Bioinformatics methods for identifying candidate disease genes

Marc A. van Driel[1] *and Han G. Brunner*[2]*

[1] Molecular Biology Department, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands
[2] Department of Human Genetics, University Medical Centre Nijmegen, Geert Grooteplein 10, Nijmegen, The Netherlands
* *Correspondence to*: Tel: + 31 24 3614017; Fax: + 31 24 3668752; E-mail: H.Brunner@antrg.umcn.nl

Date received (in revised form): 28th December 2005

Abstract

With the explosion in genomic and functional genomics information, methods for disease gene identification are rapidly evolving. Databases are now essential to the process of selecting candidate disease genes. Combining positional information with disease characteristics and functional information is the usual strategy by which candidate disease genes are selected. Enrichment for candidate disease genes, however, depends on the skills of the operating researcher. Over the past few years, a number of bioinformatics methods that enrich for the most likely candidate disease genes have been developed. Such *in silico* prioritisation methods may further improve by completion of datasets, by development of standardised ontologies across databases and species and, ultimately, by the integration of different strategies.

Keywords: *bioinformatics, candidate disease gene prediction*

Introduction

Currently, with the increase in accessible data and the development of novel molecular biology techniques, new methods for the identification of disease genes are evolving. Linkage studies and mutation screening are becoming easier and the number of identified (disease) genes is increasing rapidly. 2003 saw the completion of the human genome sequence and the number of genes is now set to 20,000–25,000.[1,2] With all the genetics technology in place, identification of disease-related mutations in Mendelian single-gene disorders mainly depends on having the right patients and families. The genetic analysis of complex diseases still remains a difficult task, however, and most genes for multifactorial disease remain to be discovered.

Genetic mapping by linkage is a mainstay of human genetics research. While positional information reduces the number of genes that are candidates for causing the disease, this reduction is often not sufficient for rapid disease gene identification.

The aim of candidate gene prioritisation methods is to choose those genes for detailed mutation analysis that are most likely to be the cause of the disease. This is especially relevant since positional methods may leave up to 100 different genes as candidates. Hence additional information to be used for prioritisation is essential.

Databases have become a core source for today's gene hunters. Retrieval systems such as the National Center for Biotechnology Information's Entrez,[3] the Sequence Retrieval System[4] and Maarten's Retrieval System[5] provide easy and fast access to a collection of frequently used databases. The main focus of these retrieval systems is to fetch a set of database entries that meet the user query. Identification of a disease gene is most likely to be successful when positional and functional routes are integrated. Integration of data based on genomic context, such as in the University of California, Santa Cruz genome browser and Ensembl,[6,7] resulted in step by step interfaces (eg EnsMart[8]) which extract data based on chromosomal position, gene expression[9] and gene ontology (GO).[10] Enrichment for disease candidate genes using these database interfaces, however, depends heavily on the operation skills of the researcher. Alternative methods have been developed systematically to explore datasets for the most likely candidate disease genes. This paper presents an overview of such methods and their accessibility.

Candidate disease gene identification methods

The methods developed for candidate disease gene identification use different data sources and strategies.

Perez-Iratxeta *et al.* developed the Genes2Diseases (G2D) method.[11] This searches Medline abstracts for MeSH-C (phenotypic features) and MeSH-D (chemical objects)

terms.[12,13] Co-occurrence of MeSH-C/D in the Medline abstracts was considered to be related to the association between the chemical and the phenotypic terms. Sequences in the RefSeq database[14] are used to associate MeSH-D from the sequence Medline links with the GO functional annotation of the same sequence.[10] This creates the possibility of associating phenotype terms with functional terms via the chemical terms. Literature on a disease can be screened for MeSH-C terms, which are then used to determine the association between the disease and the genes with the GO annotation. The system was tested on 450 diseases that had previously been mapped to a specific locus but without a particular gene assigned. The resulting scores were compared with 100 diseases for which the gene was known.[11] On average, Perez-Iratxeta *et al.* tested 30-megabase candidate regions. Assuming 20,000–25,000 human genes,[1,2] and an average gene density of one gene per 120 kilobases, an 8–31-fold enrichment was calculated for this method. The G2D method was recently extended with expressed sequence tag data. The system for phenotype input was also improved, which reduces the prior clinical knowledge required to be entered. This new version of G2D performs better, mostly because more databases are used with larger datasets. Both versions of the method are available online (http://www.bork.embl-heidelberg.de/g2d/; http://www.ogic.ca/projects/g2d_2/).

Whereas G2D uses information extracted from the medical literature, POCUS, developed by Turner *et al.*, uses functional, domain annotation and gene expression data to prioritise candidate disease genes.[15] The method assumes that genes involved in the same disease will share GO annotation, protein domains and a similar expression pattern. A scoring system that includes these sources allows one to rank genes in the candidate disease regions. POCUS seeks over-representation of functional annotation between loci for the same disease. Larger candidate regions are *a priori* more likely to share similarities and are thus less likely to generate gene connections that are statistically significant. The method was tested with 29 diseases and achieved an enrichment of 12–42-fold. The method cannot be used online, but the POCUS scripts can be downloaded.

