BIOINFORMATICS

Vol. 18 Suppl. 2 2002
Pages S110–S115

A similarity-based method for genome-wide prediction of disease-relevant human genes

J. Freudenberg and P. Propping

Institute of Human Genetics, Bonn University Hospital, Wilhelmstr. 31, D-53111, Bonn, Germany

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

Motivation: A method for prediction of disease relevant human genes from the phenotypic appearance of a query disease is presented. Diseases of known genetic origin are clustered according to their phenotypic similarity. Each cluster entry consists of a disease and its underlying disease gene. Potential disease genes from the human genome are scored by their functional similarity to known disease genes in these clusters, which are phenotypically similar to the query disease.

Results: For assessment of the approach, a leave-one-out cross-validation of 878 diseases from the OMIM database, using 10672 candidate genes from the human genome, is performed. Depending on the applied parameters, in roughly one-third of cases the true solution is contained within the top scoring 3% of predictions and in two-third of cases the true solution is contained within the top scoring 15% of predictions.

The prediction results can either be used to identify target genes, when searching for a mutation in monogenic diseases or for selection of loci in genotyping experiments in genetically complex diseases.

Contact: jan.freudenberg@uni-bonn.de

INTRODUCTION

Identification of genes, whose products play a role in monogenic or genetically complex diseases, is a major aim in the analysis of the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). In the genetic analysis of monogenic diseases, mutations cosegregating with the disease phenotype are searched directly in potentially disease relevant genes. In the genetic investigation of complex diseases, associated single nucleotide polymorphisms (SNPs) are searched previous to actually disease causing mutations (Risch and Merikangas , 1996). When testing for mutations or genetic variations underlying human disease, investigations of both monogenic and complex diseases rely on potentially disease relevant genes, termed in the following as *candidate genes*. The term *positional candidate* is used for genes, which are located in a genomic region, that is considered suspicious by linkage analysis studies (Lander and Kruglyak, 1995). The term *functional candidate* is used for genes, which are assumed because of the molecular role of the gene product.

In this paper, a similarity-based algorithm for functional scoring of candidate genes, which may be relevant for an arbitrary query disease, is introduced. The algorithm starts from the assumption, that phenotypically similar diseases are caused by similar molecular mechanisms. Phenotypically similar diseases are defined by their similar clinical features. For a query disease of unknown genetic cause, clusters of phenotypically similar diseases with known underlying disease genes are computationally identified. Genes of similar function to the respective known disease genes are then suggested as candidate genes for the query disease.

The predictions can be used to home in to the more presumable candidate genes, when searching for a mutation in a monogenic disease. In genetically complex diseases, high scoring candidate genes may be selected in SNP-genotyping experiments and consecutive relative risk scoring.

To our knowledge, only one other purely computational method for prediction of disease relevant genes from the clinical features of a disease has been published elsewhere (Perez-Iratxeta *et al.*, 2002). This other method uses text-mining for linking disease phenotypes with underlying molecular mechanisms.

MATERIALS AND METHODS

Computing phenotypic similarity between human diseases

The OMIM-morbid map lists all diseases contained in the OMIM-database (Hamosh *et al.*, 2000). These diseases are attributed manually according to their phenotypic appearance, using the indices ‘periodicity’, ‘etiology’, ‘tissue’, ‘age of onset’ and ‘mode of inheritance’. Of course attributes are not distinct for each disease entry and

Table 1. Diseases from the OMIM database are indexed according to their episodic occurrence, primary etiology, primary tissue, mode of inheritance and age of onset. Here a set of four example entries is shown

Index Disease	**episodic**	**etiology**	**tissue**	**onset**	**inheritance**
Epilepsy, nocturnal frontal lobe (600513)	*yes*	*regulatory, metabolic*	*cns*	*to puberty*	*autosomal dominant*
Colorectal cancer (16806)	*no*	*neoplastic*	*gastro-intestinal*	*late adult*	*autosomal dominant*
Argininemia (107800)	*no*	*metabolic*	*cns, liver*	*first year*	*autosomal recessive*
Duchenne muscular dystrophy (310200)	*no*	*degenerative*	*muscles*	*to puberty*	*x-chromosom*

occasionally arbitrary decisions have to be made. A set of examples of indexed diseases are shown in Table 1.

The index 'episodic' of a disease is a boolean variable, indicating an episodic occurrence of a disease in contrast to a linear progression. Typical examples of episodic disease are bipolar affective disorder or epilepsy.

