Association of genes to genetically inherited diseases using data mining

Carolina Perez-Iratxeta, Peer Bork & Miguel A. Andrade

Published online: 13 May 2002, DOI: 10.1038/ng895

Although approximately one-quarter of the roughly 4,000 genetically inherited diseases currently recorded in respective databases (LocusLink[1], OMIM[2]) are already linked to a region of the human genome, about 450 have no known associated gene. Finding disease-related genes requires laborious examination of hundreds of possible candidate genes (sometimes, these are not even annotated; see, for example, refs 3,4). The public availability of the human genome[5] draft sequence has fostered new strategies to map molecular functional features of gene products to complex phenotypic descriptions, such as those of genetically inherited diseases. Owing to recent progress in the systematic annotation of genes using controlled vocabularies[6], we have developed a scoring system for the possible functional relationships of human genes to 455 genetically inherited diseases that have been mapped to chromosomal regions without assignment of a particular gene. In a benchmark of the system with 100 known disease-associated genes, the disease-associated gene was among the 8 best-scoring genes with a 25% chance, and among the best 30 genes with a 50% chance, showing that there is a relationship between the score of a gene and its likelihood of being associated with a particular disease. The scoring also indicates that for some diseases, the chance of identifying the underlying gene is higher.

To support and rationalize the manual association of known or inferred functional features of genes to the phenotypic features of a disorder, we have developed a data-mining system, based on fuzzy set theory[7], which makes inferences using information from biological and medical literature. We have applied the system to the prioritization of candidate genes for 455 genetically inherited diseases for which no underlying gene has yet been assigned.

The first phase of the data-mining process (see Methods for a more complete description) involves combining the information from MEDLINE and a protein sequence database to derive relationships between pathological conditions and terms describing protein function. We used a three-step procedure. (i) We computed the associations between pathological conditions and chemical terms using MEDLINE, a database of indexed journal citations and abstracts of the biomedical literature, which currently contains more than 11 million entries (Fig. 1). We consider the relationship between associated terms as strong if they occur together in many abstracts. (ii) We calculated the relationships between chemical terms and terms describing protein function. We used the NCBI RefSeq database[1], which contains more than 10,000 genes whose function is annotated with terms from a controlled functional vocabulary (Gene Ontology, GO[6]; Fig. 1). Experimental evidence is provided for each protein-function annotation by a pointer to MEDLINE. We consider that an annotated gene relates its functional terms to the chemical terms found in the linked bibliography. (iii) We combined the associations of

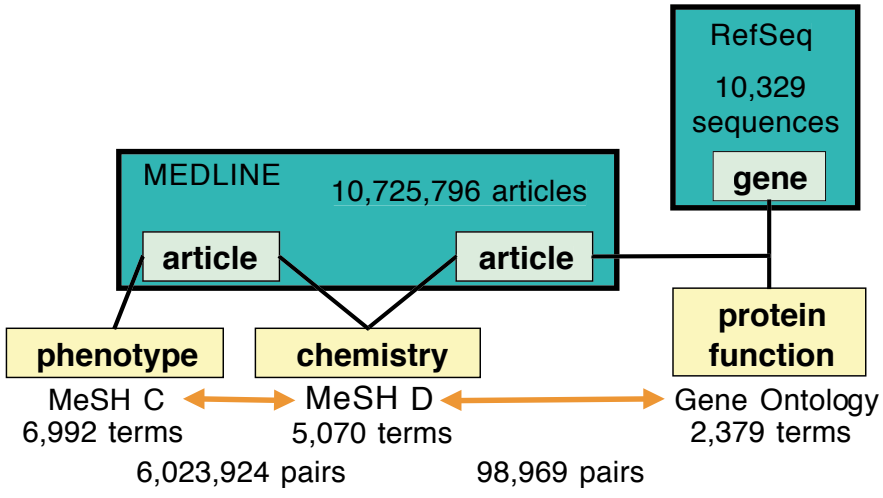


Fig. 1 Components used for deriving associations between phenotypic features and gene functions. Green boxes represent databases, yellow boxes indicate the associated concepts and red arrows represent the associations used. The MEDLINE version was obtained in February 2001 from the US National Library of Medicine. It contains 10,725,796 references. Many of these references were annotated with the controlled set of MeSH C and MeSH D terms. Of these references, 1,380,733 contained at least one MeSH C term and one MeSH D term. The derived links from MeSH C to MeSH D terms gave us phenomenological relations through the biological knowledge deposited in MEDLINE. The set of 10,329 RefSeq sequences[1] was annotated with a variety of 2,379 GO terms. Each annotation with a GO term is linked to a MEDLINE entry. This allowed us to produce links between GO terms to MeSH D terms, indicating phenomenological relationships between chemical entities and protein functionality.

