ARTICLE

www.nature.com/ejhg

A text-mining analysis of the human phenome

Marc A van Driel¹, Jorn Bruggeman², Gert Vriend¹, Han G Brunner^{*,3} and Jack AM Leunissen²

¹Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, Toernooiveld 1, 6525ED Nijmegen, the Netherlands; ²Department of Bioinformatics, Wageningen University and Research Centre, Dreijenlaan 3, 6703HA Wageningen, the Netherlands; ³Department of Human Genetics, University Medical Centre Nijmegen, Geert Grooteplein 10, 6525GA Nijmegen, the Netherlands

A number of large-scale efforts are underway to define the relationships between genes and proteins in various species. But, few attempts have been made to systematically classify all such relationships at the phenotype level. Also, it is unknown whether such a phenotype map would carry biologically meaningful information. We have used text mining to classify over 5000 human phenotypes contained in the Online Mendelian Inheritance in Man database. We find that similarity between phenotypes reflects biological modules of interacting functionally related genes. These similarities are positively correlated with a number of measures of gene function, including relatedness at the level of protein sequence, protein motifs, functional annotation, and direct protein–protein interaction. Phenotype grouping reflects the modular nature of human disease genetics. Thus, phenotype mapping may be used to predict candidate genes for diseases as well as functional relations between genes and proteins. Such predictions will further improve if a unified system of phenotype descriptors is developed. The phenotype similarity data are accessible through a web interface at http://www.cmbi.ru.nl/MimMiner/.

European Journal of Human Genetics (2006) 14, 535–542. doi:10.1038/sj.ejhg.5201585; published online 22 February 2006

Keywords: phenome; text mining; candidate disease genes; phenotype-genotype relations

Introduction

Functional annotation of genes is an important challenge once the sequence of a genome has been completed. Gene annotation encompasses a variety of functional attributes, from structural motifs, through cellular function, to associations with specific functions and processes at the level of the organism.

Apart from descriptors at the gene and protein level, the phenotype effect of a mutated or deleted gene forms part of its functional annotation. Systematic mutation and RNA interference screens have been performed for selected

Tel: +31 24 361 4017; Fax: +31 24 366 8752;

E-mail: H.Brunner@antrg.umcn.nl

Received 23 August 2005; revised 12 December 2005; accepted 5 January 2006; published online 22 February 2006

phenotypes in *Drosophila. melanogaster, Caenorhabditis. elegans,* and *Saccharomyces. cerevisiae.*¹⁻³ Also for *Mus. musculus,* an ambitious project to mutagenize most or all genes has been conceived.⁴

Previous studies have correlated various attributes of human genes, such as predicted function or amino-acid sequence length, with the chance of causing a disease.^{5–7} However, only limited attention has been paid to the grouping of (disease) phenotypes into a matrix as a means of predicting biological relations between genes and proteins. Such systematic grouping of genes by their associated phenotypes may be referred to as phenomics.

The existence of specific disease phenotype groups suggests that phenomics is possible in humans. Also, there are large numbers of naturally occurring mutations and we have a detailed knowledge of the phenotypes that are associated.^{8,9} Qualitatively, the human mutation data set surpasses that of most model organisms, because we can detect and describe human anomalies in more detail than

^{*}Correspondence: Professor HG Brunner, Department of Human Genetics, University Medical Centre Nijmegen, Geert Grooteplein 10, 6525GA Nijmegen, The Netherlands.

in other species.¹⁰ Specific examples illustrate that individual genes that cause a given phenotype tend to be linked at the biological levels as interacting proteins, as components of a multiprotein complex, or as steps in a biochemical pathway.