Specific gene characteristics have been used in candidate disease gene identification. Sequence analysis of human/eukaryotic genes showed that human proteins with multiple long amino acid runs are more often linked with genetic disease than are shorter proteins.[16] Lopez-Bigas *et al.* found that proteins involved in genetic diseases tend to be long, conserved and without close paralogues.[17] Disease genes are more frequently found to be conserved in other species, but this might be due the preferential sequencing of known (disease) genes. The disease gene prediction system using these sequence characteristics can be accessed online (http://cgg.ebi.ac.uk/services/dgp/).

Similarly, Adie *et al.* tested sequence property analysis using alternating decision trees.[18] They found differences between random genes and disease genes based on a number of features, including: gene/cDNA/protein/3′ untranslated region length, the number of exons, distance to the adjacent gene, higher level of conservation in the mouse, signal peptide encoding and 5′ CpG islands. Tests for candidate gene identification showed 2–25-fold enrichment. Data can be accessed using the PROSPECTR web server (http://www.genetics.med.ed.ac.uk/prospectr/). The user can rank genes for their likelihood to be involved in a disease, either from a list or a genomic region. The method was recently extended with GO terms, Interpro domains and gene expression data. The SUSPECTS web server uses PROSPECTR, and allows one to rank genes for their likelihood of involvement in the disease of interest (http://www.genetics.med.ed.ac.uk/suspects/). Smith *et al.* used a comparable analysis, which found similar differences between disease and non-disease genes. Using discriminant analysis, they showed that these differences may help to predict human disease genes;[19] however, their method is not accessible online.

It is possible systematically to interrogate the multitude of gene and protein expression data that are produced by methods such as RNA expression microarray analysis and SAGE. For example, Tiffin *et al.* developed a method which uses an anatomical ontology (eVOC)[9] to integrate biomedical literature and human gene expression data.[20] The method uses eVOC as a controlled vocabulary to find anatomy terms specific for the disease in PubMed abstracts. The anatomy terms are ranked and the candidate genes are selected using the highest ranked terms. The selected candidate genes have a gene expression pattern that matches the disease associated/affected tissues. The enrichment reached is 1.5–3.0-fold and the correct gene was found in more than 85 per cent of the cases. Data and scripts are available, but there is no web interface (http://www.sanbi.ac.za/tiffin_et_al/).

The link between the tissues and organs that are affected by a genetic disease and candidate gene expression profiles have been exploited.[21,22] GeneSeeker uses human as well as mouse expression and phenotypic data stored in various databases (http://www.cmbi.ru.nl/GeneSeeker/). This information is combined with positional data for the genes from both species. The system uses different online databases rather than local data and thus mines in real-time. The GeneSeeker approach differs from the other candidate prioritisation approaches by utilising cross-species data. Knowledge of model organisms makes comparative candidate gene selection possible. This situation applies when a gene is known to cause a similar phenotype in another species. Nonetheless, a direct comparison between the phenotypes in humans and model organisms can be complicated because of differences in anatomy. Transfer of knowledge by phenotype is most straightforward in other mammalian species, such as the mouse, that are evolutionarily close to humans. An example is the disease gene identification in ectrodactyly-ectodermal dysplasia-clefting syndrome. This human disease gene was identified by a comparable phenotype in homozygous null mice.[23] A 7–25-fold enrichment of

candidate disease genes was achieved using GeneSeeker on a test set of ten diseases.[21] Recently, Bradford *et al.* presented a cross-species search system.[24] The human–mouse gene searcher enables the user to search with phenotype data from human and mouse and links this to the Mouse Gene Expression Database.[25] This tool can assist in the human–mouse phenotype mapping process. It has its own merits, and can also be implemented in GeneSeeker.

A number of groups have started to use clinical overlap between genetic diseases to cluster phenotypes, thereby allowing correlations with the functional classification of their underlying disease genes.[26] Such phenotype relationships might be a powerful method for function prediction.[27–29] The human phenotype collection and the underlying gene–phenotype relations can therefore be used as a tool for functional genomics.[30]

Freudenberg and Propping developed a method for clustering genetic diseases based on their phenotype similarity.[31] They manually attributed the disease phenotypic manifestations using the Online Mendelian Inheritance in Man (OMIM) database.[32] A similarity measure was developed to compare the phenotypes and to perform a complete linkage clustering. The approach was tested with a leave-one-out cross-validation of 878 diseases from the OMIM database, using 10672 candidate genes from the human genome. They achieved an enrichment of 7–33-fold. Unfortunately, the method is not available for other users.