The index 'etiology' is based on clinical signs and laboratory or pathological findings of a disease. The attribute list includes the terms *inflammatory*, *neoplastic*, *degenerative*, *regulatory* or *metabolic*. In cases where no clear distinction can be made, multiple attributes are allowed.

The index 'tissue' is compiled as a subset of the terms: *central nervous system*, *peripheral nervous system*, *eye*, *lens*, *cornea*, *ear*, *heart*, *lung*, *kidney*, *gastro-intestinal*, *liver*, *bone-marrow derived cells*, *endocrine tissue*, *connective tissue*, *muscle*, *skin* and *bone*. For each disease instance tissues are listed, where disease signs are primarily recognized, due to tissue damage or impairment of organ function. Secondary changes due to the course of a disease are ignored as much as possible. If necessary, multiple attributes are allowed here as well.

The mode of 'inheritance' indicates, whether a disease is inherited in *autosomal-dominant*, *autosomal-recessive*, *x-chromosomal*, *mitochondrial* or *complex* manner.

The age of 'onset' of a disease refers to the age, when symptoms are generally first noticed. Here the attributes *in utero*, *first year of live*, *up to puberty*, *early adulthood (up to 50 yrs)* and *late adulthood (above 50 yrs)* are applied. Cases, where no clear mode of inheritance or age of onset is known, are left without these attributes.

The most stringent definition of similarity requires identical attributes for each index. However, this would separate many diseases, which are phenotypically similar to a high degree. For example, clinical subtypes of a disease rely on different phenotypes, consecutively resulting in non-identical indexing. Therefore a more relaxed definition of similarity between two diseases $D1$ and $D2$ is invented, which can be expressed by the following similarity function:

$$\mathrm{sim}(D1,D2) := \sum w_i \cdot sim(D1.index_i, D2.index_i)$$

with the weights w_i specifying the contribution of a single index to the total similarity score. The weights are adjusted, that the similarity between two diseases scales to the interval [0..1[. The following similarity functions are applied for comparing the individual indices:

$$\mathrm{sim(D1.episodic,D2.episodic)} := \begin{cases} 1 \text{ if D1.episodic} = \text{D2.episodic} \\ 0 \text{ if D1.episodic} \neq \text{D2.episodic} \end{cases}$$

$$\mathrm{sim(D1.inheritance,D2.inheritance)} := \begin{cases} 1 \text{ if D1.inheritance} = \text{D2.inheritance} \\ 0 \text{ if D1.inheritance} \neq \text{D2.inheritance} \end{cases}$$

$$\mathrm{sim(D1.onset,D2.onset)} := \begin{cases} 1 \text{ if D1.onset} = \text{D2.onset} \\ 0 \text{ if D1.onset} \neq \text{D2.onset} \end{cases}$$

$$\mathrm{sim(D1.etiology,D2.etiology)} := \begin{cases} 1 \text{ if D1.etiology} = \text{D2.etiology} \\ 0 \text{ if D1.etiology} \neq \text{D2.etiology} \\ p \text{ if D1.etiology} \approx \text{D2.etiology} \end{cases}$$

$$\mathrm{sim(D1.tissue,D2.tissue)} := \begin{cases} 1 \text{ if D1.tissue} = \text{D2.tissue} \\ 0 \text{ if D1.tissue} \neq \text{D2.tissue} \\ q \text{ if D1.tissue} \approx \text{D2.tissue} \end{cases}$$

As mentioned above, indexing of diseases according to their etiology and tissue does allow multiple attributes. Partial matches indicated by $\approx$ between two sets of attributes are scored as p and q from the interval [0..1].

Phenotypically similar diseases are grouped into clusters, using a complete linkage strategy. Thus the required similarity between each pair of diseases within a same cluster is specified by a given threshold. In order to improve the quality of the disease clusters, diseases regarded similar, are placed next to each other as a starting condition of the clustering algorithm.

In the following analysis, all weights in the above proposed disease similarity function are kept fixed. The respective value is searched by a human expert, manually assessing the resulting intra-cluster similarity of diseases.