We note that current human disease phenotype data set(s) such as Online Mendelian Inheritance in Man (OMIM) are far from ideal for bioinformatics purposes as they were not designed with the intention to systematically describe phenotypes. Therefore, this study can only provide proof of concept, and certainly not a finished product. Our analyses show that, these limitations notwithstanding, there is a detectable association between phenotype clusters and the function of the underlying genes. We have further devised and used a system for assigning similarity scores, which allows all genes with known phenotypes to be compared. This approach is very different from that which uses an artificial division into predetermined entities.¹¹ We have classified over 5000 disease phenotypes in humans into a single human phenome system. Given that the human phenome reflects the biology of the system, any phenotype classification should reflect other measures of gene function at least to some extent. We have therefore compared the organization of genes within this human phenome map to their known interactions, and similarities at multiple levels including sequence, protein motifs, and assigned gene ontology (GO) functions. Our analyses show that for similar human disease phenotypes, there is indeed a consistent association at multiple levels of gene annotation.

The phenotype similarity data are accessible through a web interface (http://www.cmbi.ru.nl/MimMiner/). This interface, called the 'MimMiner', enables the user to retrieve the similarity ranking for a specific OMIM phenotype.

Methods

The OMIM database

The OMIM (November 25, 2004) database¹² contains record-based textual information, one gene or one genetic disorder per record. OMIM also contains literature references and links to other databases. We have used the full-text (TX) and clinical synopsis (CS) fields of all records that describe genetic disorders. We will refer to this combination of the TX and CS fields as a 'record'.

OMIM is a rich data set containing 16357 TX records of which 5132 describe a disease phenotype (semi-automatically selected, manually verified). The remaining records contain variation, mutation, gene/protein, or other information.

OMIM was originally designed to be read by humans, not by computer. We have automatically extracted the phenotypic features from OMIM using text analysis techniques.

Creation of 'feature vectors'

We used the anatomy (A) and the disease (C) sections of the medical subject headings vocabulary (MeSH) to extract terms from OMIM. MeSH terms and their plurals and components are concepts. MeSH provides a standardized way to retrieve information that uses different terminology to refer to the same concepts. Its size and internal hierarchical structure make it a rich dictionary, which is needed to match the OMIM texts. MeSH concepts serve as phenotype features characterizing OMIM records: every entry in the feature vectors represents an MeSH concept. The number of times the terms for a given concept are found in an OMIM record reflects the concept's relevance to the phenotype.^{13,14} Nonspecific concepts like 'syndrome' or 'disease' were excluded.

Refinement of the feature vectors

MeSH concepts can be very broad like 'Eye' or more specific like 'Retina'. MeSH includes a concept hierarchy that describes relationships such as 'Eye' – 'Retina' – 'Photoreceptors'. 'Eye' is called a hypernym of 'Retina', which in turn is a hypernym of 'Photoreceptors', etc. Conversely, 'Retina' is called a hyponym of 'Eye'. To ensure that the concepts 'Eye' and 'Retina' are recognized as similar, we use the MeSH hierarchy to encode this similarity in the feature vectors by increasing the value of all hypernyms as described in (Figure 1)

$$r_c := r_{c,counted} + \frac{\sum r_{hypo's}}{n_{hypo,c}}.$$

For any concept *c*, its relevance r_c becomes the actual count of the concept in a document $r_{c,counted}$ plus the relevance sum of the concept's hyponyms $r_{hypo's}$. This sum is divided by the number of hyponyms $n_{hypo,c}$. This equation is applied iteratively from the most detailed level in the MeSH tree, till the highest hypernym level is reached. Not all concepts in the OMIM records are equally informative. For example, 'retina pigment epithelium' occurs rarely, and thus provides more specific information than very frequently occurring terms such as 'Brain'. We allowed for differences in the importance of concept frequencies by using the inverse document frequency measure¹⁵

$$gw_c = \log_2 \frac{N}{n_c}$$

The inverse document frequency or global weight of concept c (gw_c) is the logarithm of the total number of records analyzed (N; N = 5080) divided by the number of records that contain concept c, n_c . Not all OMIM records contain equally extensive descriptions. These differences will make a comparison between records difficult because the diversity and the frequency of concepts in the larger records will exceed those in the smaller records. Equation





Figure 1 Example of concept expansion using the MeSH hierarchical structure. The concept 'Photoreceptors' (with MeSH descriptor: D010786) is found twice in an OMIM record. Expansion of this concept gives the hypernym 'Retina' (D012160). The relevance of the concept 'Retina' is derived from 'Photoreceptors' according to equation (1). 'Retina' has seven descendants or hyponyms, thus its relevance becomes $1/7 \times 2$ (for two times 'Photoreceptors'). Similarly, Retina's hypernym 'Eye' (D005123) has 11 hyponyms, so that 'Photoreceptors' contributes $1/11 \times 1/7 \times 2$ (= 0.026) to the relevance of 'Eye'. (Squared brackets indicate the MeSH tree position and a '+' sign that there are more specific concepts underneath).

