Nucleic Acids Research Advance Access published April 1, 2007

Nucleic Acids Research, 2007, ***1–11***
doi:10.1093/nar/gkm054

Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations

Steven Van Vooren[1,*], Bernard Thienpont[2], Björn Menten[3], Frank Speleman[3], Bart De Moor[1], Joris Vermeesch[2] and Yves Moreau[1]

[1]Department of Electrotechnical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium, [2]Center for Human Genetics, Leuven University Hospital, Herestraat 49, B-3000 Leuven, Belgium and [3]Center for Medical Genetics, Ghent University Hospital, MRB 2nd floor, De Pintelaan 185, B-9000 Ghent, Belgium

Received October 3, 2006; Revised October 27, 2006; Accepted January 18, 2007

ABSTRACT

Biomedical literature provides a rich but unstructured source of associations between chromosomal regions and biomedical concepts. By mining MEDLINE abstracts, we annotate the human genome at the level of cytogenetic bands. Our method creates a set of chromosomal aberration maps that associate cytogenetic bands to biomedical concepts from a variety of controlled vocabularies, including disease, dysmorphology, anatomy, development and Gene Ontology branches. The association between a band (e.g. 4p16.3) and a concept (e.g. microcephaly) is assessed by the statistical overrepresentation of this concept in the abstracts relating to this band. Our method is validated using existing genome annotation resources and known chromosomal aberration maps and is further illustrated through a case study on heart disease. Our chromosomal aberration maps provide diagnostics support to clinical geneticists, aid cytogeneticists to interpret and report cytogenetic findings and support researchers interested in human gene function. The method is available as a web application, aBandApart, at http://www.esat.kuleuven.be/abandapart/.

INTRODUCTION

Forward genetics, i.e. identification of gene mutations that underlie a phenotype of interest in a particular individual, is a key strategy to characterize gene function. In humans, where mutagenesis screens are impossible, genomic information from patients with developmental disorders can serve as the basis for disease gene discovery. Different positional cloning strategies, such as cytogenetic studies and linkage and association studies, can subsequently identify the chromosomal region where the disease gene is located.

To speed up the process of gene discovery, some attempts have been made to associate genomic rearrangements (such as subchromosomal deletions and duplications) to congenital malformations based on clinical and cytogenetic information from patients. Brewer *et al.* analyzed detailed clinical and cytogenetic information associated to a large number of autosomal deletions (1) and duplications (2) to construct a chromosome map showing associations of congenital malformations and chromosomal regions. Notably, these maps have not been updated since their publication.

Research groups with an interest in the etiology of, for example, congenital malformations often lack an extensive pool of patients to conduct large and informative association studies. Several public and private databases are being constructed to support such efforts by aggregating case reports and encouraging the exchange of patient information to complement private patient pools. Examples are the Catalogue of Unbalanced Chromosome Aberration in Man (3), the Human Cytogenetics Database and ECARUCA (www.ecaruca.net), DECIPHER (decipher.sanger.ac.uk), the Chromosome Anomaly Collection (www.som.soton.ac.uk/research/geneticsdiv/), the Mitelman Database of Chromosome Aberrations in Cancer (cgap.nci.nih.gov/Chromosomes/Mitelman), the Mendelian Cytogenetics Network DataBase (www.mcndb.org), Orphanet (www.orpha.net), etc. These efforts differ in setup but aim at aggregating chromosomal aberration information and charting phenotypes and case reports. Some catalogs are available only in print or at a licence fee, other databases require registration. Others are open and searchable by the public, but include no specific means for data mining.

*To whom correspondence should be addressed. Tel: +3216328654; Fax: +3216321970; Email: steven.vanvooren@esat.kuleuven.be

The information available in the public corpus of biomedical literature is a powerful alternative resource for patient reports and cytogenetic findings to conduct association studies. This corpus can be seen as a *de facto* genotype–phenotype association database. Moreover, it is not limited to case reports listing congenital malformations. Apart from disease related concepts, it is a rich source of information with regard to anatomy and development, systems and tissues and molecular functions and biological processes as well.

We have developed a method to automatically create chromosomal aberration maps from MEDLINE abstracts that mention (ranges of) cytogenetic bands. Through the use of multiple structured vocabularies, association with a band is not limited to a disease or syndrome, but also covers dysmorphology, human development and cell biology. The online application built on this method forms a bridge to the relevant and most current literature for further analysis by the researcher, rather than merely providing a catalog of genotype–phenotype associations. It thereby facilitates studies in the etiology of disease and the identification of disease genes. This resource is freely accessible and will stay up-to-date through regular automatic updates.

Related literature-mining methods

A number of tools and methods are currently available and offer capabilities for mining associations from literature between disease and genomic locations, although none have a scope identical to our method.