Scoring candidate genes for a disease of unknown genetic cause

After diseases of known genetic origin are clustered according to their phenotypic similarity, candidates from

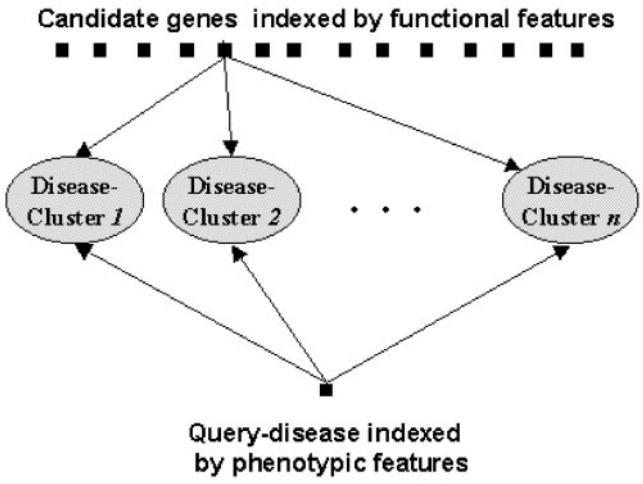

Fig. 1. Outline of the reasoning process: The clusters contain phenotypically similar diseases and their known respective underlying disease genes. For a query disease, all similar clusters are identified. Scores of candidate genes as related to known disease genes within these clusters are summed up. Top scoring candidate genes are taken as more probable candidates for the respective query disease.

the human genome are scored for these clusters. Thereto annotations of candidate genes are compared to the annotations of knowns disease genes, underlying diseases in a cluster. Thus annotations of disease genes in a phenotypic similarity cluster link a query disease to candidate genes, which are not connected to the disease yet. By combining the similarity score between the query disease and the clusters with the candidate gene scores as related to the respective clusters, candidate genes relevant for a query disease are identified. The reasoning process is outlined in Figure 1.

In the presented analysis, scoring of candidate genes relies on recent efforts, that provided Gene Ontology annotations for the human genome (Apweiler *et al.*, 2001). Gene Ontology (GO) is a controlled vocabulary of terms, describing the biological roles of molecular entities and their relationships (Gene Ontology Consortium, 2000). Totally, we obtain a set of 10 672 GO-annotated candidate genes.

If C is a candidate gene from the human genome having n GO-annotations GOA, the score of C for a disease cluster having m entries, is computed as the average ratio of the number of each GO-annotation of cluster disease genes identical to the candidate's GO-annotation and the number of the respective GO-annotation in all disease genes, scaled by the size of the cluster.

$$\text{score}(C,Cluster) := \frac{\sum_{GOA} \frac{\#GOA_i \in Cluster}{\#GOA_i}}{n \cdot m}$$

The score estimates, how well the cluster is separated from other clusters by features shared between the candidate gene with disease genes within the cluster, compared to features shared between the candidate gene and all other disease genes.

For a query disease D, we define the set $SimClusters_D$ as all these clusters, which are regarded similar to D. Subsequently the score of a candidate C from the human genome for a query disease D is computed as the sum over all scores of the candidate C for all $SimClusters_D$.

$$\text{score}(C,D) := \sum_{SimClusters_D} score(C, SimCluster_{D_j})$$

This score estimates the degree of association between an annotated candidate gene and a query disease as the sum over all similarity scores between the candidate and clusters containing diseases, which are phenotypically similar to the query disease.

The similarity between a query disease D and a disease cluster is computed as the average similarity between the disease D and all m diseases D_k contained in the cluster:

$$\text{sim}(D,Cluster) := \frac{\sum_{D_k \in Cluster} sim(D, D_k)}{m}$$

This similarity score corresponds to an average-linkage score, using the above similarity measure.

Computing functional similarity between disease genes

A functional similarity score between two disease genes $G1$, $G2$ is needed for validation of our underlying assumption. This functional similarity between two genes is computed as the average specificity *Sp()* of their shared n GO-annotations GOA:

$$\text{sim}(G1,G2) := \begin{cases} 1 \text{ if } G1 = G2 \\ \frac{\sum_{GOA_i} Sp(GOA_i)}{n} \end{cases}$$

Due to the hierarchical nature of GO, different levels of detail of an GO-annotation roughly relate to the hierarchy levels of respective GO-terms. So the hierarchy level of a GO-term is used as the specificity of an annotation in the given formula. We are aware, that the hierarchy level of a GO-term is not in complete accordance with its specificity. Thus applying the hierarchy level of a GO-term as its specificity may give only an estimate of an ideal score. However, the defined score is sufficient for our needs (see below).

