A literature network of human genes for high-throughput analysis of gene expression

Tor-Kristian Jenssen[1], Astrid Lægreid[2], Jan Komorowski[1,4] & Eivind Hovig[3]

We have carried out automated extraction of explicit and implicit biomedical knowledge from publicly available gene and text databases to create a gene-to-gene co-citation network for 13,712 named human genes by automated analysis of titles and abstracts in over 10 million MEDLINE records. The associations between genes have been annotated by linking genes to terms from the medical subject heading (MeSH) index and terms from the gene ontology (GO) database. The extracted database and accompanying web tools for gene-expression analysis have collectively been named 'PubGene'. We validated the extracted networks by three large-scale experiments showing that co-occurrence reflects biologically meaningful relationships, thus providing an approach to extract and structure known biology. We validated the applicability of the tools by analyzing two publicly available microarray data sets.

Introduction

Functional genomics is still at an early stage, but, with some genomes sequenced[1–3] and others near completion, attention is shifting towards assigning biological function to gene sequences[4]. DNA microarrays[5] and other high-throughput gene and protein assays will be critical tools for gene-expression analysis. The ability to use existing knowledge is fundamental to scientific discovery, but unsupervised methods for data analysis, such as hierarchical clustering, leave the user to incorporate background knowledge[6]. The large number of genes that can be included in such studies motivates the implementation of automated methods for use of existing knowledge.

The fact that a substantial amount of biomedical knowledge is recorded in only free-text form and, as such, is not readily available for computerized analysis has inspired research on methods for automated extraction of biomedical knowledge[7]. Many have focused on protein-protein[8–10] and gene-protein interactions[11] or other specific relationships between molecular entities, for example, cellular localization of proteins[12], molecular binding relationships[13], and interactions between genes or proteins and drugs[14]. A fundamental problem to be solved is the recognition of biomedical nouns or noun phrases (for example, gene[15] and protein names[16]). Noun recognition can also be done using predefined dictionaries, as is often the case for index-based information-retrieval systems. Keyword indexing has been used to annotate proteins[17] and was recently proposed for construction of co-occurrence networks of genes in human[18] and *Saccharomyces cerevisiae*[19]. Text mining of functional links based on document similarity is another strategy that has been used to extract and annotate relationships between genes[20].

Here, we present the completion of a full-scale literature network for 13,712 human genes extracted from the titles and abstracts of over 10 million article records from the MEDLINE citation database (http://www.ncbi.nlm.nih.gov/PubMed/) of the National Library of Medicine (NLM). We constructed the network from the co-occurrence of gene symbols or short gene names in the title or the abstract of a common article record. The method is based on the assumption that if two genes are co-mentioned in a MEDLINE record there is an underlying biological relationship. As co-occurrence may reflect many kinds of interactions, we annotated the network to better appreciate the nature of the extracted relationships. The annotation consisted of linking genes to terms from the MeSH index (http://www.nlm.nih.gov/mesh/meshhome.html) and terms from the Gene Ontology (GO) database[4]. The extracted information can be obtained through a set of web tools (http://www.PubGene.org) to be used for analysis of gene-expression data. The database and the tools are collectively named PubGene and are publicly available.

We evaluated the quality of the network by manual examination of 1,000 randomly chosen pairs of genes and by comparison with the Database of Interacting Proteins (DIP) and the Online Mendelian Inheritance in Man (OMIM) database. Following the analysis of the network, we used the web tools to analyse publicly available microarray data[21,22] and although our approach was limited to named genes, the automatic integration of background knowledge made PubGene a complement to conventional clustering analysis.

Results

Automated indexing of named human genes

We collected publicly available nomenclature information for human genes from the HUGO Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/), LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), the Genome Database (http://www.gdb.org/) and GENATLAS (http://www.citi2.fr/GENATLAS/). The resulting gene nomenclature database

Departments of [1]Computer and Information Science, and [2]Physiology and Biomedical Engineering, Norwegian University of Science and Technology, Trondheim, Norway. [3]Department of Tumour Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo, Norway. [4]Department of Computer and Information Science, Linköping University, Sweden. Correspondence should be addressed to E.H. (e-mail: ehovig@radium.uio.no).