(3) was used to (partly) correct for these record size differences. $^{\rm 15}$

$$r_c = 0.5 + 0.5 \cdot \frac{r_c}{r_{mf}}$$

The local weight of concept *c* in a record is a function of the concept's frequency r_c divided by the frequency of the most occurring MeSH concept in that record, r_{mf} . The three feature vector corrections were applied in the order of equations (1)–(3).

Comparing OMIM records

The similarity between OMIM records can be quantified by comparing the feature vectors that are expanded and corrected (equations (1)-(3)). Similarities between feature vectors were determined by the cosines of their angles (equation (4)).¹⁶

$$s(x,y) = \frac{\sum_{i=1}^{l} x_i y_i}{\sqrt{\sum_{i=1}^{l} x_i^2} \sqrt{\sum_{i=1}^{l} y_i^2}}$$

The similarity between the feature vectors X and Y (s(X, Y)) is a function of their respective concept frequen-

cies x_i and y_i . The index *i* runs from 1 to the number of MeSH concepts l.

Phenotype-genotype correlations

The matrix of all pairwise vector similarities was denoted the phenomap. A subset of this phenomap containing the 1653 OMIM records for which the causative gene and protein are known was used as a starting point for determining the relation between phenotypic similarities on the one hand and genotypic similarities on the other. All 1653 phenotypes associated with a protein in the UniProt database¹⁷ were then compared to four genotyperelated data sets. The average of 10 randomized phenomaps was used as a control for background signal. Feature vectors were randomly permutated using Fisher–Yates shuffling¹⁸ before computation of the phenomap.

The PFAM database¹⁹ was used to determine whether pairs of genes share similar domains.

To check if genes are similar at the sequence level, we compared the proteins associated with the 1653 phenotypes from the UniProt database in an all-against-all Smith–Waterman analysis²⁰ (Blosum-90). Sequence pairs with an alignment e-value better than 10^{-6} were considered similar.²¹

Protein-protein interactions were extracted from the interaction section of the HPRD database²² and used to

check whether the proteins are part of the same complex or interact in any other way.

The GO database²³ and the GO annotations $(GOA)^{24}$ were used to determine if two genes fall within the same functional category.

Two genes/proteins were considered related when they shared at least three GO terms at the sixth annotation level.²⁵ Annotations at the more detailed levels (level 7, 8, etc) were converted to the corresponding annotation at the sixth level. For example, the RDS protein [UniProt:P23942] has the detailed annotation 'visual perception' at level 7 [GO:0007601], which gets converted to the more general 'sensory perception of light' [GO:0050953] at level 6.

Clustering

Multidimensional scaling methods did not show evident clusters, which makes every clustering attempt intrinsically subjective. In order to objectively visualize the relations between the phenotypes, we performed average-link clustering that results in a tree structure.²⁶ This tree can be explored at http://www.cmbi.ru.nl/MimMiner/.

Results

Feature vectors

A total of 5132 OMIM records (31%) describe a disease phenotype, and their TX and CS fields were analyzed for the presence of concepts from the anatomy (A) and disease (C) sections in the MeSH thesaurus. For 5080 OMIM records, we could match one or more MeSH terms.

Comparing OMIM records

The 5080 × 5080 pair-wise feature vector similarities form the phenomap. The distribution of these similarity scores is shown in Figure 2a. We used multi-dimensional scaling methods to test whether the phenotypes form distinct groups. The goal of these methods is to detect meaningful underlying dimensions from high-dimensional data. For computational reasons, we used a sample that contains only the OMIM records with a CS field. Principal component analysis, classic multidimensional scaling, and independent component analysis implementations from the 'R' software package²⁷ were used, but none of the methods showed evident groups. This suggests that the human phenotypes when based on MeSH terms form a continuous spectrum rather than distinct classes (data not shown).