Similarly, Cantor *et al.* clustered OMIM[32] records based on the clinical synopsis section.[33] They reduced the disease characteristics to 50 categories. In a test of two diseases, they found relationships at the genotype level. Since the authors only intended to establish proof of principle on using OMIM for phenotype clustering, they did not systematically analyse phenotype–genotype relationships, and their system cannot be accessed directly.

Masseroli *et al.* developed the GFINDer system. This web tool allows one to mine the annotation information from several databases for a given set of sequence identifiers.[34] Filter parameters are set manually in the system to select disease genes, and statistical analysis can be performed. Recently, the clinical synopsis of OMIM was integrated into the GFINDer system (http://www.bioinformatics.polimi.it/GFINDer/). Phenotype data were normalised and structured in order to obtain two controlled vocabularies suitable for computational purposes. The absence of a predefined strategy makes the efficiency of the system heavily dependent on the operating researcher. The authors presented only a few selected examples of their method, which makes it difficult to estimate the enrichment.[34]

van Driel *et al.* devised a method for comparing phenotypes derived from the OMIM database that uses a textual similarity measure by an automated full text-analysis technique, rather than predefined term classes, and analysed the phenotype–genotype relationships.[35] They found that phenotype similarity scores, which are based on automatic quantification, correlate positively with a number of measures of gene function, including protein sequence, similarity of shared protein motifs, functional annotation and direct protein–protein interaction. The data support the idea that phenotypic relationships may be used as indicators of biological and functional interactions at the gene and protein levels. The phenotype–phenotype ranking scores can be searched online (http://www.cmbi.ru.nl/MimMiner/). The method can be used to study the phenotypic relationships at the genotype level, by which the phenotype becomes a tool for functional genomics.[30] The aim was not to enrich a specific region for candidate genes, but the data can be used for this purpose.

Future: Integration and standardisation

The various methods for identifying my candidate disease genes in humans cover different concepts. They use functional and literature data, gene-specific characteristics, anatomy-based gene/protein expression data or phenotype comparison analyses. In light of the comparable enrichment levels achieved with the different methods, it is likely that they can complement each other.

The results discussed here suggest that the phenotype is a powerful source for revealing biological function and that special attention is needed for the standardisation of the description of phenotypes.[30] Various approaches to a more systematic description of phenotype data have been proposed and await further development.[36,37] Essential to the improvement of the candidate disease gene identification methods will be the establishing of standard vocabularies that can be used across databases and species. A further challenge will be to develop, refine and integrate these methods into a system that aids in elucidation and understanding of the mechanisms of (complex) disease.

References

1. International Human Genome Sequencing Consortium (2004), 'Finishing the euchromatic sequence of the human genome', *Nature* Vol. 431, pp. 931–945.
2. Larsson, T.P., Murray, C.G., Hill, T. *et al.* (2005), 'Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery', *FEBS Lett.* Vol. 579, pp. 690–698.
3. Schuler, G.D., Epstein, J.A., Ohkawa, H. *et al.* (1996), 'Entrez: Molecular biology database and retrieval system', *Methods Enzymol.* Vol. 266, pp. 141–162.
4. Etzold, T. and Argos, P. (1993), 'SRS — An indexing and retrieval tool for flat file data libraries', *Comput. Appl. Biosci.* Vol. 9, pp. 49–57.
5. Hekkelman, M.L. and Vriend, G. (2005), 'MRS: A fast and compact retrieval system for biological data', *Nucleic Acids Res.* Vol. 33, pp. W766–W769.
6. Kent, W.J., Sugnet, C.W., Furey, T.S. *et al.* (2002), 'The human genome browser at UCSC', *Genome Res.* Vol. 12, pp. 996–1006.