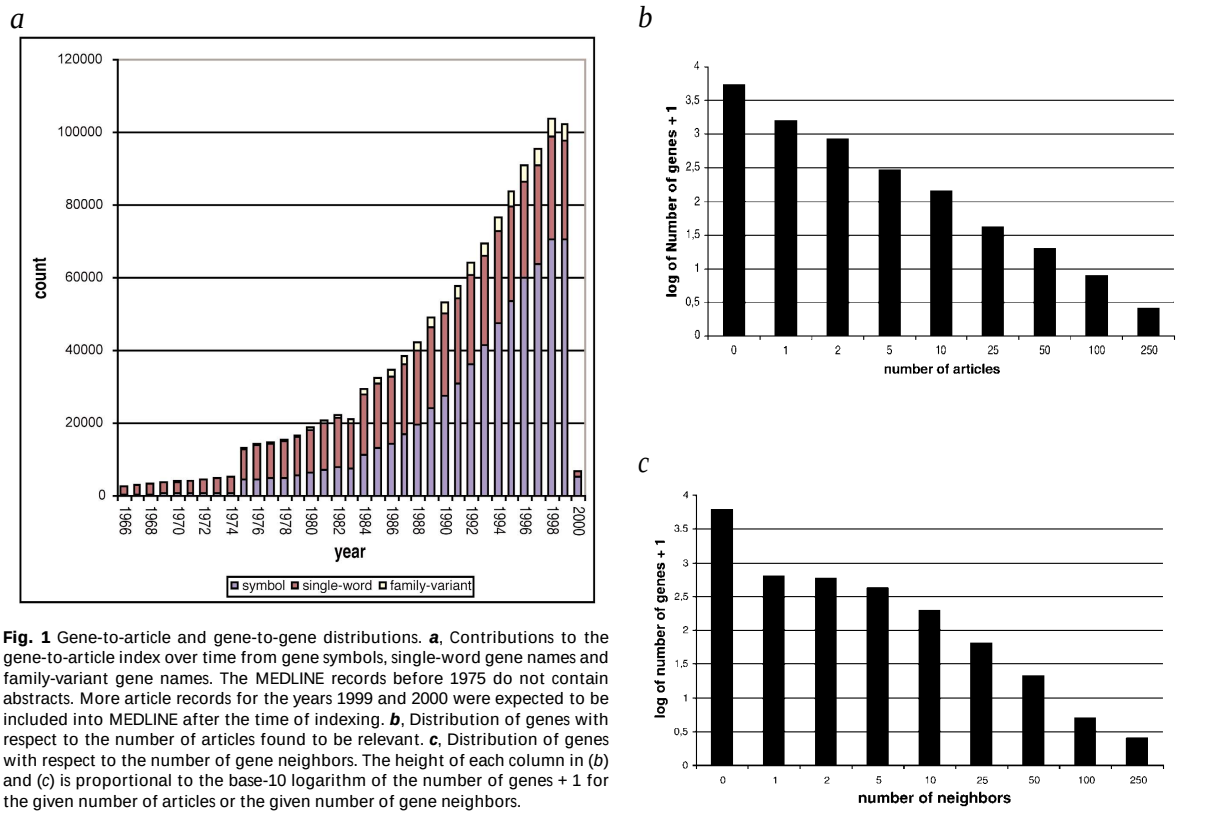


Fig. 1 Gene-to-article and gene-to-gene distributions. ***a***, Contributions to the gene-to-article index over time from gene symbols, single-word gene names and family-variant gene names. The MEDLINE records before 1975 do not contain abstracts. More article records for the years 1999 and 2000 were expected to be included into MEDLINE after the time of indexing. ***b***, Distribution of genes with respect to the number of articles found to be relevant. ***c***, Distribution of genes with respect to the number of gene neighbors. The height of each column in (*b*) and (*c*) is proportional to the base-10 logarithm of the number of genes + 1 for the given number of articles or the given number of gene neighbors.

contained 13,712 genes, with each gene identified by a primary symbol. We obtained 9,722 primary gene symbols from the HUGO database, and 2,729, 1,239 and 358 additional (primary) gene symbols from the LocusLink, GENATLAS and Genome Database databases, respectively. Because many of the official symbols defined by the HUGO Nomenclature Committee have been recently introduced, many authors had not referred to genes by their official gene symbol. Thus, in addition to the primary symbols, we collected gene names and alternative symbols. The standardization of gene nomenclature has also resulted in a number of previously official gene symbols being withdrawn (that is, their status as primary gene symbols has been revoked). Of the gene symbols found as primary symbols in LocusLink, GENATLAS and the Genome Database, 63, 63 and 352, respectively, had been withdrawn by the HUGO Nomenclature Committee.

We found gene occurrences by searching for gene symbols and short gene names. Occurrences of gene symbols and short gene names were mapped to primary gene symbols through the nomenclature information. The short names were names consisting of a single word (for example, insulin) or names of the form 'family variant', in which the family stem consisted of a single word (for example, cyclin E2). We analyzed 10,125,978 MEDLINE records from 1966 to the present and found 1 or more gene symbols in 1,964,717 (19.4%) of them. Counting each symbol found in one or more places in a record as one occurrence resulted in 3,534,061 gene-symbol occurrences. Many symbols have been used to refer to more than 1 gene and, of the 24,443 symbols in our database, 2,796 were ambiguous in the sense of having associations to multiple genes (for example, *ALR*, *MTS1* and *PBP*). We treated each symbol occurrence as a possible occurrence of any of the genes for which the symbol had been listed as a primary symbol or an alternative symbol. This gave 8,920,666 putative gene occurrences. Some gene symbols coincided with common abbreviations in other contexts (for example, *II*, *IV* and *ABO*). After mapping occurrences of gene symbols to genes, only 885,146 gene occurrences based on symbols remained. We mapped occurrences of short gene names into occurrences of the corresponding genes. Fig. 1*a* illustrates the evolution of the total number of gene occurrences the three gene-term sources contributed to the total gene-article index.

Literature co-occurrence associates biologically related genes

After constructing the gene-article index, we used it to compute a network of genes by linking two genes if they occurred in the same article. Graphically, we represented each gene in the database by a node in the network and created a connecting link between every pair of genes that co-occurred (Fig. 1*b*). As an indication of strength, we gave each pair of genes a weight equal to the number of articles in which the pair was found (Fig. 1*c*). The network contained 139,756 pairs of such related genes, with a total occurrence weight of 1,087,757. Among the 13,712 genes, 7,512 had one or more neighbors, and 710 genes had literature references but no neighbors. Of the 5,490 genes that were not found in any articles, 5,202 genes had a status of 'reserved' or 'provisional'.

To examine the extent and nature of gene pairs being over-represented or incorrectly assigned in the PubGene network, we carried out a randomized experiment by drawing 500 gene pairs with a weight of 1, and 500 pairs with a weight of 5 or more. The 1,000 gene pairs were classified into 7 categories (Table 1). Six categories covered pairs with meaningful biological relationships and one category, labeled 'incorrect', covered pairs for which no relationship was found or for which it was obvious that the association was incorrect. The proportions of incorrect pairs were 40% and 29% for the low-weight and high-weight categories,

Table 1 • Types of gene relationships found in PubGene

	Count	
Relationship	W1	W5+
cell biology	43	24
expression correlation	151	183
histology	22	66
homology	29	75
chromosome mapping	53	6
other	4	5
incorrect	198	141

We randomly selected 1,000 pairs of genes and manually analyzed these to obtain an assessment of what kinds of biological relationships are reflected and what types of errors are being made by connecting genes by co-occurrence of gene terms. We randomly selected 500 pairs from pairs with weight 1 (W1) (that is, pairs of genes that had been found in exactly one article) and also randomly selected the other 500 pairs from pairs with weight greater than or equal to 5 (W5+). The proportions of correct links in the two categories were 60% and 72% for W1 and W5+, respectively. Pairs for which no relationship was found or the link was obviously incorrect were analyzed further (Table 2).

respectively. The distributions of errors are shown (Table 2). There were essentially three types of errors in the higher weight group: symbols belonging to more than one primary gene symbol, very general symbols coinciding with general acronyms, and very short gene names. In contrast, in the low-weight group, we observed a number of different types of errors, with a significant contribution from general symbols.