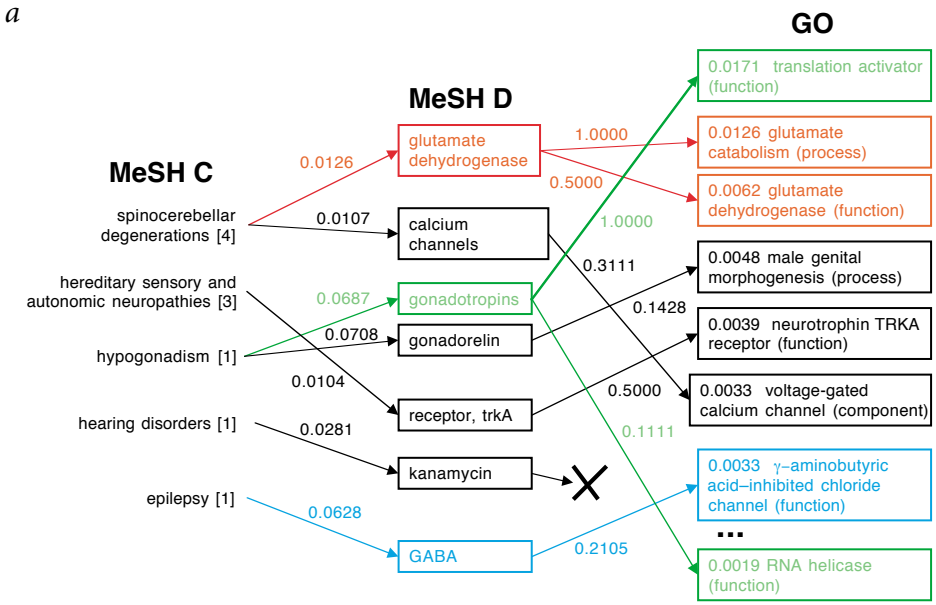
European Molecular Biology Laboratory, Meyerhofstr.1, Heidelberg 69012, Germany and Max Delbruck Center for Molecular Medicine, Department of Bioinformatics, PO Box 740238, D-13092, Berlin-Buch, Germany. Correspondence should be addressed to P.B. (e-mail: bork@embl-heidelberg.de).

functional terms to chemical terms with the previously established associations of pathological conditions to chemical terms, to derive the aforementioned relations between pathological conditions and protein-function terms (Fig. 1).

Next, for each of the 455 diseases with chromosomal mapping information, protein-function terms are associated by combining the medical terms found in the literature regarding the disease with the set of previously computed relationships. We then score the strength of the relation of the RefSeq sequences to the disease according to their functional annotation (hereafter referred to as GO score; see Methods).

Finally, we prioritize the candidates for a given mapped disease by carrying out a sequence comparison between the respective region (on average 30 Mb) and the set of scored RefSeq sequences. The hits in the region are then sorted according to the GO score of the RefSeq homologous sequence. This means that candidates in the respective region may be associated with a disease phenotype through a homologous sequence in the RefSeq set. Owing to the heterogeneity of the literature and gene annotation, the maximum GO score observed in the RefSeq set changes from disease to disease. To have a measure independent of this effect, we used a score relative to the distribution of GO scores (*R* score). For a sequence with a given GO score, its *R* score is computed as the fraction of sequences in the RefSeq set having a higher GO score; sequences scoring well have an R score of zero or close to zero, irrespective of the disease considered.

To test the performance of our system, we analyzed 100 genes for which disease-causing mutations had already been reported (see Web Fig. A online). The disease-related gene was identified in 55 cases (being on average among the best-scoring 3% of genes). The computations were done automatically excluding the papers related to the disease, to ensure that we would not find the relation directly from the papers that described the connection of the gene to the corresponding disease (see Methods).