G2D (4) is a method for the prioritization of genes according to their relation to inherited disease. It allows a user to enter an OMIM disease identifier and a genomic region of interest. Through sequence and biomedical database analysis, G2D then identifies genes potentially associated with the disease.

HCAD (5) (Human Chromosome Aberration Database) is a web-based text-mining tool supporting analysis of human breakpoint data by mining the scientific literature to generate information on all human breakpoints.

Korbel *et al.* mine MEDLINE to identify clusters of gene-phenotype associations based on information on prokaryotic genomes (6). The results are not available through a web interface.

Tiffin *et al.* use an anatomical ontology to integrate text mining of biomedical literature and data mining of available human gene expression data (7). Their method prioritizes candidate genes according to their expression in disease-affected tissues.

iHOP (8) (information Hyperlinked Over Proteins) uses genes and proteins from multiple organisms as hyperlinks between sentences and abstracts to access and navigate PubMed.

MimMiner (9) is restricted to mining the OMIM database and ranks related phenotypes for a given phenotype or OMIM identifier. GeneSeeker (10) is a related tool that aims at the identification of genes underlying human genetic disorders by combining data on cytogenetic locus, phenotypes and expression patterns, to generate a list of candidate genes.

GFINDer (11) mines text data present in OMIM to annotate genes with gene ontology concepts and statistically selects relevant annotation categories. Phenotype descriptions are normalized to handle synonymy and are hierarchically structured.

Our method relates to these approaches but differs in several aspects. First, instead of extracting MEDLINE references linked to OMIM entries, or mining only text present in OMIM, MEDLINE abstracts are directly mined for cytogenetic bands and biomedical concepts. While curated databases offer high quality annotations and hence reduce noise, the use of abstracts allows mining to be more complete and up to date. Second, gene prioritization tools like G2D build an internal representation for the disease or phenotype under study through the intermediate association of MeSH and GO terms. This allows relating genes to phenotypes by means of chemicals, molecules, etc. In our method, this internal association process is rendered explicit through the choice of controlled vocabularies that allow the user to elucidate overrepresented associations between loci and concepts. Third, most of these tools offer only a disease-specific approach (in some cases using other annotations internally) while aBandApart explicitly allows for additional perspectives or user interests, such as dysmorphology, anatomy, development, molecular function, etc.

ABandApart is a novel analysis method based on abstracts present in MEDLINE for cataloguing biomedical concepts according to their association with chromosomal bands, which can be considered as a cytogenetic approach to genotype–phenotype correlation. Rather than prioritizing candidate genes, it focuses on cytogenetic bands and offers a portal into relevant literature. Through its different approach and goal, it can be considered complimentary to tools that already exist.

MATERIALS AND METHODS

Three elements are necessary to automatically build a chromosomal aberration map from MEDLINE abstracts: (1) identification of cytogenetic bands, (2) identification of concepts from multiple vocabularies and (3) assessment of the statistical overrepresentation of a concept among the abstracts relating to a band.

To discover overrepresented association between concepts and cytobands, we must first locate cytogenetic band identifiers and concepts from the vocabularies (and their synonyms) in the MEDLINE corpus. We have extended Lucene (12), a high-performance text-indexing engine written in Java, to parse all MEDLINE abstracts and extract cytogenetic bands, ranges of bands and biomedical concepts that are present in our different structured vocabularies.

Identification of cytogenetic bands

The International System for Human Cytogenetic Nomenclature (ISCN) gives a universal terminology of the description of chromosomal anomalies based on cytogenetic staining techniques (13). This nomenclature guarantees that all chromosomal anomalies are reported in a standardized way. Hence, reports in literature typically mention bands to delineate a genomic region at various levels of cytogenetic resolution. Because of this specific nomenclature, bands can be unambiguously extracted from text in the majority of cases. A similar approach is adopted in HCAD, where the nomenclature for translocations is used.

Although band patterns delineate chromosomal regions at a less detailed resolution than markers, base-pair positions, BAC clone identifiers, or genes, this approach is advantageous because of its effectiveness. Indeed, in most cases, chromosomal deletions and duplications have so far been resolved and reported only if their size was of the order of a cytogenetic band. Also, more accurate identifiers of genomic location are not used frequently or consistently enough in abstracts to construct a large and reliable mapping between genomic location and literature.

A range is a delineation of consecutive cytobands, possibly even spanning a centromere. Whenever such a range is encountered in an abstract, all the intermediate cytobands are associated to the abstract as well. A custom ontology resolves all bands in a range: a document mentioning 1p21.2-q23.1 will be annotated to all bands in between. In addition, an association to a certain abstract is transferred from a certain cytoband upwards through different levels of cytogenetic resolutions. This implies documents mentioning 3q26.32 will be annotated to 3q26 as well.