To examine the extent and nature of under-represented gene pairs in the PubGene network, we extracted information from the Database of Interacting Proteins (DIP; http://dip.doe-mbi.ucla.edu/). The DIP contained 169 human pairs of interacting proteins recorded from articles published in peer-reviewed journals within the time span covered by PubGene. We mapped the 171 proteins included in the 169 protein pairs to corresponding genes in PubGene, giving 169 pairs of human genes (Table 3). PubGene contained 51% of the DIP pairs, a more than sixfold improvement over a random experiment (assuming sampling without replacement of an equal number of pairs as found in PubGene from all possible pairs over the 171 genes). The DIP-derived gene pairs not detected by PubGene were further analyzed to find the main reasons for under-representation of gene pairs. A total of 49 references covered all missed interactions in the DIP database (Table 4).

The predominant problems were due to three main integrity issues: insufficient synonym lists, synonym case variation, and complex gene families with immature or complex naming conventions. These problems mainly caused a reduction in true gene pairs, and did not create 'noise'. As DIP is based on protein names not necessarily included in the PubGene name lists, it is expected that synonym problems will be elevated in such a comparison. As very few interactions are missed due to lack of citation in the title or abstract, however, an even better representation is expected from optimizing the indexing procedure.

To examine PubGene performance in comparison with a very rich independent information source, we selected from the OMIM database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) all genes that have been mapped to a chromosomal location. For each of these, we analyzed the textual description in OMIM to locate OMIM numbers for other mapped genes. This resulted in 19,157 ordered pairs of genes. PubGene correctly incorporated 45% of the OMIM pairs, an 88-fold increase over random sampling (Table 3). OMIM record information content varies depending on the editor. In some records, the information is biased towards gene mapping and diseases, and will therefore include gene associations that are not recorded in PubGene, whereas in others very little information is offered.

It is difficult to assess precisely the expectations from these experiments, as the manually curated databases only contain subsets of the PubGene data not necessarily reflected in titles or abstracts of MEDLINE records. But the numbers of interactions in DIP and OMIM contained in PubGene reflect that PubGene captures substantial amounts of the existing biological information on protein-protein interactions and on gene mapping and disease.

Literature associations highlight background knowledge for signature genes in patient sample data

The microarray data set of Alizadeh *et al.*[21] contains mRNA expression measurements across 96 normal and malignant lymphocyte samples, and represents the most informative and richest data set on human material so far published. The authors presented an analysis based on unsupervised hierarchical clustering analysis of the data. They proposed the term 'signature gene cluster' as an operational definition to indicate genes that are coordinately expressed and thus cluster together. Of the six identified signatures thought to characterize distinct cell types or biological processes, the germinal center B-cell (GC-B) signature highlights a main finding using the hierarchical clustering technique: germinal center (GC) B-cells represent a specific stage of B-cell differentiation, distinct from activation of blood B-cells. Using the GC-B signature and the signature of activated B-cells, diffuse large B-cell lymphomas were divided into two subgroups having features of either germinal center B-cells or activated blood B-cells; these subgroups were also distinguished by clinical outcome.

To explore the correlation between unsupervised clustering and the supervised PubGene approach in a large, biologically relevant data set, we extracted the data for tonsil GC B-cells and activated B-cells isolated from healthy individuals. The publicly available data subset contains measurements on 4,026 clone spots that can be mapped, through the IMAGE clone-IDs, to 1,302 named genes in PubGene. For the two cell types, we calculated the mean log-ratio for each gene across the samples in each group and submitted the group differences to the expression analysis tool. The cell-type mean value represented an extra layer of abstraction from the original data, and was obtained to highlight group-specific biological differences between the two cell types. The 50 genes corresponding to the 50 most up- or downregulated literature sub-networks included 7 (28%) of the 25 named genes in the GC-B signature (*BCL6*, *BMP7*, *CD24*, *CD38*, *E2F5*, *MME* and *MYBL1*). This is a sevenfold increase compared with a random experiment (assuming randomly sampling 50 genes without replacement from all 1,302). Moreover, 39 of 50 genes identified by PubGene were cluster designations (CDs; data not shown), which are surface cell markers generally used to define

Table 2• Categories of incorrectly linked pairs in PubGene

	Count	
Category	W1	W5+
symbol, other gene	19	40
symbol, gene other species	8	3
symbol, cell line	20	4
symbol, other biomedical concept	15	1
symbol, general	82	43
symbol, other	20	4
short name	34	46
total number of incorrect links	198	141

The two groups of pairs were the same as in Table 1. For each group, we analyzed and classified reasons for incorrect links. "Symbol, other gene" denotes incorrect links due to a symbol associated with several genes. "Symbol, gene other species" denotes incorrect links due to a symbol that had been used in another species. "Symbol, cell line" denotes symbols that had also been used as names for cell lines, and "symbol, other biomedical concept" covers other biomedical concepts (for example, diseases or drugs). The category "symbol, general" includes symbols that had been used in a number of contexts (for example, P1 had been used to refer to postnatal day 1). "Symbol, other" includes other incorrect links caused by symbol confusion that are not in any of the above categories. "Short name" denotes incorrect links caused by short gene names that have other uses (for example, in most of the manually checked abstracts medulloblastoma had been used to refer to the tumor type and not the gene).

Table 3 • Comparison of PubGene with manually curated databases

	Database	
	DIP	OMIM
Number of genes	171	6,404
Number of actual links	169	19,157
Number of possible links	14,535	37,350,432
Number of actual links found in PubGene	86	8,585
Number of all links found in PubGene	1,052	187,226
Number of expected actual links in PubGene	< 13	< 97
p value	<E–10[a]	< E–7[b]
Improvement over random	> 6	> 88

[a]p(x≥86)=1–p(x<86)<1–p(x≤39)<1.1E–11.

[b]p(x≥8585)=1–p(x<8585)<1–p(x≤160)<7.4E–8.

Signal-to-noise ratio as assessed by comparison with the manually curated databases DIP and OMIM is shown. Links in DIP were found by selecting pairs of interacting proteins in which both proteins were human (determined by the SWISS-PROT ID; http://www.expasy.ch/sprot/sprot-top.html). Participating proteins were linked to human genes through information in SWISSPROT, OMIM and the PubGene nomenclature compilation. As DIP includes self-interactions, we included 'pairs' with the same gene, when comparing against the DIP database. Links in OMIM were found by analyzing the text part of the OMIM record for each mapped gene in OMIM. Assuming a process of sampling without replacement, we calculated the number of actual links from DIP and OMIM expected to be found in a random sample with the same number of pairs as the number of links in PubGene.

lymphocyte differentiation stages. The products of these genes are known to be significant regulators of lymphocyte functions. Using the Alizadeh web analysis tool (http://llmpp.nih.gov/lymphoma/index.shtml), we were able to link 21 of 50 top-ranked genes to one of the B-cell signatures found by hierarchical clustering. The distribution of the 21 genes among the 3 B-cell signatures (Table 5) illustrates the correlation between the supervised PubGene approach and the unsupervised clustering approach.