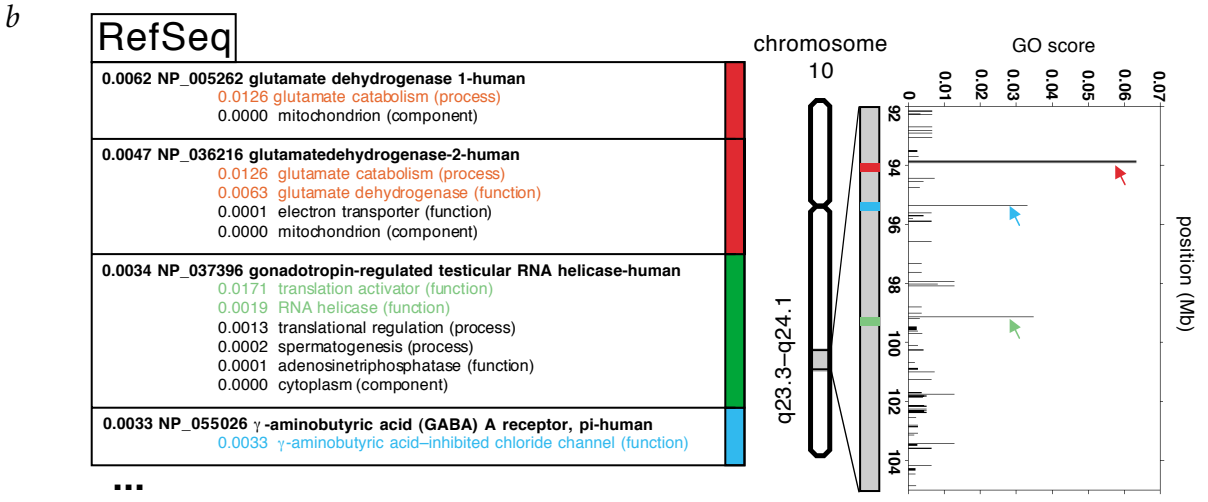


Fig. 2 Example of the analysis of 'spinocerebellar ataxia-8, infantile, with sensory neuropathy'. The red/blue/green colored boxes indicate different protein functions. ***a***, Spinocerebellar ataxia-8 was linked to a number of medical terms (left column, MeSH C terms; the number in brackets indicates the number of papers in which the terms were present). Using the previously derived relations, we could evaluate the more-related functional terms (right column, GO terms). For example, the strength of the association of the pair (hypogonadism, translation activator) is defined by a path through the MeSH D term 'gonadotropins', which gives a value of 0.0687 × 1.0000 = 0.0687 (see Methods). Taking into account that the MeSH C term 'hypogonadism' was associated with one of four papers describing the disease, the corresponding value for the weighted association $\mu_{T_2}{}^{w}$ (hypogonadism, translation activator) is 0.0687 × 1/4 = 0.0171. As this value is the maximum $\mu_{T_2}{}^{w}$ pointing to 'translation activator' from any MeSH C, the corresponding GO term 'translation activator' receives a score $\mu_{T_2}{}^{w}$ (translation activator) of 0.0171. ***b***, Regions of homology to the 10,000 RefSeq sequences were searched on a region of chromosome 10 (positions 92–105 Mb) where the disease had been mapped. The four sequences that best match the disease, according to their annotations (GO terms, indicated as in *a*) and for which homology was found, are shown on the left. The first two sequences are homologous and therefore point to the same region. The graph on the right indicates the GO score of the matching sequence versus the position in the hit in the region. The color (red/green/blue) to the right of the sequence annotation box indicates the region of the match. The three hits pointed in the graph by the arrows stand well above the rest: one of them is likely to correspond to the gene underlying this disease.

Diseases for which genes have been associated in the past tend to be well characterized—that is, there is more information retrievable from the literature for those diseases than for those more recently discovered. To determine the extent of this effect, we tested the system with diseases for whom underlying genes were identified during the years 2000 and 2001 (a total of 27), using a version of MEDLINE without literature corresponding to that period (see Web Note A online). In 7 of the 27 cases, the newly identified genes were not located in the corresponding chromosomal region where the disease had been mapped (as given by LocusLink), and could thus not be detected. This was probably a result of the incompleteness of the current draft of the human genome. In another 10 cases, the disease-related gene gave no hit to a RefSeq sequence with GO annotation revealing any significant association to the disease. Of the remaining 10 cases, the correct gene was among the best-scoring 5% of genes in four cases, and among the top 15% in another five. As might be expected, the analysis of more recently resolved diseases is more difficult. This effect was probably influencing the performance of the analysis of the set of 455 diseases, but can be easily detected using our scoring system.

As there is a correlation between the *R* score of a candidate gene and its likelihood to be the target gene, we can create a list of candidates for each disease to have a high chance of identifying the underlying gene. The benchmark indicates that, for example, there is a 25% chance of finding the target gene within the candidates with an *R* score below 0.01, or a 50% chance for candidates with an *R* score below 0.05. For the 455 unknown cases, this corresponds to an average of 8 and 30 genes, respectively (see Web Note A online).