Constructing a test set for assessment of disease gene predictions

As described above, disease entries from the OMIM database are indexed according to their clinical phenotype

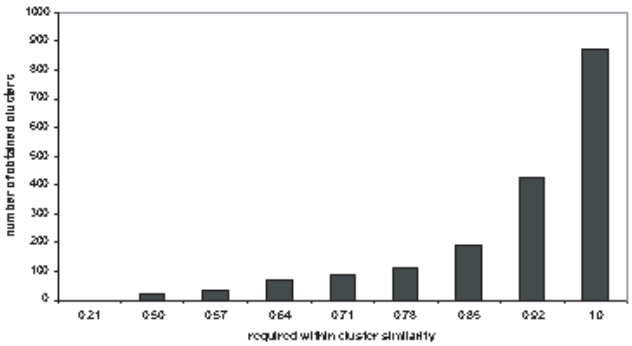

Fig. 2. The number of phenotypic similarity clusters is shown for different required pairwise within cluster-similarity. Thresholds are scaled linear to the interval [0..1]. As expected, the number of obtained clusters increases sharply with the required minimum similarity.

for the presented experiments. The sample of indexed entries, where a certain underlying gene has been annotated by GO-terms, contains 878 entries. Each entry consists of a disease and an underlying disease relevant gene. This set of disease entries is taken as test-set for assessment of our method. Thus totally 878 known disease genes are to be identified from the above mentioned 10672 GO-annotated genes.

RESULTS

Testing the underlying assumption

The algorithm starts from the assumption, that phenotypically similar diseases are caused by similar molecular mechanisms. This assumption relates to the observation by Jiminez-Sanchez *et al.* (2001), who show that functionally similar genes, when mutated, have a characteristic pattern of human disease.

To assess the assumption that phenotypically similar diseases are caused by similar molecular mechanisms, a complete linkage clustering of disease entries from the OMIM-database with respect to their phenotypic indices is performed, as described above. In an analog fashion, a complete linkage clustering of disease entries with respect to the functional similarity of their underlying disease genes is performed. Consecutively each disease entry belongs to a unique phenotypic similarity cluster as well as a unique functional similarity cluster. Of course both size and content of the clusters depend on the required similarity between entries within a cluster. As expected, the total number of both the phenotypic disease similarity clusters (Figure 2) and the functional disease gene similarity clusters (data not shown) increases with more stringent within cluster similarity required.

If the assumption that phenotypically similar diseases are caused by similar molecular mechanisms holds true, membership of disease entries in both types of clusters are dependent on each other. To evaluate this dependency, the actual counts of disease entries are recorded in a *n x m* contingency table. A chi-square test of statistical independence between phenotypic cluster memberships and functional cluster memberships is performed.

Table 2. Error probabilities of dependence between disease entry memberships in phenotypic similarity clusters and functional similarity clusters. The top line shows the required within cluster phenotypic similarity, the left line shows the required within cluster functional similarity. A stable region in parameter space showing high significance is found in the central part of the table

similarity threshold	**0.57**	**0.64**	**0.71**	**0.78**	**0.85**
0.27	0.16	0.85	0.12	0.66	0.67
0.36	0.18	0.06	0.003	0.02	0.17
0.45	0.12	0.04	0.02	0.004	0.12
0.54	0.53	0.40	0.11	0.004	0.13
0.63	0.69	0.96	0.20	0.12	0.59

Table 3. Different thresholds are applied for computing the similarity between a query disease and the respective disease clusters (top row). The fraction of diseases from the 878 diseases in the test set, where at least one cluster is recognized as similar, is shown in the bottom row

similarity threshold	**0.64**	**0.71**	**0.78**	**0.85**	**0.92**
fraction of diseases	0.99	0.98	0.97	0.92	0.66

In Table 2, error probability values for rejection of the null hypothesis of independence between the memberships of disease entries in phenotypic similarity and functional similarity clusters depending on different similarity thresholds are tabulated. A stable region in parameter space is detected, where the assumption appears to be significant. Using the similarity definitions given above, this regions contains the phenotypic similarity thresholds [0.64..0.78] combined with the functional similarity thresholds [0.36..0.45].

Prediction of disease-relevant genes

GO-annotated candidate genes can be scored for a query disease, if at least one disease cluster is recognized as similar to a query disease. This means, if for certain parameter values no disease cluster at all is recognized as similar to a query disease, no candidate genes are scored for this disease. Table 3 shows for different similarity thresholds the fraction of diseases, where at least one disease cluster is recognized as similar. The threshold similarity between a query disease and a cluster is set to be the required similarity within a cluster respectively.

