp. 1224–1229, 2001.

[9] C. Perez-Iratxeta, P. Bork, M. A. Andrade, "G2D: a Tool for Mining Genes Associated with Disease", *BMC Genetics*, vol. 6, pp. 45, 2005.

[10] C. Perez-Iratxeta, P. Bork, M. A. Andrade, "Update of the G2D Tool for Prioritization of Gene Candidates to Inherited Diseases", *Nucleic Acids Research*, vol. 35, pp. W1–W5, 2007.

[11] T. Strachan, A. Read, "Human Molecular Genetics 2", *BIOS Scientific Publishers Ltd*, 1999.