Beyond the intrinsic limitations of the system described above, there are complications derived from the application of functional assignment by sequence similarity: (i) domain problems, (ii) presence of paralogous genes, (iii) identity thresholds and (iv) the present status of the human draft sequence (see Web Note A online).

Despite its limitations, the system should be useful for assigning priorities to candidate genes. The selection of candidates depends on both the fuzzy associations and the sequence-similarity analysis. This implies that the system's performance is of a varying nature. The derived relations are often obvious, the most straightforward being based on a strong similarity between the candidate gene and another gene known to produce a slightly different variant of the disease. In some cases, less apparent associations can be verified easily by examining the literature from which the association originated. In other cases, however, the relationship is hidden in a small number of papers and could be overlooked without the explicit suggestion of the system. Moreover, the system detects even weak associations that may, for example, only be based on the similarity of a domain within the protein. In such a case, a background of worse associations of other genes in the region (as indicated by their *R* scores) might call for further exploration of these weak associations. Finally, the system can overcome some of the problems caused by current gene prediction schemes that seem to be too conservative and tend to overlook genes[8]. For details and examples, see Web Note A online.

The use of this system can be demonstrated with one of the 455 cases analyzed, spinocerebellar ataxia-8 (LocusLink id 3648). The medical term most frequently associated with papers about this disease is 'spinocerebellar degenerations' (Fig. 2*a*). The related protein-function terms refer to the neurotransmitter glutamate, because levels of glutamate dehydrogenase (Gdh), an enzyme central to glutamate metabolism, are significantly reduced in individuals with neurological disorders affecting the cerebellum and its connections[9]. Accordingly, the top candidate gene is human glutamate dehydrogenase 1 (*GLUD1*), located in this region (Fig. 2*b*).

The data-mining system described here is highly dependent on the information associated with both genes and diseases, and relies on homology—that is, it is impossible to estimate the exact accuracy of an individual prediction. Thus, the prioritization given by the system requires a manual inspection of the context of both the disease and the candidate genes. For that purpose, the detailed data regarding the analysis of each disease, including the criteria for selection of candidates, can be examined using a public web-based server (see Methods). We expect that the performance of the system will improve in future regular updates as a result of the advent of additional literature, better gene and protein annotation, enhanced releases of the human genome sequence and a widespread production of standardized expression data.

Our approach evaluates and scores multiple associations between gene functions and monogenic disease phenotypes. However, the system could be adapted to detect other relationships between phenotype and genotype. For example, it could also propose the association of multiple genes to a disease or to other phenotypes, such as longevity, from a whole genome. Even if hundreds of candidates are proposed, recent developments in large-scale gene screening techniques (see, for example, ref. 10) will make their analysis feasible.

Methods

Inference system based on fuzzy relations. To identify and score relationships between terms, we used an approach from fuzzy set theory[7]. In this application, three different sets of items are related: the Medical Subject Headings (MeSH) terms of the 'Diseases' category (*C*), the MeSH terms of the 'Chemical & Drugs' category (*D*) and the set of 'Gene Ontology' terms (*GO*). A subset of MeSH terms were excluded because they were noninformative (see Web Note A online). Two different fuzzy binary relations were defined in $C \times D$ and $D \times GO$, referred as $R(C,D)$, and $S(D,GO)$, respectively. They were used to measure the 'degree of association' between medical and chemicals terms, $R(C,D)$, and between chemicals and protein-function terms, $S(D,GO)$. We assume that two terms are highly related in some context if they appear frequently together. Accordingly, the strength of term association is estimated by counting the co-occurrences of both items in the same 'transaction'. The value of the membership function for the (x,y) pair in the fuzzy binary relation $Q(X,Y)$ is

$$\mu_Q(x, y) = \frac{|X \cap Y|}{|X \cup Y|}$$

where $X = C$ and $Y = D$, or $X = D$ and $Y = GO$, and $|.|$ denotes set cardinality. In the case of $R(C,D)$, the set of transactions considered was the MEDLINE subset of abstracts annotated with both C and D MeSH terms. For the computation of $S(D,GO)$, each LocusLink entry was considered as a transaction relating its GO annotation terms with the MeSH D terms indexed on the MEDLINE abstracts linked to that entry. Next, to obtain the associations between C and GO terms, we define another fuzzy binary relation in $C \times GO$, $T(C,GO)$, given by the 'max-product composition' of $R(C,D)$ and $S(D,GO)$, with membership function:

$$\mu_T(c, go) = \max_{d \in D} \{\mu_R(c, d) \cdot \mu_S(d, go);\ c \in C,\ go \in GO\}$$

for every (c,go) in $C \times GO$. $T(C,GO)$ models the association between symptoms or manifestations of diseases with protein-function terms.