For each OMIM record, the most similar of the other 5079 records was identified (Figure 2b). We asked whether moderately similar phenotype pairs might still yield reasonable hypotheses. For gene-finding in humans, this appeared to be the case in a number of instances. Even in the 0.3–0.4 bin, individual cases displayed potentially relevant phenotypic similarities. For example, 'Fibromuscular Dysplasia of Arteries' [OMIM:135580] and 'Cardio-

myopathy, Familial Hypertrophic' [OMIM:192600] have a phenotypic similarity score of 0.31, the ninth most similar score. The observation that these conditions are clinically inter-related is supported by two case reports.^{28,29}

Next, we asked if clinically different conditions caused by mutation of the same gene were likely to be detected as having considerable phenotype overlap by our system. Indeed, approximately 40% of phenotypes caused by allelic mutations were more similar to each other than to any other phenotype in the data set (data not shown).

The conclusion is that the more phenotypes resemble each other, the more likely they are to share an interaction or pathway. However, we cannot define a general cutoff for similarity scores because even low-scoring OMIM records



Figure 2 Histograms of phenotype similarity scores. (a) Histogram of all pairwise phenotype similarity scores of the 5082 phenotype records. The vertical axis is logarithmic; most phenotype–phenotype pairs have a low similarity score. (b) The best scores for all phenotypes in the disease phenotype datas et (nearest-neighbor similarity).

can occasionally contribute to our understanding of the diseases.

Phenotype-sequence similarity correlations

We asked whether the similarity at the phenotype level predicts similarity in gene/protein function. The causative gene/protein is known for 1653 of the 5080 OMIM records that describe a phenotype (33%). The corresponding 1653 × 1653 phenomap was extracted from the total phenomap. These 1653 phenotypes linked to 2168 corresponding protein sequences (1401 unique sequences). Sequences were extracted from the UniProt database and used to calculate all-against-all Smith–Waterman sequence

alignments. Figure 3a shows the fraction of significant sequence alignment similarities as a function of the phenotype similarity scores. The percentage of phenotype pairs for which the causative proteins are similar increases with increasing phenotype similarity score from 0.6% to a maximum of 26.6%. Approximately half of these are owing to different mutations in the same gene causing similar phenotypes. For example, 'Robinow–Sorauf syndrome' [OMIM:180750] and the related 'Saethre–Chotzen syndrome' [OMIM:101400] are both caused by a mutation in TWIST1 [UniProt:Q15672]. The other relations are owing to mutations in different genes that share sequence similarity. For example, the 'Rufous Oculocutaneous



Figure 3 Phenotype similarity *versus* genotype relation categories. (a) Sequence similarity (Smith–Waterman, threshold e-value 1e-6). (b) Sharing of at least one PFAM domain. (c) Protein–Protein interactions according to the HPRD database. (d) Sharing of three or more GOA at the sixth or more detailed level. The 0.8-0.9 and 0.9-1.0 phenotype similarity bins suffer from low counts. The average signal of 10 randomized phenomaps is at the level of the two lowest bins (data not shown).

Albinism' [OMIM:278400] and 'Albinism, Oculocutaneous, Type IB' [OMIM:606952] phenotypes show 0.68 phenotypic similarity. These diseases are caused by mutations in the TYRP1 [UniProt:P17643] and TYR [UniProt:P14679] proteins, respectively, that are 43% identical at the sequence level. Mutations in TYRP1 also cause 'Oculocutaneous Albinism Type 3' [OMIM:203290].