7. Hubbard, T., Andrews, D., Caccamo, M. *et al.* (2005), 'Ensembl 2005', *Nucleic Acids Res.* Vol. 33, pp. D447–D453.
8. Kasprzyk, A., Keefe, D., Smedley, D. *et al.* (2004), 'EnsMart: A generic system for fast and flexible access to biological data', *Genome Res.* Vol. 14, pp. 160–169.
9. Kelso, J., Visagie, J., Theiler, G. *et al.* (2003), 'eVOC: A controlled vocabulary for unifying gene expression data', *Genome Res.* Vol. 13, pp. 1222–1230.
10. Ashburner, M., Ball, C.A., Blake, J.A. *et al.* (2000), 'Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium', *Nat. Genet.* Vol. 25, pp. 25–29.
11. Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2002), 'Association of genes to genetically inherited diseases using data mining', *Nat. Genet.* Vol. 31, pp. 316–319.
12. Wheeler, D.L., Barrett, T., Benson, D.A. *et al.* (2005), 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res.* Vol. 33, pp. D39–D45.
13. Lipscomb, C.E. (2000), 'Medical Subject Headings (MeSH)', *Bull. Med. Libr. Assoc.* Vol. 88, pp. 265–266.
14. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005), 'NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res.* Vol. 33, pp. D501–D504.
15. Turner, F.S., Clutterbuck, D.R. and Semple, C.A. (2003), 'POCUS: Mining genomic sequence annotation to predict disease genes', *Genome Biol.* Vol. 4, p. R75.
16. Karlin, S., Brocchieri, L. and Bergman, A. (2002), 'Amino acid runs in eukaryotic proteomes and disease associations', *Proc. Natl. Acad. Sci. USA* Vol. 99, pp. 333–338.
17. Lopez-Bigas, N. and Ouzounis, C.A. (2004), 'Genome-wide identification of genes likely to be involved in human genetic disease', *Nucleic Acids Res.* Vol. 32, pp. 3108–3114.
18. Adie, E.A., Adams, R.R. and Evans, K.L. (2005), 'Speeding disease gene discovery by sequence based candidate prioritization', *BMC Bioinformatics* Vol. 6, p. 55.
19. Smith, N.G. and Eyre-Walker, A. (2003), 'Human disease genes: Patterns and predictions', *Gene* Vol. 318, pp. 169–175.
20. Tiffin, N., Kelso, J.F., Powell, A.R. *et al.* (2005), 'Integration of text and data-mining using ontologies successfully selects disease gene candidates', *Nucleic Acids Res.* Vol. 33, pp. 1544–1552.
21. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P. *et al.* (2003), 'A new web-based data mining tool for the identification of candidate genes for human genetic disorders', *Eur. J. Hum. Genet.* Vol. 11, pp. 57–63.
22. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P. *et al.* (2005), 'GeneSeeker: Extraction and integration of human disease-related information from web-based genetic databases', *Nucleic Acids Res.* Vol. 33, pp. W758–W761.
23. Celli, J., Duijf, P., Hamel, B.C. *et al.* (1999), 'Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome', *Cell* Vol. 99, pp. 143–153.
24. Bradford, I., Winter, R., Evans, C. *et al.* (2005), 'Human-mouse gene searcher: A tool to assist discovery of malformation-associated genes by using phenotype databases', *Bioinformatics* Vol. 21, pp. 408–409.
25. Ringwald, M., Eppig, J.T., Begley, D.A. *et al.* (2001), 'The Mouse Gene Expression Database (GXD)', *Nucleic Acids Res.* Vol. 29, pp. 98–101.
26. Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001), 'Human disease genes', *Nature* Vol. 409, pp. 853–855.
27. Spranger, J. (1985), 'Pattern recognition in bone dysplasias', *Prog. Clin. Biol. Res.* Vol. 1200, pp. 315–342.
28. Annunen, S., Korkko, J., Czarny, M. *et al.* (1999), 'Splicing mutations of 54-bp exons in the COL11A1 gene cause Marshall syndrome, but other mutations cause overlapping Marshall/Stickler phenotypes', *Am. J. Hum. Genet.* Vol. 65, pp. 974–983.
29. van Steensel, M.A., Buma, P., de Waal Malefijt, M.C. *et al.* (1997), 'Oto- spondylo-megaepiphyseal dysplasia (OSMED): Clinical description of three patients homozygous for a missense mutation in the COL11A2 gene', *Am. J. Med. Genet.* Vol. 70, pp. 315–323.
30. Brunner, H.G. and van Driel, M.A. (2004), 'From syndrome families to functional genomics', *Nat. Rev. Genet.* Vol. 5, pp. 545–551.
31. Freudenberg, J. and Propping, P. (2002), 'A similarity-based method for genome-wide prediction of disease-relevant human genes', *Bioinformatics* Vol. 18, pp. S110–S115.
32. Hamosh, A., Scott, A.F., Amberger, J. *et al.* (2002), 'Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res.* Vol. 30, pp. 52–55.
33. Cantor, M.N. and Lussier, Y.A. (2004), 'Mining OMIM for insight into complex diseases', *Medinfo* Vol. 11, pp. 753–757.
34. Masseroli, M., Galati, O. and Pinciroli, F. (2005), 'GFINDer: Genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists', *Nucleic Acids Res.* Vol. 33, pp. W717–W723.
35. van Driel, M.A., Bruggeman, J., Vriend, G. *et al.* (2006), 'A text-mining analysis of the human phenome', *Eur. J. Hum. Genet.* Vol. 14, pp. 535–542.
36. Freimer, N. and Sabatti, C. (2003), 'The human phenome project', *Nat. Genet.* Vol. 34, pp. 15–21.
37. Biesecker, L.G. (2005), 'Mapping phenotypes to language: A proposal to organize and standardize the clinical descriptions of malformations', *Clin. Genet.* Vol. 68, pp. 320–326.