To estimate the extent of complementarity of the two approaches, we characterized the biological basis of the GC-B signature using the PubGene tools. We started with the 25 named genes found using hierarchical clustering by Alizadeh *et al.* and extracted a network from PubGene based on the named GC-B signature genes. Of the 25 signature genes, 5 did not have neighbors in PubGene, and therefore would not be visible in a network. The remainder and the most important literature neighbors (among the 1,302 genes with expression data) were all connected (Fig. 2*a*), thus displaying graphically an underlying biological relationship between these genes. We then linked the signature gene list to disease MeSH terms to search for diseases associated with the signature genes. Among the top-ranked terms were those related to Fragile X and Angelman syndromes, lymphoma, leukaemia and tuberculosis (data not shown). Fragile X and Angelman syndromes ranked ahead of lymphoma, because *FMR2* and *HERC2* are loci associated with these syndromes. *FMR2* is downregulated by repeat expansion and methylation[23], and its protein product shows similarity to that of *MLLT2*, which is involved in translocations found in acute lymphoblastic leukemia cells. Moreover, we noted that *MLL* was also listed as a close neighbor of the signature gene *MME* (CD10). *MLL* is not a signature gene (that is, it was not found by the clustering analysis), but is distinctly upregulated in tonsil GC B cells, and has been found translocated to a number of genes, including *FMR2*. Transcription was the most significant ontology term obtained by the genes in the GC-B signature, indicating that a large number of the signature genes are transcriptional regulators.

PubGene rapidly focused the extracted biological attention for the GC-B signature genes towards central GC B-cell processes. *DNTT* is upregulated as a signature gene. This gene is involved in normal V(D)J immunoglobulin recombination in B-cells. A number of the genes identified tend to be translocated in lymphomas (*FMR2*, *MLL*, *BCL6* and *BCL7A*), as disclosed by the MeSH terms and their neighboring positions in expression networks. All of these genes are upregulated in GC-B cells (Fig. 2*a*), most likely a reflection of hypermutability being a process of immunoglobulin variation. Current models of immunoglobulin variation suggest that both the recombination and hypermutability processes take place in the germinal center[24,25]. The fact that these genes were solely identified by the PubGene approach demonstrates that the current PubGene index may be used to identify gene networks not identified by clustering and to classify genes according to biological processes. In parallel to our analyses, it was confirmed that the process of hypermutability is a major discriminant between the germinal center versus activated B cell-like lymphomas[26], as an ongoing somatic mutation process was shown in the lymphomas of the GC-B type.

Detection of complex coregulatory patterns between biologically related genes

The gene-expression data set of Iyer *et al.*[22] contains 8,613 mRNA measurements over 12 time points. We used a publicly available data subset containing 517 clones, selected from genes whose transcription levels changed substantially after serum stimulation of human fibroblasts. Using the IMAGE clone-IDs, we were able to map expression data to 340 named human genes. To examine the biological relationships between similarly upregulated genes, we used PubGene to identify literature associations between the 340 genes and ranked sub-networks according to their content of highly upregulated genes.

The highest-scoring network at the time point of one hour (Fig. 2*b*,*c*) depicts the upregulation of the typical immediate early genes *FOS*, *JUNB* and *EGR1*, which all encode transcription factors involved in the cellular response to mitogenic stimuli[27]. Four genes associated with these transcription factors by literature co-citation were also strongly upregulated at this early stage of the fibroblast serum response (*IL6*, *PDGFRB*, *FGF7* and *SERPINE1*). Superimposing expression levels from time point eight hours onto the same gene network revealed that the genes encoding the angiogenesis-promoting factors *FGF2* and *VEGF* (ref. 28) followed a more delayed course of gene induction. Moreover, at eight hours, *EGR1* transcript levels were well below the levels at time point zero (67% of levels at 0H). Also the levels of *FOS* and *JUNB* had considerably decreased relative to the one hour levels. Activation of *FOS* and *JUNB* had also considerably decreased relative to one hour. Although still 123% and 185% of zero hour levels, transcript levels were only 9.7% and 29% of one-hour levels for *FOS* and *JUNB*, respectively. This illustrates how PubGene can visualize complex co-regulatory patterns of gene expression and simultaneously highlight biological relationships relevant for these patterns.

Extending the analysis for time point 1H showed that, among the ten highest-scoring networks, there were several other networks similar to the network shown in Fig. 2*b*,*c*. In addition, we found networks containing the upregulated genes encoding

Table 4 • Reasons for under-representation of DIP derived gene pairs

Reason	Number of articles
missing in abstract	5
symbol case mismatch	22
missing alternative symbol	24
complex protein (gene) family name	9
other	5
sum	65

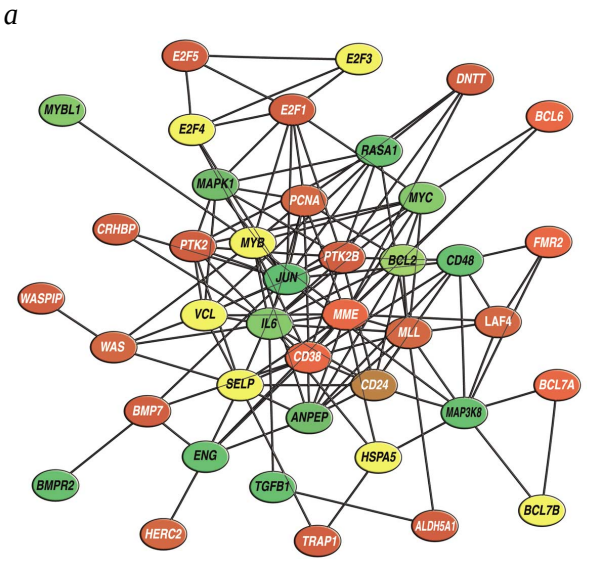
The MEDLINE records of articles referred to in DIP as references for the DIP pairs not found in PubGene were manually examined to determine the causes for this under-representation. One or more reasons were assigned to each article according to what caused the omission of the gene pair corresponding to the protein pair documented in DIP. Examples of proteins (genes) with complex names are, for instance, 14-3-3 family, G-coupled protein receptors and integrins.

nuclear receptors NR4A2 and NR4A3 (two members of the steroid/thyroid hormone family and presently with less precisely defined functions), and a network showing that *DUSP1*, which is involved in signal transduction, is upregulated at one hour. The latter network revealed literature connections from *DUSP1* to the cell-cycle regulatory genes *CCNA2*, *CCND1* and *CDKN1A*, whose gene expression levels are still unchanged at this time point.