Gene scoring based on Gene Ontology annotation. We consider the set of proteins from RefSeq[1] annotated with GO terms. Given a particular disease, a score for each gene, based on its GO terms, is computed (GO score). The set of abstracts *A* dealing with the disease is considered, and

the corresponding subset of MeSH C terms (C_A) is extracted. If one manifestation is characteristic of the disease, it will occur more often in A than a casual symptom. Thus, the frequency of occurrence in A of each MeSH C term is considered, and the obtained weighting vector is incorporated into the model through its semi-scalar product by $\mu_T(c,go)$, resulting in a weighted fuzzy relation $T^W(C_A,GO)$. To assign a score to each GO term, we compute the second projection of $T^W(C_A,GO)$, which is a fuzzy set in GO whose membership function is given by:

$$\mu_{T_2^W}(go) = \max_{c \in C}\{\ \mu_{T^W}(c, go); go \in GO\}.$$

The value of this score is associated with each single GO term, and this is used to score and rank the whole set of RefSeq genes that receive a score equal to the average of their GO term scores (GO score).

Homology searches in the chromosomal region. We extracted the cytogenetic map location from LocusLink[1]. We obtained the corresponding base coordinate positions from the Golden Path[5] human genome assembly. In the benchmarks, a region of 30 Mb was taken around the disease-related gene. For the 455 non-associated diseases, regions smaller than 10 Mb were expanded to this size to compensate for possible marker dislocations (see Web Fig. B online). The RefSeq set is ordered as a hash list by the GO scores. The best-scoring 30% of sequences are compared with the chromosomal region (masked for low-complexity regions) using TBLASTN[11].

Prioritizing the candidates. The sequences identified in that region under an E-value threshold of 10e–10 are prioritized according to the GO score of their corresponding homologous sequence in RefSeq. To make the scoring given to the analysis of different diseases comparable, we introduced the R score, which accounts for the fraction of GO annotations of the RefSeq set giving a better GO score.

Automating queries in MEDLINE. The starting point for the analysis of a disease is selecting a set of papers to obtain MeSH C terms that describe the phenotype of the disease. To set up an automated protocol, we simply used the name of the disease for a query in MEDLINE. If less than five abstracts were selected by the query, the query was simplified step-wise until a minimum of five abstracts was obtained. We used the [tw] flag after each single word to carry out a search only in 'text words' (title, abstract, and MeSH terms). Synonym tables created by MeSH developers apply to this search.

URL. Detailed data regarding the analysis of each disease can be examined using a public web-based server (see http://www.bork.embl-heidelberg.de/g2d/).

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank Y.P. Yuan, J. Reina, D. Torrents, M. Suyama and other members of our group for helpful discussions. We are grateful to the US National Library of Medicine for kind licensing of MEDLINE, to NLM annotators for their extensive work in annotating MEDLINE papers with MeSH terms, and to the developers of RefSeq, LocusLink and Gene Ontology.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 28 December 2001; accepted 22 April 2002.

1. Pruit, K.D. & Maglott, D.R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
2. Hamosh, A., Scott, A.F., Amberger, J., Valle, D. & McKusick, V.A. Online mendelian inheritance in man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).
3. Garcia, C.K. *et al*. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science* **292**, 1394–1398 (2001).
4. Zhou, B., Westaway, S.K., Levinson, B., Johnson, M.A., Gitschier, J. & Hayflick, S.J. A novel pantohenate kinase gene (*PANK2*) is defective in Hallervorden-Spatz syndrome. *Nature Genet.* **28**, 345–349 (2001).
5. Lander, E.S. *et al*. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
6. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
7. Zimmermann, H.J. *Fuzzy Set Theory and its Applications* 3rd edn (Kluwer Academics, Boston, 1996).
8. Hogenesch, J.B. *et al*. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**, 413–415 (2001).
9. Plaitakis, A., Flessas, P., Natsiou, A.B. & Shashidharan, P. Glutamate dehydrogenase deficiency in cerebellar degenerations: clinical, biochemical and molecular genetic aspects. *Can. J. Neurol. Sci.* **20**, S109–S116 (1993).
10. Cargill, M. *et al*. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
11. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).