Predictions appear to be most powerful in cases, where different members of a disease family are caused by mutations in different genes. For example, different forms of colorectal neoplasms are related to Beta Catenin (IPI00017292), Bax Protein (IPI00023992), APC Protein (IPI00012391), Tumorantigen P53 (IPI00025087), PMS Protein Homolog1 (IPI00005541), PMS1 Protein Homolog 2 (IPI00005543), Transforming Protein N-Ras(IPI00000005), Tranforming Protein P21B (IPI00000010), Colorectal Mutant Cancer Protein (IPI00011957), DNA Mismatch Repair Protein MLH1 (IPI00011957), DNA Mismatch Repair Protein MSH2 (IPI00017303), Chloride Anion Exchanger (IPI00031036), TGF-Beta Receptor Type II (IPI00020431) and the Tumor Suppressor Protein DCC (IPI00016422). Mutations in the respective genes may produce different phenotypes of intestinal neoplasms. Both the disease phenotype and disease genes functions are clearly similar to each other. The same holds true for other diseases, such as epilepsy, cardiomyopathy or peripheral neuropathy.

At the other end of the spectrum, prediction results appear to be poor for syndromal phenotypes, such as Wiskott–Aldrich Syndrome, Beckwith–Wiedemann Syndrome or Prader–Willi Syndrome. This may be either due the partly understood causal mechanisms or weak phenotypic indexing because of the more complex phenotype. However, no general rule can be recognized, on what kind of diseases the systems succeeds or fails in either case.

Benchmarking prediction of disease-relevant genes

A leave-one-out cross validation is performed on the above described test-set. This means for each disease, clusters of phenotypic similar diseases are recomputed, disregarding the respective test disease. The respective underlying disease gene is regarded as unknown and the above set of 10 672 GO-annotated candidate genes is scored by our method.

Predictions are then ordered according to their scores in a descending manner. Top scoring genes may be taken as more probable candidates for a query disease.

In Figure 3 the cumulative sum of true predictions contained in a candidate set is plotted over the threshold size of the candidate set. Different curves relate to different cluster similarity threshold parameters.

The diagonal shows number of true predictions, if candidate genes were picked by chance alone. The difference of the plotted curves and the diagonal relates to the average amount of information, which is gained about candidate genes by our method. Using high stringent thresholds obtains high numbers of true predictions in a relatively small sized candidate set. Thus high stringent thresholds give the most confident predictions. However, in a significant fraction of diseases, no candidates are scored at all.

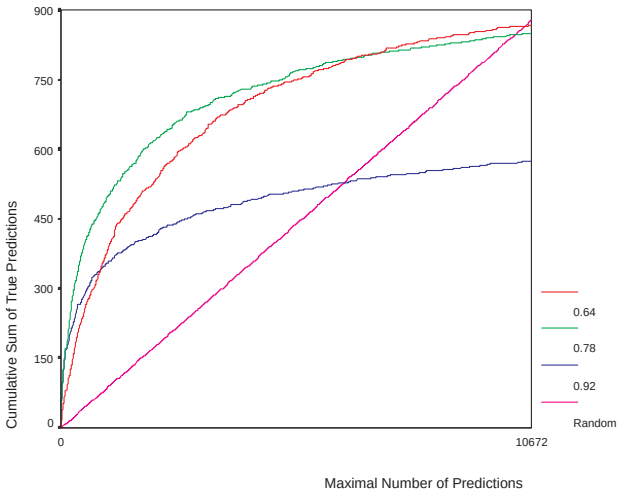

Fig. 3. The cumulative sum of true solutions contained in the candidate set is plotted over its threshold size. Totally 878 test diseases are to be predicted from 10672 candidate genes. Different curves show different required cluster similarity thresholds. The diagonal indicates the number of correct predictions contained in the candidate set by chance alone.

If a medium stringent threshold is used for definition of phenotypic similarity, both a high number of true positive members in the candidate sets as well as high fraction of predicted test diseases is achieved. Low stringent thresholds are able to score the set of candidate genes for the all test diseases, but the quality of the respective predictions is reduced.

Using high stringent thresholds, in about one-third of predicted diseases (193 of 580) the true solutions is contained within the top scoring 1.5% of candidates (160 of 10 672). Using medium stringent thresholds, in about one-third of cases (284 of 851) the true solutions is contained within the top scoring 3% of candidates (321 of 10 672) and in two-third of cases (568 of 851) the true solutions is contained within the top scoring 15% of candidates (1600 of 10 672). Thus most confident predictions are obtained based on small sets of highly similar diseases to a query disease. If such highly similar diseases are not available, sensible results can still be obtained using relaxed similarity thresholds.