The most upregulated literature network at time point six hours (Fig. 2*d*,*e*) contained genes largely classified as cytokines, growth factors and hormones. Using PubGene to look up relevant MeSH-terms for the 12 genes (Table 6), we found that 'angiogenesis' co-occurred with the highest fraction of genes (10/12). These data (Table 6) are similar to those obtained by Iyer *et al.*[22], who used background knowledge to assign biological function to the genes involved in the fibroblast serum response. Our indexing strategy permits rapid profiling of genes through the distribution of MeSH terms, as well as identification of strong associations between genes and biological processes. In effect, this represents a quantitative approach to bringing background knowledge into gene-expression analysis.

Discussion

Several linguistic and statistical methods have been applied to information retrieval and information extraction in biomedicine. Parsing, tagging of parts of speech and estimation of keyword distributions are computationally expensive compared with term recognition. By choosing a simplistic method of computing a network from term co-occurrences, we aimed to create a literature-wide as well as a genome-wide view of the current knowledge about human genes. By not attempting to detect directly occurrences of particular types of relationships, thus prioritizing perspective over detail, it was possible to obtain a global view of the literature of the human genome. To our knowledge, there have been no other reports on completed work on a comparable scale. Stapley and Benoit[19] also extracted a gene network based on term co-occurrence, but they created a network of genes



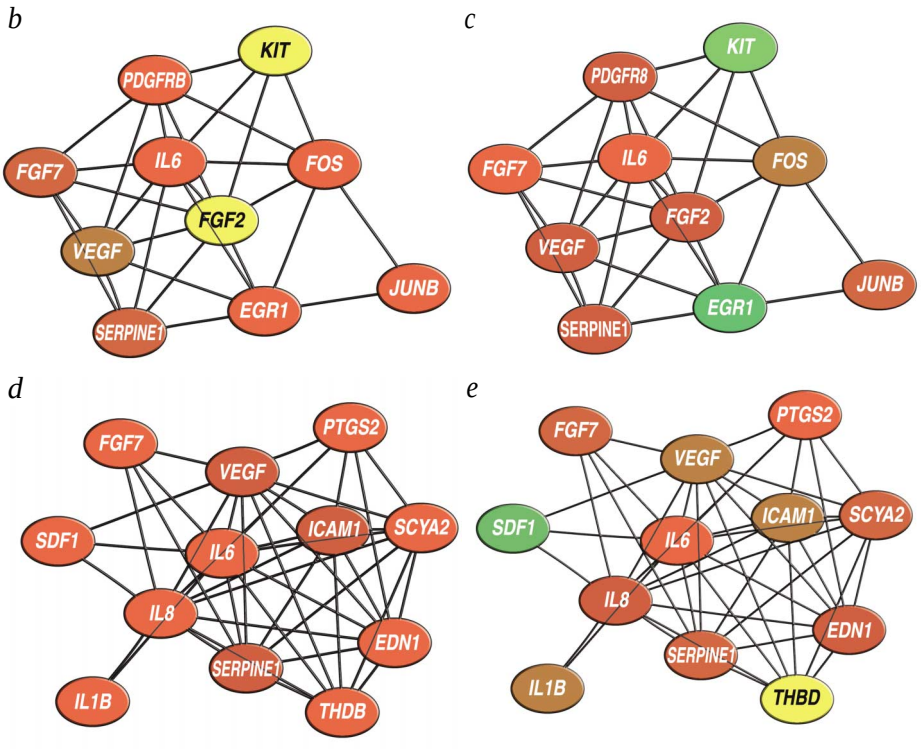


Fig. 2 Literature networks of genes found relevant in gene expression data analysis. We submitted gene symbols and corresponding expression data to PubGene to identify co-expressed groups of genes associated in the literature. Lines connect genes that have co-occurred in one or more articles. Annotations reflecting the number of literature co-citations have been suppressed for readability. ***a***, Network of the genes in the GC-B signature. Colors represent relative expression values when comparing tonsil GC B-cells with activated blood B-cells. Yellow reflects no difference between the two groups, whereas red/green reflects genes that are more/less expressed in tonsil GC B-cells than in activated blood B-cells. ***b***,***c***, Literature network of genes highly upregulated at time point one hour (1H) in the fibroblast serum response. ***d***,***e***, Literature network of genes highly upregulated at time point 6 h (6H) in the fibroblast serum response. We used PubGene to score sub-networks with the closest neighbors of each gene based on gene-expression levels ranking the networks according to their content of highly upregulated genes. At 1H, a three-gene network containing *JUNB*, *EGR1* and *FOS* was ranked as the most upregulated one. An expanded version, including the literature neighbors of these genes, is shown color-coded using expression levels from 1H (*b*) and expression levels from 8H (*c*). *JUNB*, *EGR1* and *FOS* have lower expression levels at 8H. The most upregulated network at 6H contained the neighborhood around *IL8*. This network contained 16 genes, of which the 12 most important in terms of upregulation and literature proximity to *IL8* are shown (*d*,*e*). *d*, The 6H-network color-coded with expression levels at 6H. *e*, The same network color-coded with expression levels at 1H. *b*–*e*, The gene-expression data encoded in the colors were ratios at the given time point relative to serum-starved, growth-arrested fibroblasts used as the reference (time point 0H). Strong up/downregulation is shown as bright red/green color, whereas genes with unchanged expression levels are yellow. *a*, GBC signature network. *b*, 1H network at 1H. *c*, 1H network at 8H. *d*, 6H network at 6H. *e*, 6H network at 1H.