Based on this premise, we constructed a map that links MEDLINE abstracts to cytogenetic bands. This highly specific map was then used to characterize individual cytogenetic bands based on the content of the abstracts they are linked to. As the contents of the literature indices underlying aBandApart are updated regularly, the validation is based on a version of the tool that was frozen at the state of MEDLINE on 6 September 2005. Within that MEDLINE corpus, we identified 36 092 abstracts mentioning at least one cytogenetic band or range of bands. From this set, 293 808 associations between bands and concepts were extracted. Nearly 60 000 publications are added to the MEDLINE corpus every month. Hence, the number of abstracts and associations is expected to grow steadily as the system is continuously brought up to date.

A potential source of concern for the text-mining algorithm is that man is not the only organism for which banding patterns can be discerned through cytogenetic staining. Band nomenclatures also exist for other organisms. Genome architecture differs among species, which implies that assertions on human genotype–phenotype correlations are contaminated by literature dealing with nonhuman organisms for which a similar band pattern nomenclature is used. To assess the importance of this problem, we need to know the prevalence of documents dealing with nonhuman species in our corpus. We considered the complete set of documents that mention one or more cytogenetic bands and indexed this set using a vocabulary of both common and scientific organism names based on English animal-related lists (nouns and adjectives), as well as the NCBI taxonomy (www.ncbi.nlm.nih.gov/Taxonomy/). From this vocabulary, 489 distinct terms and phrases were detected at least once in the document set. The most frequently occurring species are shown in Table 1.

Table 1. The most frequently occurring species in a set of 36 082 cytogenetic MEDLINE abstracts mentioning cytogenetic bands

Rank	Phrase	Rank	Phrase
14 865	Human	126	Pig
3664	Mouse	107	Primates
1252	Rat	98	Papillomavirus
590	Rodent	70	Cat
474	Hamster	70	Bacteria
240	Bovine	68	Zebrafish
214	Melanogaster	67	Sheep
183	Chicken	63	Canine
178	Porcine	63	Troglodytes
135	Rabbit	61	Monkey

Note that the results from Table 1 do not imply that 14 865 documents discuss human cases and 3664 documents discuss mouse: on the one hand, the term *human* does not necessarily occur in all abstracts on human. On the other hand, the terms *human* and *mouse* can co-occur, since some abstracts discuss patients as well as model organisms. Although the mere occurrence of terms and phrases relating to organisms does not clearly elucidate the topic of a document, this brief analysis allows us to estimate how species are distributed as subjects of documents.

A clear majority of all references to organisms in our test corpus is human. The second most frequent organism is mouse and is referenced four times less often in the test documents. However, it does not add noise to the cytogenetic band detection because its band-staining patterns are indicated with capital letters followed by a number. The third most frequent organism is rat, as *rat* occurs in 3.47% of the test document set. As the rat chromosome nomenclature closely follows the human cytogenetic nomenclature (14), abstracts dealing with rat band patterns are a potential source of contamination—however, they represent only a small fraction of the abstracts.

The problem is further reduced because of at least two reasons. First, only a fraction of these rat-related documents actually contaminate the genome-to-literature map. We manually verified a random sample of 30 documents containing the term *rat*. Only a third contained cytogenetic bands that indeed referred to the rat genome, the other documents all contained bands that referred only to the human genome. This suggests that contamination of the genome-to-literature map by nonhuman band patterns is smaller still. Second, not all bands stand the risk of contamination. Human bands at high resolution (e.g. 4q15.32) do not occur in rat. In addition, for chromosome 1 (for example) and at the same cytogenetic resolution for rat and human, only 12 of 21 rat bands and only 12 of 24 human bands occur in both nomenclatures.

This brief analysis shows that the contamination effect must be kept in mind, but does not weigh significantly on the results of our method.

Vocabularies

Geneticists, pediatricians or physicians in general, dysmorphologists, molecular cell biologists and etiologists are all interested in making genotype–phenotype correlations. They have however each a different focus—for example, a different level of emphasis on clinical practice versus molecular biology research. To retrieve knowledge that is interesting to a specific researcher at a given time, we increase the specificity of the text-mining results by limiting its scope through controlled lists of concepts derived from biomedical vocabularies and ontologies.

These lists or sets of linked concepts confine the results of our information extraction method to the current interest of the researcher: different domain-specific vocabularies define from which perspective to annotate the genome. The available options include dysmorphology, anatomy-specific, gene- or protein-centered, gene ontology and disease-related perspectives on the literature. An overview is shown in Table 2.

Words as well as phrases are detected as concepts. In the case of ontologies, no relational information is kept, except from synonymy, which is taken into account when applicable (e.g. with LDDB as a vocabulary, the occurrence of the phrase *small head* will trigger an association to *microcephaly*).

The choice of controlled vocabularies is crucial to the scope and applicability of this method. The vocabulary sources were selected with both a research and diagnostics perspective in mind. For example, options range from a rather general heritable disease vocabulary (OMIM) to a specific Dysmorphology concept hierarchy (LDDB). Also, each vocabulary focuses on a different level of biological detail, from small (molecular, biochemical) over intermediate (cellular and tissue level) to large (organs and anatomy).