Many proteins have multiple functional domains. If two proteins share a functional domain, then mutations may disrupt the same process and thereby lead to similar phenotypes. This can be true, even if the full protein sequences do not share significant sequence identity. Figure 3b shows the percentage sequence pairs that share a PFAM domain as a function of the phenotype similarity scores. The majority of pairs that share a PFAM domain (67% or 119/181, disallowing the same gene in the 0.5-0.6bin) also share significant overall sequence similarity. Only in a minority did sequence comparison based on domain sharing add new information compared to overall sequence similarity. For some phenotype pairs, the proteins lack significant overall sequence identity, but do share one or more common structural features. For instance, 'Long Qt Syndrome 3' [OMIM:603830] is caused by a mutation in SCN5A [UniProt:Q14524] and shares phenotypic characteristics with 'Jervell And Lange-Nielsen Syndrome' [OMIM:220400] that can be caused by a mutation in KCNQ1 [UniProt:P51787]. These two proteins have an 'Ion Transporter' domain in common [PFAM:PF00520].

Phenotype-protein interaction correlations

We checked which of the 1653×1653 pairs have an interaction described in the HPRD (Figure 3c). Although the HPRD dataset is sparser than the other datasets, and thus reveals fewer relations, 54% of these HPRD relations were not yet detected by the sequence alignments or the PFAM analyses. To check whether possible relations can still be found in the lower similarity ranges, we performed a detailed inspection on the 212 protein interactions that are listed in the HPRD for pairs that have a phenotypic similarity between 0.3 and 0.4. Over 50% of the relations suggested a plausible reason for the phenotypic similarities. For example, 'Wiskott-Aldrich Syndrome' [OMIM:301000] and 'Fleisher Syndrome' [OMIM:307200] are caused by mutations in WAS [UniProt:P42768] and BTK [Uni-Prot:Q06187], respectively. WAS and BTK do not share significant sequence similarity, nor do they share a PFAM domain. However, both are involved in cell growth regulation and cytoskeleton processes.³⁰ WAS is phosphorylated by BTK, leading to activation of the Arp2/3 complex.³¹ In general, despite its sparse nature, the HPRD provided biologically plausible information.

Phenotype-functional process correlations

Various excellent databases describing pathway information are available, for example, KEGG,³² BRENDA,³³

Reactome,³⁴ etc. Most of these databases focus on metabolic pathways, whereas less than 10% of the OMIM phenotype records relate to metabolic disorders. To get an impression of possible functional relations between genes and proteins, we compared their GOA. We defined GO similarity by the sharing of at least three GOA at the sixth or more detailed GO level.²⁵ The signal we find is well above the average of 10 randomized matrices ($\sim 7\%$) over all bins. The percentage of pairs that share three or more GOA increased as a function of the phenotypic similarity (Figure 3d). Compared to the three other sets, 76% of the relations were new. A random set of 50 of the 786 proteins that share three or more GOA in the 0.5-0.6 bin were inspected in more detail. A plausible reason for the phenotypic similarities was found in 37 of these 50 cases. Unlike in Figure 3a-c, less than half of the relations are owing to a defect in the same gene, which suggests the criterion of sharing three GOA at the sixth level is less stringent than in the other data sets.

Nevertheless, sharing GOA can be regarded as a relatively nonspecific characterization of gene function with a high noise level (Figure 3d). Overall, when the genotype by phenotype relations were normalized using random phenomaps, the HPRD was most efficient in providing nonrandom gene–gene relationships (Figure 4).

Discussion

We have developed a text-mining approach to map relationships between more than 5000 human genetic



Figure 4 Histogram of normalized genotype relations as a function of phenotype–phenotype similarity. All values are corrected for random information levels. Although the HPRD data set contains fewer relations than the other sets, the normalized signal is more than two times as strong, reflecting the quality of the HPRD data set. The 0.8–0.9 and 0.9–1.0 phenotype similarity bins suffer from low counts.