Table 5 • B-cell signature cluster associations for the 50 highest ranked genes according to PubGene expression analysis

Signature	Number of genes all	member	similar
activated peripheral B-cell	37 (87)	2	3
resting peripheral B-cell	21 (81)	2	9
germinal center B-cell	25 (112)	3	2
sum	83 (270)	7	14

For each of the 50 genes, we investigated associations to the 3 B-cell signatures using similarity of the clustering patterns of the 20 genes having most similar expression patterns (using Pearson correlation as given by the analysis tool of Alizadeh *et al.*[21]). The 'all' column shows the number of genes in the respective signatures, with the number of clones in parentheses. The 'member' and 'similar' columns show the number of genes among the 50 highest ranked genes that could be classified as a member of a B-cell signature, or whose majority of most closely related genes (clones) were from the given B-cell signature. A large number of the genes could in this way be associated with one of the B-cell signatures (42%), as opposed to 6% by pure chance, reflecting the B-cell nature of the samples being submitted.

for yeast extracted from 2,524 MEDLINE documents chosen on the basis of being from 1997 or 1998 and containing the MeSH term '*Saccharomyces cerevisiae*'.

There are several limitations to our methodology. The use of MEDLINE records restricts the relationships that can be found to those mentioned in titles and abstracts. The dependency of the underlying text material is well illustrated by the increase in the number of relevant articles from 1974 to 1975, when abstracts were first included. Nevertheless, the use of MEDLINE records has the advantage of making an explicit basis for defining co-occurrence. A principal question in this respect is whether correctly determined co-occurrences in the title and abstract precisely reflect meaningful relationships between genes. Based on the 1,000 pairs we investigated, our answer, albeit preliminary, is 'yes'. This is because all incorrect pairs were explained by synonym or name confusion, and not because the gene terms referred to genes where no relationship was mentioned in the text. This conclusion was also supported by Stapley and Benoit[19].

Although our experiments support the premise that abstract co-occurrence reflects meaningful biology, they also demonstrate problems generated by ambiguous gene names and symbols. The use of a thesaurus for information retrieval has been debated[29]. An advantage is that it may increase recall (the percentage of all correct occurrences that are detected), as it is not necessary to explicitly mention all alternative names. Conversely, ambiguities arise when gene terms are associated with several genes in the nomenclature database (thesaurus) and when a gene term has been used in other contexts. This may degrade precision (the percentage of detected occurrences that are correct). Manually editing (parts of) the thesaurus and resolving ambiguities may reduce this problem. Given that a large proportion of the incorrect links were caused by a relatively small number of problematic terms (data not shown), we expect that major improvements can be made by manually verifying and, if necessary, deleting some of these terms from synonym lists. We believe this may provide a level of precision similar to that of a completely manual process, but at a much lower cost. Our strategy to resolve symbol ambiguities was to accept symbol occurrences as gene occurrences only when one (or more) of the words from at least one of the names of the gene occurs with the symbol. Our experiments indicated that this was not sufficiently stringent. Improving the precision and recall of the indexing procedure is an ongoing effort, but as perfect accuracy (perfect precision and perfect recall) is presently beyond our reach, we are left with a compromise between precision and recall. So far, we have prioritized recall, as it is easier for the user to relate to noise that is present than to ponder relationships that are missing.

Stapley and Benoit argued that gene terms (symbols), due to their specificity, are superior to natural language searching for information retrieval[19]. Our choice of using gene symbols and short gene names was based on efficiency, and in our experience gene terms are not as specific; this may in part be attributed to organism-specific nomenclatures. We have noted a long list of gene symbols that are poorly designed with respect to information retrieval (for example, 'II', 'IV', 'KD', 'SD', 'AS', 'A1' and 'ABO'). It would be beneficial if specificity were considered in the design of future official symbols, but most symbols enter the literature as a convenient abbreviation of a full name associated with a gene[30]. Moreover, such considerations would only benefit information retrieval in new literature and provided the official nomenclature is actually used. More elaborate indexing procedures may raise precision, but most likely would be at the expense of efficiency, and, because the gene nomenclature is rapidly changing, it is critical to be able to rebuild the indices periodically. One method to improve the filtering procedure may be to estimate a word distribution for MEDLINE records relevant to (human) genetics and simply eliminate gene term occurrences from articles with deviating word distributions.

The idea of estimating a word distribution, as well as the overall aim of our work, is related to the work of Shatkay *et al.*[20]. They used document similarity based on estimated word distributions to link genes through a set of so-called 'kernel' documents. Using one kernel document per gene, they first identified a core of documents containing the union of the sets of the 50 documents most similar to any kernel document. This core was then trimmed to contain

Table 6 • MeSH terms associated with selected genes

MeSH Term	Gene *IL8*	*IL6*	*SCYA2*	*ICAM1*	*VEGF*	*EDN1*	*SERPINE1*	*PTGS2*	*THBD*	*FGF7*	*SDF1*	*IL1B*
blood coagulation	4	–	–	12	1	4	3	–	94	50	–	–
chemotaxis	331	57	173	79	11	6	1	–	–	–	26	–
fibrinolysis	2	5	1	1	1	2	294	–	44	–	–	–
hemostasis	–	4	1	1	1	3	57	–	35	–	–	–
inflammation	45	76	7	37	5	3	7	3	2	–	–	–
angiogenesis	44	21	1	24	740	14	35	6	3	5	–	–
wound healing	3	17	4	2	11	3	8	1	1	4	–	–
gene total articles	2,932	11,149	1,048	5,084	1,858	3,985	2,359	866	1,254	348	104	31

Using the gene to MeSH term index in PubGene, we immediately obtained an overview of the degree of association between the genes in Fig. 2*d*,*e* and important biological processes relevant for the fibroblast serum response. Each cell in the table contains the number of articles relating the gene in the given column with the MeSH term in the given row. Note that the process 'blood coagulation' was searched for using the MeSH terms 'blood coagulation factor inhibitor' and 'blood coagulation factor inhibitor blood' in addition to 'blood coagulation', and that 'angiogenesis' was searched for with the term 'neovascularization'. The bottom row shows the total number of articles found in the PubGene index for each gene. The columns are arranged in the order of decreasing strength of association with *IL8*, which was the defining gene for the network shown in Fig. 2*d*,*e*. The selection of MeSH terms was based on the report of Iyer *et al.*[22]. This analysis, however, can easily be adapted to find the most relevant MeSH terms for an arbitrary set of genes without relying on prior knowledge. It is obvious from the table that *IL1B* is heavily under-represented in the Pubgene index, hence also its association to *IL8* and the MeSH terms. Being one of the first variants of interleukin discovered, *IL1B* would be expected to have more references than both *IL6* and *IL8*. The explanation can be found in terms of naming conventions. Indeed, the number of Medline hits for 'interleukin 1 beta' exceeds 20 thousand, whereas the primary symbol *IL1B* is found in less than 20 for the subset of Medline relevant for comparison.

only documents found for at least two kernel documents. Genes were then linked if their kernel documents had similar sets of related documents in this core set, and the intersecting documents were used to annotate the link. This approach surpasses the problems with gene-term ambiguity, but has two problems related to scalability. The first is that of choosing the best kernel document for each gene. To be most useful, the kernel document should be descriptive of the gene function, and it is likely that the most descriptive document will change over time. The second problem is that density estimation is computationally expensive.