The sources for these vocabularies were chosen based on how authoritative they are in their respective field. Several of these vocabularies have already proven their value in previous work on gene profiling and prioritization. For example, the GO-derived vocabularies boost prioritization performance in Endeavour (15), our web-based method for candidate gene prioritization by genomic data fusion. Additionally, GO, MeSH and OMIM vocabularies have proven their merit in TXTGate, a web tool in support of previous work on text-based gene and gene group profiling (16). The dysmorphology vocabulary is also widely used in its field: first, the Oxford Medical Dictionary dysmorphology and neurology databases that build on the LDDB taxonomy are a widely used clinical reference. Second, LDDB is the elementary dysmorphology taxonomy within DECIPHER, the Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources, developed and hosted at the Wellcome Trust Sanger Institute (decipher.sanger.ac.uk).

Below, individual aBandApart vocabularies, their size, construction and origin are discussed in detail. The vocabularies themselves can be obtained from the authors upon request.

First, the OMIM- and LDDB-derived vocabularies provide hereditary disease and dysmorphology phenotype-specific views.

Second, a number of vocabularies have been constructed to mine literature at different levels of detail. From GO, vocabularies at the molecular and cellular level are constructed. TDMS offers tissues, organs and systems. Finally, two anatomical

Table 2. Different controlled vocabularies in aBandApart

Name	Function	Example	Size
MeSH	Medical subject headings	Chemicals, medical concepts	16.998
GO.B	Biological processes	'Cell growth', 'signal transduction'	1.120
GO.C	Cellular components	'Proteasome', 'nucleus'	402
GO.M	Molecular functions	'ATPase activity'	701
GO.E	Gene ontology	All of the above	2.170
LDDB	London dysmorphology database	'Microcephaly' or 'small head'	808
OMIM	Genetic disorders	'Attention deficit hyperactivity disorder'	1.716
CBIL	Human anatomy	'Heart muscle'	303
OHDA	Embryo development	'Early stage, fetus'	380
TDMS.s	Systems, tissues and sites	'Cardiovascular system'	392
TDMS.l	Microscopic lesions	'Disseminated intravascular coagulation'	204

A total of 11 vocabularies are present, shown above with an example concept and the number of concepts in each vocabulary.

vocabularies are provided, one of which is development specific.

Finally, medical and chemical terms and synonyms based on MeSH make up a general purpose vocabulary.

OMIM. This database is a catalog of human genes and genetic disorders that focuses primarily on heritable genetic diseases. Although OMIM does not contain direct information on chromosomal aberrations, it is relevant and useful as a resource of hereditary disease phenotypes. From its downloadable textual information, we have extracted a vocabulary of disease-related concepts out of which 1642 entries occur in our cytoband-related subset of the PUBMED corpus.

London Dysmorphology Database. Most clinical geneticists are familiar with the Oxford Medical Databases. LDDB contains information on 3428 dysmorphic syndromes and has a hierarchically structured feature vocabulary which we have manually annotated with synonymous phrases to increase recall in our method. In our band annotated corpus, 796 dysmorphologies are annotated through 1286 synonyms. This dictionary is an authoritative source (17) of information about dysmorphic and neurogenetic syndromes.

Gene Ontology. GO provides consistent descriptions of gene and gene-product attributes in the form of three structured controlled vocabularies that each provide a specific angle of view (biological processes, cellular components and molecular functions). The GO effort is deliberately term centered to allow for uniform queries across different databases. Our method does incorporate synonymy information from GO. GO is built and maintained with the explicit goal of applications in text mining and semantic matching in mind. Hence, the gene ontology is an ideal source for domain-specific views in our method and makes up four controlled vocabularies: (a) the whole set of GO concepts, for general associations to gene and gene-product attributes; (b) cellular components, which may include anatomical structures (e.g. *rough endoplasmic reticulum* or *nucleus*) or a gene-product group (e.g. *ribosome*, *proteasome* or a *protein dimer*) (18), (c) biological processes, defined as series of events accomplished by one or more ordered assemblies of molecular functions and (d) molecular functions, which describe activities at the molecular level.

TDMS tissue and lesions vocabularies. At another level up from the molecular and cellular scale, specific vocabularies are provided that are geared at organs, tissues and systems. Two vocabularies have been extracted from phrase lists used in a laboratory data acquisition system set up at the USA National Institutes of Health. The word lists of their toxicology data management system are subset in a vocabulary with microscopic lesions on the one hand and a vocabulary with microscopic sites, systems, tissues and organs on the other hand. This allowed us to complete the set of vocabularies ranging from the very small (molecular functions) over spatially larger concepts (cellular locations) to tissues and organs, which are part of the TDMS vocabularies. At the macroscopic end of this spectrum, CBIL offers anatomical structures.