540

disease phenotypes from the OMIM database. OMIM does not contain a standardized system for scoring phenotypes. We decided to use OMIM for this initial analysis because it is still by far the most comprehensive of available databases. The resulting phenotype matrix has a number of characteristics that suggest that it (or similar systems based on more comprehensive phenotyping efforts) might be a useful addition to functional genomics tools such as the HPRD and KEGG. As expected, we find that different phenotypes associated with mutations of a single gene show considerable overlap. Such allelic conditions are each others best phenotypic hit in 40% of the cases. After exclusion of allelic conditions, there remained a positive correlation between phenotypic similarities on the one hand, and sharing of gene sequence, protein motifs, functional annotation, and known protein interactions on the other (Figure 3). In these comparisons, we found a relatively high level of noise for GO. We believe that our finding that all gene attributes correlate with phenotype sharing can be explained by assuming that human phenotypes reflect disturbance of functional modules, more than of individual genes. One striking finding was that biologically meaningful relationships were mostly detected in the small fraction of the phenotype relations with a similarity score greater than 0.4 (Figure 3). The combined data suggest that we may indeed use phenotypic relationships as general indicators of biological and functional interactions at the gene and protein levels.

Several applications can be envisaged for the phenomap. First, our analysis suggests that the phenotype matrix may aid in the prediction of candidate genes for the 3400 traits listed in OMIM whose molecular basis remains to be defined. Second, it is conceivable that one would take phenotypic relationships as the starting point for biochemical and cell biological experiments in order to prove a suspected link at the gene and protein levels. Experiments of this type have already been shown to be successful (eg polycystic kidney disease (PKD1, PKD2), tuberous sclerosis (TSC1, TSC2), breast and ovarian cancer (BRCA1, BRCA2), and Fanconi anemia (FANCA-G)).^{35–38}

Finally, there may well be a point in pursuing large-scale phenotype analyses using more precise measures of the phenotypes themselves. OMIM was not designed as a structured database for phenotype analysis. Indeed, it does not contain rules for feature assignment and most of the phenotype information collected by our text-mining approach derives from free text fields. A more standardized method for phenotype description including frequency estimates for each feature would greatly increase the yield of the analyses of genotype–phenotype correlations.³⁹

Phenotype clustering reflects the modular nature of human disease genetics. Thus, the phenomap may be used to predict candidate genes for diseases. Such predictions will further improve if a unified system of phenotype descriptors is developed.

Acknowledgements

We thank M Huynen, B Snel, H van Bokhoven and R van der Heijden for fruitful discussions. M Oti, M Hekkelman, T Hulsen and J van Rooij are acknowledged for providing expert technical assistance. This work was supported by the Netherlands Organization for Scientific Research (NWO), SON/Unilever (326756), and the EU FP6 Programme (LHSG-CT-2003-503265).

Databases

http://www.ncbi.nlm.nih.gov/omim/; http://www.nlm.nih.gov/mesh/ meshhome.html; http://www.uniprot.org; http://www.sanger.ac.uk/ Software/Pfam/; http://www.hprd.org/; http://www.geneontology.org/; http://www.ebi.ac.uk/GOA/.

References

- 1 Boutros M, Kiger AA, Armknecht S *et al*: Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* 2004; **303**: 832–835.
- 2 Giaever G, Chu AM, Ni L et al: Functional profiling of the Saccharomyces cerevisiae genome. Nature 2002; 418: 387-391.
- 3 Rual JF, Ceron J, Koreth J *et al*: Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res* 2004; **14**: 2162–2168.
- 4 Auwerx J, Avner P, Baldock R *et al*: The European dimension for the mouse genome mutagenesis program. *Nat Genet* 2004; **36**: 925–927.
- 5 Jimenez-Sanchez G, Childs B, Valle D: Human disease genes. Nature 2001; 409: 853-855.
- 6 Lopez-Bigas N, Ouzounis CA: Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 2004; **32**: 3108–3114.
- 7 Perez-Iratxeta C, Bork P, Andrade MA: Association of genes to genetically inherited diseases using data mining. *Nat Genet* 2002; **31**: 316–319.
- 8 Brunner HG, van Driel MA: From syndrome families to functional genomics. *Nat Rev Genet* 2004; 5: 545–551.
- 9 Freimer N, Sabatti C: The human phenome project. Nat Genet 2003; 34: 15–21.
- 10 Donnai D, Read AP: How clinicians add to knowledge of development. *Lancet* 2003; 362: 477-484.
- 11 Katsanis N, Worley KC, Gonzalez G, Ansley SJ, Lupski JR: A computational/functional genomics approach for the enrichment of the retinal transcriptome and the identification of positional candidate retinopathy genes. *Proc Natl Acad Sci USA* 2002; **99**: 14326–14331.
- 12 Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002; **30**: 52–55.
- 13 Jenssen TK, Laegreid A, Komorowski J, Hovig E: A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001; **28**: 21–28.
- 14 Andrade MA, Valencia A: Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 1998; 14: 600–607.
- 15 Wilbur WJ, Yang Y: An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med* 1996; **26**: 209–222.
- 16 Hand DJ, Mannila H, Smyth P: *Principles of data mining*. Cambridge, MA: MIT Press, 2001.
- 17 Apweiler R, Bairoch A, Wu CH *et al*: UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004; **32** (Database issue): D115–D119.
- 18 Fisher RA, Yates F: *Statistical tables for biological, agricultural and medical research.* Edinburgh: Oliver and Boyd, 1938.
- 19 Bateman A, Coin L, Durbin R *et al*: The Pfam protein families database. *Nucleic Acids Res* 2004; **32** (Database issue): D138–D141.