CONCLUSION AND FUTURE WORK

We present a new approach for prediction of disease relevant human genes: a query disease is linked to its potentially underlying disease genes, using known genetic causes of similar diseases. Therefore a similarity measure between the phenotypes of two diseases is defined. Diseases of known genetic origin are then clustered according to their phenotypic similarity. Each cluster entry

consists of a disease and the respective underlying disease gene. Candidate genes are scored for each cluster, by comparing features of the candidate gene to features of disease genes in the cluster. Candidate genes relevant for a query disease are identified by combining these scores with the similarity between the query disease and the disease clusters.

Improvement of the presented method may result from more detailed and revised phenotypic indexing of diseases, as well as a more systematic search of the parameter space, to optimize the applied similarity functions. In the presented validation study, clusters of phenotypically similar diseases are computed in an automatic fashion. Instead, manual editing of disease clusters may lead to improved results, because more detailed biomedical knowledge can be applied.

Presently large scale approaches investigating genetic variation underlying genetic disease are under way worldwide (Heil *et al.*, 2002). As shown here, it is possible to significantly decrease the amount of required experimental resources by including easily accessible knowledge into an experimental set-up. However, computational prediction of disease relevant genes must be regarded as an extremely hard problem, with probably no biomedical optimal solution attainable at all. More than 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. Evidence independent to our predictions is given for example as positional information, which is experimentally generated by family linkage studies. These studies are successful to a point, where a relatively high number of positional candidate genes remain. Consecutively preference might be given to these candidates, which are both predicted by the presented method and which are supported by linkage studies. However, it has to be kept in mind, that positional evidence is probabilistic by its nature.

The presented work has been made possible by the high quality annotations of the human genome and proteome, provided by the NCBI (Hamosh *et al.*, 2000) and the EBI (Apweiler *et al.*, 2001). We expect further improvement of the method with the further increase of quantity and quality of the human genome annotation. An optimistic bias in the presented benchmarking surely results from the fact, that known genes underlying disease in the OMIM database tend to have been the subject of more empirical research and are more likely to be well annotated. However, our approach is not restricted to GO-annotations of candidate genes. Instead, experimental data could be used as annotating features, linking candidate genes and known disease genes. For example gene expression patterns could be used as features, to score a candidate for a disease cluster. This approach would extend the gene expression analysis method proposed by Zien *et al.* (2000), who score high order knowledge about molecular pathways with respect to expression data.

Since most knowledge gathered in OMIM relates to monogenic diseases, one may ask, whether prediction of candidate genes for complex diseases is out of reach for our approach. This is true in the sense, that predictions presently rely on the diseases gathered in OMIM. However, also monogenic forms of complex diseases can give important clues to the underlying mechanisms.

A lot of research in bioinformatics is based on the fact, that biological data can be grouped with respect to meaningful similarities. In this context, clustering of human diseases according to a combination of both clinical appearance and underlying molecular mechanisms still needs further exploration. So it might also be interesting, to draw explicit molecular knowledge from the disease clusters. The presented work should be seen as a small step of ongoing work, using computational methods to relate human diseases to their molecular basis.

ACKNOWLEDGEMENTS

This work was supported by the German National Genome Research Network (J.F.) We thank the anonymous reviewers for their helpful comments on the manuscript. We appreciate the work of data annotators at the EBI and NCBI.

REFERENCES

International Human Genome Sequencing Consortium (2001) Initial Sequencing and Analysis of the Human Genome. *Nature*, **409**, 860–921.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Lander,E. and Kruglyak,L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.*, **11**, 241–247.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human disease. *Science*, **273**, 1516–1517.

Perez-Iratxeta,C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genet.*, (June 2002 advance online publication).

Hamosh,A. *et al.* (2000) Online Mendelian Inheritance in Man (OMIM). *Human Mutation*, **15**, 57–61.

Apweiler,R. *et al.* (2001) Proteome analysis database: online application of InterPro and ClusSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.*, **29**, 44–48.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Jiminez-Sanchez,G. *et al.* (2001) Human disease genes. *Nature*, **409**, 853–854.

Heil,A. *et al.* (2002) An automated computer system to support ultra high throughput SNP genotyping. *Pac. Symp. Biocomput.*, **7**, 41–52.

Zien,A. *et al.* (2000) Analysis of gene expression data with pathway scores. *Proceedings of the International Symbosium on Molecular Biology, 2000*. pp. 407–417.