In the long run, term co-occurrence detection should be considered a temporary solution and, as we expect linguistic and statistical methods be improved in robustness and efficiency, the overview map represented by PubGene can be refined and enhanced. A number of more specialized tools based on the PubGene concept may be envisaged. For example, by combining predefined keywords (such as disease states or chemical agents) with the gene index in the same way as MeSH terms were linked to genes, composition of specialized microarrays may be largely automated. Integration with unsupervised methods for gene expression analysis is also a priority. The feasibility as well as the applicability is suggested by our analyses, and the combination is expected to result in increasingly user-acceptable interfaces to the evermore-increasing complexity of biological knowledge. Linking literature information to sequence analysis may develop this even further. The large number of interactions amenable to automated extraction will increase with the increased efforts of post-genomic molecular research and with increased availability through, for example, full-text open repositories. PubGene, with its present indexing strategy, rich and varied information content and analytical flexibility, can incorporate more of the available biological knowledge for high-throughput gene-expression analysis than any other analytical tool available.

Methods

The PubGene gene database. We downloaded gene symbols and names for human genes and merged the information into a list of unique human genes, each identified by a primary symbol. For each official gene symbol in the HUGO NC database, we created a gene record and added literature aliases. We added additional gene records for primary symbols from LocusLink, GDB and GENATLAS (when these were not found with the status 'withdrawn' in HUGO).

The gene-article index. We processed the MEDLINE records year for year. For each article record, we extracted occurrences of genes by finding gene symbols and short gene names. We first split the text into separate words, whereupon each word was compared against a list of symbols or names. We required gene symbols to match with correct case, whereas we allowed gene names to match regardless of case. We found occurrences of 'family variant' names by first looking for occurrences of the family stem word and then looking for a variant designation immediately before or after. To check if a word was a symbol, a one-word gene name or a family stem, we used a hash table. Becasue hash tables can be implemented with O(1) access time (see Cormen *et al.*[31] for an explanation of big-O notation), all symbol and name occurrences can be found in O(n) time, where n represents the length of the text (for example, the number of article records if a maximum combined length on titles and abstracts is assumed).

As a preliminary step, we considered each occurrence of a gene symbol as a potential occurrence of all of the genes to which the symbol had been associated. To resolve ambiguities we implemented a filtering procedure to remove incorrect associations between genes and articles. For each gene, the filtering procedure looked at the articles for each of the listed symbols. If the symbol was longer than l_1=4 characters and was found in less than k=10 articles (within the particular year being processed), all articles for this symbol were kept for the gene. Otherwise, the procedure checked each article record to see if the title or abstract mentioned any words from any of the long names of the gene. The procedure kept articles associated with symbols longer than l_2=2 that also contained w=1 word from at least 1 of the long names. If the symbol was only two characters long, it was required that the title or abstract contained more than w words from at least one of the long names. l_1, l_2, k and w are parameters. The given numbers reflect default settings. For each gene-article association, the procedure checked, at most, all the words in all the names of the gene. Assuming that the number of names associated with a gene is less than some constant and that the length of each name is also less than some constant, each gene-article association can be checked in O(1) time by precomputing the association between words and articles. Moreover, if we assume that there is a fixed upper limit of the number of genes that need to be checked for each article, the complexity of the filtering procedure is O(n+m), where n is the number of articles and m is the number of genes.

We mapped occurrences of gene names to occurrences of genes through the map between symbols and genes. That is, each name was associated with a symbol linked to one or more genes. We interpreted all occurrences of short gene names as occurrences of all thus-associated genes.

The implemented procedure has complexity O(n+m), where n is the length of the analyzed text and m is the size of the gene database to be indexed. In other words, the procedure is linear in the size of the data. As an indication of the actual processing time, on a UNIX server (multi-processor server running SunOS 5.8 with 4×400 MHz UltraSPARC-II processors) concurrently running other processes, the year 1998 was indexed in less than 5 hours. The actual processing time depends on the system load.

The gene-gene network and gene-term maps. We found links between genes by associating genes that co-occurred in an article record. For each gene, the algorithm considered all articles for this gene, and, for each article, found all other genes that also occurred in this article. For each other gene that was also linked to the article, the algorithm created a link of weight 1 or added 1 to the weight of a pre-existing link.

We also constructed the gene-term relations by linking genes to terms through a common set of articles. We linked every gene to any term co-associated to a common article. The MeSH term index links MeSH terms to articles and was obtained from NLM, whereas we created the GO ontology-term index and the disease index with the same software used to index gene symbols. Assuming the use of O(1) lookup hash tables for the relations between genes and articles, terms and articles and the inverses, the algorithm for constructing gene-gene and gene-term relations has complexity O(n*m), where n is the number of genes (or maximum of number of genes and number of terms) and m is the average number of articles associated to each gene (or term).

The gene network browser. Given a particular query gene, the network browser extracts its neighborhood in the whole network and creates a small sub-network with the most important neighbors of the gene. The graphs are generated dynamically in real time with adjustable parameters determining the size of the gene neighborhood to be shown. The extraction algorithm is a breadth-first search, prioritizing links with higher weight. The Neato program in the Graphviz software (AT&T; http://www.research.att.com/sw/tools/graphviz/) is used to create the two-dimensional layout. The extraction algorithm and the Neato layout-algorithm are both of complexity O(n^2), where n is the size of the graph.