- 20 Smith TF, Waterman MS: Identification of common molecular subsequences. J Mol Biol 1981; 147: 195-197.
- 21 Brenner SE, Chothia C, Hubbard TJ: Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA* 1998; **95**: 6073–6078.
- 22 Peri S, Navarro JD, Amanchy R *et al*: Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003; **13**: 2363–2371.
- 23 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
- 24 Camon E, Magrane M, Barrell D *et al*: The gene ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* 2003; **13**: 662–672.
- 25 Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci USA* 2003; **100**: 8348–8353.
- 26 Sokal RR, Mitchener CD: A statistical method for evaluation of systematic relationships. Univ Kansas Sci Bull 1958; 28: 1409–1438.
- 27 R-Development-Core-Team R: A language and environment for statistical computing. Austria: Vienna, 2004.
- 28 Safioleas M, Kakisis J, Manti C: Coexistence of hypertrophic cardiomyopathy and fibromuscular dysplasia of the superior mesenteric artery. *N Engl J Med* 2001; **344**: 1333–1334.
- 29 Scully RE, Mark ÉJ, McNeely WF, McNeely BU: Case records of the Massachusetts general hospital: Case 9-1995. *N Engl J Med* 1995; 332: 804–810.

- 30 Welch MD, Mullins RD: Cellular control of actin nucleation. *Annu Rev Cell Dev Biol* 2002; **18**: 247–288.
- 31 Torres E, Rosen MK: Contingent phosphorylation/dephosphorylation provides a mechanism of molecular memory in WASP. *Mol Cell* 2003; **11**: 1215–1227.
- 32 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004; **32** (Database issue): D277–D280.
- 33 Schomburg I, Chang A, Ebeling C *et al*: BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004; **32** (Database issue): D431–D433.
- 34 Joshi-Tope G, Vastrik I, Gopinath GR *et al*: The Genome Knowledgebase: a resource for biologists and bioinformaticists. *Cold Spring Harb Symp Quant Biol* 2003; **68**: 237–243.
- 35 Chen J, Silver DP, Walpita D *et al*: Stable interaction between the products of the BRCA1 and BRCA2 tumor suppressor genes in mitotic and meiotic cells. *Mol Cell* 1998; **2**: 317–328.
- 36 Qian F, Germino FJ, Cai Y, Zhang X, Somlo S, Germino GG: PKD1 interacts with PKD2 through a probable coiled-coil domain. *Nat Genet* 1997; **16**: 179–183.
- 37 van Slegtenhorst M, Nellist M, Nagelkerken B *et al*: Interaction between hamartin and tuberin, the TSC1 and TSC2 gene products. *Hum Mol Genet* 1998; 7: 1053–1057.
- 38 D'Andrea AD, Grompe M: The Fanconi anaemia/BRCA pathway. *Nat Rev Cancer* 2003; **3**: 23–34.
- 39 Biesecker LG: Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clin Genet* 2005; **68**: 320–326.