CBIL anatomy. To focus on structures of larger scale than cellular and tissue levels, an anatomy-specific vocabulary was extracted from the hierarchical controlled vocabulary of anatomy terms from the computational biology and informatics laboratory at the University of Pennsylvania. The controlled vocabulary is based on anatomy terms taken from the mouse gene expression database at the Jackson Laboratory and was extended to incorporate human anatomy. It was then further revised in a number of areas, such as the haematolymphoid system and the brain.

Ontology of Human Developmental Anatomy. The Edinburgh Human Developmental Anatomy (19) lists the tissues present during the first 50 days after conception. This vocabulary is based on detailed anatomy information and standard named tissues for analysis of normal and abnormal human embryos. Space-associated data is included. Hunter *et al.* based this anatomical ontology on literature and on a detailed examination of histological material. It includes all the basic tissues recognizable to an experienced histologist and was designed for describing tissue at a fairly fine resolution (e.g. in gene expression experiments).

Medical Subject Headings. MeSH is the National Library of Medicine's controlled vocabulary thesaurus. From it, we constructed a vocabulary that takes into account all phrases up to six terms and maps all narrow and equal synonyms, leaving out broad synonyms. Apart from the 22 997 descriptors and their synonyms, over 150 000 entries and synonyms from the separate Supplementary Concept thesaurus are included as well, adding a general focus of chemical records to the vocabulary. Terms and phrases range from general to specific and constitute a general purpose vocabulary with broad coverage of the biomedical field. We recommend this vocabulary for first exploration of a genomic region and for when none of the specific vocabularies described above apply.

Statistical overrepresentation

Cytogenetic bands and concepts can occur together in a single document just by chance. First, consider an abstract where one band is mentioned together with one disease and that this disease is then compared to a second disease. Merely relying on co-citation within single documents would have such an abstract cause a spurious association between the band and the second disease. Second, a similar situation occurs when a document discusses several bands and contains multiple, loosely related case reports. This situation implies that we cannot accept a genotype–phenotype association based on the mere co-occurrence of the genomic location identifier (a cytogenetic band) and a concept from one of the vocabularies. Our method reports all co-occurrences together with a *P*-value indicating how much confidence an association deserves. To quantify this level of overrepresentation, we assume a hypergeometric distribution as a model.

The hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of C draws from a finite population without replacement. The population has labeled (success) and unlabeled items. The hypergeometric distribution describes $P(O)$, the probability that in a sample of C distinctive objects drawn from the global population, exactly O objects are labeled.

In the context of this work, the question is whether more papers link a cytoband b (e.g. '4p16.3') to a concept c (e.g. 'microcephaly') than one might expect by chance. If this is the case, the link between the concept and the cytoband can be thought of as being overrepresented in the text corpus.

Let A be the total number of abstracts annotated to all known cytobands and concepts. This is the size of the PUBMED sub-corpus in our text indices. Let C be the number of papers containing concept c or its synonyms and B the number of papers associated to b as described in the identification of cytogenetic bands section. We want to qualify the strength of the link between band b and concept c.

By only inspecting abstracts from the corpus that are linked to concept c, in fact, a draw is performed with C abstracts in it where some are and some are not linked to band b. Let O_{bc} be the observed number of papers that are associated to band b and mention concept c or one of its synonyms. To know whether the number of b-linked papers in that drawn sample is unusually large, we need to know the probability of drawing O_{bc} papers or more extreme outcomes. This corresponds to calculating the cumulative probability $P(X \geq 0)$ and can be calculated by the cumulative distribution function of a hypergeometric random variable X with parameters as described.

Since the hypergeometric distribution is a discrete probability distribution, the cumulative probability can be calculated easily by adding all corresponding single probability values. This probability constitutes a *P*-value since it is the probability of seeing something as extreme or more extreme than what was observed.

The *P*-value is then given by the hypergeometric cumulative distribution function

$$p_{bc} = 1 - H_{\mathrm{cdf}}(O_{bc} \mid A, B, C) \qquad \mathbf{1}$$

$$= 1 - \sum_{i=0}^{O_{bc}-1} \frac{\binom{B}{i}\binom{A-B}{C-i}}{\binom{A}{C}} \qquad \mathbf{2}$$

$$= \sum_{i=O_{bc}}^{\min(B,C)} \frac{\binom{B}{i}\binom{A-B}{C-i}}{\binom{A}{C}} \qquad \mathbf{3}$$

The *P*-value p_{bc} is the probability that we observe by chance O_{bc} documents or more that associate band b to concept c. It is the probability of observing O_{bc} or more documents linked to band b when drawing C concept-related documents without replacement from a corpus of A abstracts. Symmetrically, it equals the probability of observing O_{bc} or more documents linked to concept c when drawing B band-related documents.