The gene expression and literature score. Gene-expression data are superimposed on the literature network to compute an expression score for each gene in the assay. The input format of the data is a list of pairs (g_i, v_i), where g_i is a gene (symbol) and v_i is an expression measurement (for example, a log of a ratio or the difference between two log-ratios if comparing two conditions). This score is computed by extracting a literature neighborhood for the gene, similarly as in the network browser, and calculating a score from the expression data of the genes in the neighborhood. The score is then computed as

$$f\left(\tfrac{1}{n}\sum_i h(x_i)\right)$$

where f is the absolute value function or the identity function, n is the number of genes, h is absolute value or identity (typically, at most one of f and h is chosen to be the absolute value function), and x_i is an expression

value or the average expression value of two genes connected by a link in the neighborhood cluster. The score parameters can thus be set to prioritize up-/downregulation, concomitant up- and downregulation, or no change, and if summation is carried out over links in the network, densely linked neighborhoods can be prioritized. For instance, to find the most up- or downregulated cluster (not concomitant) one would use the expression,

$$abs(\tfrac{1}{n}\sum_i x_i)$$

that is, find clusters with high absolute value of the mean value. After ranking the clusters by score, the tool generates images for the highest scoring clusters. The scoring algorithm has complexity $O(n^2)$, where n is the size of the score-neighborhood. This algorithm is repeated for each gene in the expression data set. Thus, if N is the number of genes, the whole computation is $O(N \log N + Nn^2)$, assuming that sorting the N scores requires $O(N \log N)$ time.

The gene-term association strength. Given a set of genes, the weighted relation between genes and terms can be used to compute term-relevance for individual terms. For a given term, the score with respect to a set of genes is calculated by combining scores from individual genes. The strength of association between a term and a gene can be computed in absolute terms as the number of co-associations, or in relative terms as the number of co-associations for the term divided by the number of co-associations of the highest scoring term for that gene. For each term, association strengths from each gene can be combined by addition or multiplication to derive a score of the term relative to the set of genes.

Symbol lookup. In the PubGene database, genes are identified and accessed by their primary symbol. Given the abundance of synonyms in common use, we created a nomenclature lookup tool that can be used to find official symbols for genes having information matching a given query. The query is matched against the nomenclature database using the regular expression matching facility of Perl to find all occurrences of the query. The query can be used to search among synonyms, long names, chromosome locations, UniGene clusters or any subset of these fields.

Acknowledgments

We thank the National Library of Medicine for access to MEDLINE; D. Tjeldvoll for help with installation and use of the Graphviz software, and contributions to programming on early versions of the web-interface; H.-C. Aasheim and Ø. Fodstad for discussions; and S. Bade, W.P. Kuo, S. Vinterbo and D. Warren for comments on the manuscript. This work was supported in part by grants from the Norwegian Cancer Society. T.K.J. was supported by grant 134422/410 from the Norwegian Research Council.

Received 5 July 2000; accepted 26 March 2001.

1. Adams, M.D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195 (2000).
2. The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–2018 (1998); errata: **283**, 35 (1999); **283**, 2103 (1999); **285**, 1493 (1999).
3. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
4. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
5. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
6. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
7. Andrade, M.A. & Bork, P. Automated extraction of information in molecular biology. *FEBS Lett.* **476**, 12–17 (2000).
8. Blaschke, C., Andrade, M.A., Ouzounis, C. & Valencia, A. Automatic extraction of biological information from scientific text: protein-protein interactions. in *Intelligent Systems for Molecular Biology* 60–67 (AAAI Press, Heidelberg, 1999).
9. Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* **5**, 541–552 (2000).
10. Humphreys, K., Demetriou, G. & Gaizauskas, R. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.* **5**, 505–516 (2000).
11. Sekimizu, T., Park, H.S. & Tsujii, J. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. in *Genome Informatics Workshop* 62–71 (Universal Academy Press, Tokyo, 1998).
12. Craven, M. & Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. in *Intelligent Systems for Molecular Biology* 77–86 (AAAI Press, Heidelberg, 1999).
13. Rindflesch, T.C., Rayan, J.V. & Hunter, L. Extracting molecular binding relationships from biomedical text. in *Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics* 188–195 (Association for Computational Linguistics, Seattle, 2000).
14. Rindflesch, T.C., Tanabe, L., Weinstein, J.N. & Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* **5**, 517–528 (2000).
15. Proux, D., Rechenmann, F., Julliard, L., Pillet, V. & Jacq, B. Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction. in *Genome Informatics Workshop* 72–80 (Universal Academy Press, Tokyo, 1998).
16. Fukuda, K., Tsunoda, T., Tamura, A. & Takagi, T. Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.* **3**, 705–716 (1998).
17. Andrade, M.A. & Valencia, A. Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. in *Intelligent Systems for Molecular Biology* 25–32 (AAAI Press, Halkidiki, 1997).
18. Jenssen, T.-K., Komorowski, J., Lægreid, A. & Hovig, E. Pubgen: Discovering and visualising gene-gene relations. in *Currents in Computational Molecular Biology* (eds. Miyano, S., Shamir, R. & Takagi, T.) 48–49 (Universal Academy Press, Tokyo, 2000).
19. Stapley, B.J. & Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.* **5**, 529–540 (2000).
20. Shatkay, H., Edwards, S., Wilbur, W. & Boguski, M. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. in *Intelligent Systems for Molecular Biology* 317–328 (AAAI Press, San Diego, 2000).
21. Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
22. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
23. Gu, Y., Shen, Y., Gibbs, R.A. & Nelson, D.L. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nature Genet.* **13**, 109–113 (1996).
24. Jager, U. *et al.* Follicular lymphomas' BCL-2/IgH junctions contain templated nucleotide insertions: novel insights into the mechanism of t(14;18) translocation. *Blood* **95**, 3520–3529 (2000).
25. Stamatopoulos, K. *et al.* Molecular insights into the immunopathogenesis of follicular lymphoma. *Immunol. Today* **21**, 298–305 (2000).
26. Lossos, I.S. *et al.* Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. *Proc. Natl. Acad. Sci. USA* **97**, 10209–10213 (2000).
27. Bravo, R. Growth factor-responsive genes in fibroblasts. *Cell Growth Differ.* **1**, 305–309 (1990).
28. Parsons-Wingerter, P., Elliott, K.E., Clark, J.I. & Farr, A.G. Fibroblast growth factor-2 selectively stimulates angiogenesis of small vessels in arterial tree. *Arterioscler. Thromb. Vasc. Biol.* **20**, 1250–1256 (2000).
29. Aronson, A.R., Rindflesch, T.C. & Browne, A.C. Exploiting a large thesaurus for information retrieval in *Proceedings of the 6th Applied Natural Language Processing Conference* (Rockefellar University Press, New York, 1994).
30. White, J.A. *et al.* Guidelines for human gene nomenclature. *Genomics* **45**, 468–471 (1997).
31. Cormen, T.H., Leiserson, C.E. & Rivest, R.L. *Introduction to Algorithms* 1028 (MIT Press, Cambridge, Massachusetts, 1990).