It is important to note that for small numbers of concepts and documents, the *P*-values possibly provide a distorted view on the actual relevance of the band–concept association. Even though *P*-values for small counts still correctly represent the probability of observing this or a higher number of co-citations, it is clear that the *P*-value should be regarded with caution. The web application will show actual counts with each *P*-value, to allow the user to assert confidence in the association at hand. For a detailed discussion of this issue, see the Discussion section.

Relation to other distributions. When the population size is large compared to the sample size the hypergeometric distribution is approximated reasonably well by a binomial distribution. This approach is computationally less intensive. Both distributions were compared. Although a binomial approach proved justifiable, the hypergeometric distribution was chosen because it did not prove detrimental to performance. All *P*-values are precalculated at indexing time.

The hypergeometric test based on the hypergeometric distribution is identical to a one-tailed Fisher's exact test. This can be verified by writing down the 2 × 2 contingency table.

Statistically related tools. Our method uses different domain vocabularies as concept sources. Gene ontology as a whole, together with its three sub-branches, constitute four of our vocabularies. A range of existing tools operate on gene ontology alone to identify overrepresented concepts for (groups of) genes that result from expression array experiments. In these tools, the hypergeometric distribution and binomial approximation are prominent statistical methods. The hypergeometric approach and the equivalent Fisher's exact test constitute a standard approach in the majority of the tools, as discussed by Khatri *et al.* (20). In this review paper, they further state that although different distributions are used in different tools, it seems that in most cases the differences between the models are not dramatic.

Gentleman *et al.* describe the use of a hypergeometric distribution in GOStats and GOHyperG (21) to find concepts from gene ontology that are overrepresented for genes. Our work follows a similar philosophy, though applied to a series of unstructured vocabularies and literature co-citations of bands and concepts instead of genes and GO-terms.

FunSpec (funspec.med.utoronto.ca) inputs a list of yeast gene names and outputs a summary of overrepresented concepts. The tool calculates *P*-values using the hypergeometric distribution.

Other tools use the hypergeometrical distribution or equivalent Fisher's exact test for finding overrepresented concepts from gene ontology, including Fatigo (22), GObar (23), GoMiner (24), GOToolBox (25), GeneMerge (26), GOTree Machine (27), Ontology-Traverser (28), GOCluster (29) and GOHyperGAll in R BioConductor (30).

In BlastSets, Barriot *et al.* use the hypergeometric distribution to calculate the probability of having at least the observed number of elements in common between two sets of sequences for which biological relationships are inferred from different data sources.

Web application

We constructed a web application to illustrate and publicize our method and to make validation efforts reproducible. The tool functions in two directions.

On the one hand, users indicate a cytogenetic band on a genome view. These identifiers can also be entered manually. The tool will characterize this band with statistically overrepresented vocabulary concepts found in the literature. Users indicate which controlled vocabulary is to be used, according to their current research interest. For example, when aBandApart is queried with 4p16.3 and a disease vocabulary, the most significant concepts are *achondroplasia*, *Wolf-Hirschhorn syndrome*, *Huntington disease*, *multiple myeloma*, *cherubism*, *dwarfism* and *hypochondroplasia*, all of which are disorders confirmed to be associated to that region.

On the other hand, users can start from a concept and query the database for statistically overrepresented chromosomal regions. If the concept is not found, the application will suggest alternatives with similar spelling. Overrepresented bands are listed together with their *P*-values and the raw counts that were used to calculate each *P*-value. The highly overrepresented bands are highlighted in red on the same genome chart that is used for input of cytogenetic bands. Links to relevant literature are provided with the cytoband profile.

RESULTS

To illustrate our approach, we discuss results for searches related to heart disease. A detailed validation of our method follows as we discuss the performance on a set of 90 known gene–disease associations. We conclude by evaluating the correspondence of our results to chromosomal aberration maps composed by Brewer *et al.* (1,2).

Heart disease

We now illustrate the approach by querying the system for *heart* while selecting CBIL, the human anatomy vocabulary. The concept *heart* has a total of 1324 documents associated to it. The five most relevant hits are shown in Table 3.

Table 3. Five most relevant hits for query *heart* on vocabulary CBIL

Band name	BC	B	*P*-value
22q11	164	1092	0
22q11.2	83	755	1.28e−26
20p12	19	113	3.03e−10
21q22.2	16	171	5.88e−06
7q11.23	20	301	1.12e−04

The concept *heart* has a total of 1324 documents associated to it. The four columns show the hit, the number of documents that are linked to both band and concept, the number of documents linked to the band (hit) and the *P*-value.

Table 4. Highly significant hits (*P*-value <0.01) for query 7q11.23 on vocabulary CBIL

Concept	BC	B	*P*-value
Valve	5	51	8.23e−7
Connective tissue	6	96	2.64e−6
Aorta	5	70	5.43e−6
Metencephalon	1	2	3.92e−5
Heart	20	1324	1.12e−4
Hepatocyte	3	79	1.58e−3
Carotid artery	1	10	1.71e−3
Pons	1	13	2.92e−3
Tonsil	1	14	3.40e−3
Artery	3	120	7.06e−3
Penis	1	22	8.34e−3
Cardiovascular system	1	22	8.34e−3
Brain	23	2267	9.16e−3
Skeletal muscle	9	664	9.78e−3
Midbrain	1	24	9.88e−3

The band 7q11.23 has a total of 301 documents associated to it. The four columns show the hit, the number of documents that are linked to both band and concept, the number of documents linked to the concept (hit), and the *P*-value.

A very strong correlation is found for 22q11 and specifically, 22q11.2. Closer examination of these first two hits reveals that this association relates to the well-known DG/VCFS syndrome (DiGeorge/velocardiofacial syndrome). The zero *P*-value occurs because DG/VCFS, known as the 22q11.2 deletion syndrome, is the most common chromosomal deletion syndrome found in humans (32). Cardiac defects are strongly penetrant in those patients. The third best association, linking *heart* to 20p12, is corroborated by literature on the Alagille syndrome (33), a pleiotropic disorder with involvement of the liver, heart, skeleton, eyes and facial structures. The fourth, 21q22.2, is identified through literature analysis as a chromosomal region critical for heart defects related to Down syndrome (34). The fifth most relevant result is 7q11.23. When 7q11.23 is submitted as a query with the CBIL anatomy vocabulary, a link with the cardiovascular system is apparent. Results with highly significant *P*-values ($P < 0.01$) are shown in Table 4.

As an illustration of how working with different domain vocabularies can be beneficial, we characterized the same 7q11.23 band through different vocabularies. From the perspective of dysmorphology, through vocabulary *LDDB*, the highest ranking concept is *supravalvular aortic stenosis*. Other cardiovascular concepts occur, together with *anxiety* and *mental retardation*, suggesting involvement of the central nervous system. The latter is confirmed through use of the disease-related vocabulary, OMIM, linking the genomic location to the Williams–Beuren syndrome. To elucidate an underlying molecular function for this anomaly, the same query was submitted with the GO molecular function vocabulary. The highest ranking concept, *elastin*, is assigned a near zero *P*-value. Indeed, the majority of Williams–Beuren syndrome (WBS) patients have been shown to have a microdeletion within 7q11.2 including the elastin gene, leading to disorganized pre-elastic and mature elastic fibers (35). Through this brief discussion we have illustrated how different domain vocabularies each provide a specific view towards a genotype–phenotype association.

NIH data set—Genes and Disease

The online NIH book *Genes and Disease* (www.ncbi.nlm.nih.gov/books/), discusses a set of genes and the diseases that they are known to cause. With each genetic disorder, the underlying mutations are discussed, along with clinical features and links to key web sites. Over 80 genetic disorders have been summarized in this resource, which we use as positive controls in the validation of our method.

For chromosome 1, results are shown in Table 5. The first two columns show the gene name and disease as they occur in the NIH book. The disease name is the search term that was used to test our method. In some cases, spelling variants were used. Further columns indicate whether (**H**) the method assigned a highly significant *P*-value ($P < 0.01$) to the band to which the disease is actually associated, (**S**) whether it assigned a significant *P*-value ($P < 0.05$), (**P**) whether it delineated the band precisely; i.e. at the maximum level of karyotype resolution (4p16.1 is more precise than 4p16) and (**T**) whether it rated the band as the most significant candidate for this disease, ranking higher or as high as all other bands.

Table 5. NIH book validation for chromosome 1

Gene	Disease/concept	H	S	P	T	NIH	Top	*P*-value
UROD	Porphyria cutanea tarda	1	1	0	1	1p34.1	1p34	0.70E−4
GBA	Gaucher disease	1	1	1	1	1q21	1q21	2.41E−22
GLC1A	Glaucoma	1	1	1	0	1q24.3	1q24	2.21E−26
HPC1	Prostate cancer	1	1	1	0	1q25.3	8p22	0.00E−0
PS2	Alzheimer disease	0	1	1	0	1q42.13	1q42.1	0.24E−2

On this chromosome, five disease genes are annotated. Further columns indicate whether (**H**) the method assigned a highly significant *P*-value (<0.01) to the band to which the disease is actually associated, (**S**) whether it assigned a significant *P*-value (<0.05), (**P**) whether it delineated the band at the maximum level of karyotype resolution and (**T**) whether it rated the band as the most significant candidate for this disease, ranking higher or as high as all other bands.

A validation of our method with the disease-related genes on other chromosomes is provided as supplementary material.

Our method assigns a significant *P*-value ($P < 0.05$) to 84 out of 93 (over 90%) gene-linked diseases discussed in the NIH book data set. Of these, 80 (or 86%) are assigned a highly significant *P*-value ($P < 0.01$). For 57 (or 61%) of these genetic diseases, the cytogenetic band containing the causative gene was reported with the most significant *P*-value of all reported bands. These results can be verified through the supplementary material or reproduced through the aBandApart web application.

Eight diseases were not significantly linked to the band containing the causative gene. Most of these misses are explained by the fact that the concept is not in any of the domain vocabularies (6 of 9 misses). This occurs with complex or overly detailed concepts (e.g. *gyrate atrophy of the choroid and retina*) or chemical compounds (e.g. *steroid 5-alpha reductase*, *alpha-1-antitrypsin deficiency*). Although the concept *multiple endocrine neoplasia* does not occur in any of the vocabularies, the NIH band for this disease does show an relatively high number of cancer-related concepts.

Second, misses can also be explained by the fact that there exists no literature in the MEDLINE corpus associating a concept or any of its synonyms to the band in question. This is the case for the CKN1 gene, where no abstracts link the Cockayne syndrome to 5q12 and for the Zellweger syndrome, where no literature links it to 12p13.3.

Finally, although a band is found, it is sometimes not assigned a significant *P*-value. This is the case for *diabetes*, which our method only weakly links to 7p13. Diabetes has putative causative links to many genomic regions.

Congenital malformations

To further validate our methodology, we evaluate its agreement with chromosome maps of autosomal deletions and duplications composed by Brewer *et al.* (1,2). In this work, clinical and cytogenetic information from the human cytogenetics database was used to associate different congenital malformations to nonmosaic single contiguous autosomal deletions and duplications. We have assembled a list of 63 malformation-to-band associations that the authors deemed statistically highly significant. Brewer *et al.* classified malformations in seven categories: cardiac, central nervous system, craniofacial, gastrointestinal, genitourinary, ocular and skeletal and limb malformations.

Out of 63 malformation-associated bands deemed significant by Brewer *et al.*, 44 were assigned a significant *P*-value by our method (70%), 35 were given a highly significant *P*-value (56%). Five associations were detected but not given a significant *P*-value. Of the 14 associations made by Brewer *et al.* that were not detected by our method, one was missed because of different phrasing of *agenesis of corpus callosum* in literature and 13 were missed because no abstracts were found linking band and malformation. Detailed results for all 13 cardiac anomalies discussed by Brewer et al. are shown in table 6. The full validation is provided as supplementary material.

DISCUSSION

aBandApart links phenotype information to genomic aberrations at the level of cytogenetic bands. We assessed that significant *P*-values yielded by the method are supported by known cytogenetic aberrations and by published malformations and diseases.

With regard to our text-mining methodology, one point worth noting is that MEDLINE abstracts are used instead of the full text of the corresponding article. Although full text articles are increasingly made available through centralized repositories and open access initiatives, harvesting full text is not possible for all publications because of technical and legal restrictions. Although the potential difference in information present in full text must be kept in mind (for example, the surplus of sequence-related data reported in full text versus abstract was proved to be significant in an earlier study (36)), the use of abstracts alone is justified because they summarize the key information from a paper (for example, as keywords (37)).

Regarding the statistical methodology, it is again worth stressing that the hypergeometric approach can yield small *P*-values for associations that not necessarily deserve to be marked as meaningful. This is the case for very small numbers of concepts and documents. For example, associations of *4p16.3* to both *broad nasal tip* and *microcephaly* are flagged as significant by this method; the first based on one co-citation in 2 documents, the latter on 11 co-citations in 322 documents. Even though both resulting *P*-values correctly represent the probability of observing this or a higher number of co-citations in a statistical sense, it is clear that the *P*-value in the first situation should be regarded with more caution. One option could be to use a regularized estimator that penalizes more strongly associations involving few documents. We decided against this choice because such associations can be meaningful. In the case of association through few documents, individual abstracts must be reviewed to confirm the potential associations and avoid overreliance on the *P*-value. To allow an informed decision on the actual significance of an association between a band and a concept, the web application also indicates the actual counts that were used to calculate the *P*-value. This raw count information is crucial to the interpretation of results from the web tool: *P*-values must always be evaluated in the light of the counts mentioned in the 'Links' column directly to the left. The caption of the result table explicitly mentions the meaning of each field.

Table 6. Congenital malformation validation

Malformation	Band	Type	*P*-value <0.01	*P*-value <0.05
Aortic stenosis	11q23-24	del		
Hypoplastic left heart	11q23-25	del	✓	✓
Hypoplastic left heart	16q11-12	dup		
Patent ductus arteriosus	16q22	dup	✓	✓
Pulmonary stenosis	20p13-11	del	✓	✓
Pulmonary stenosis	22q11	del	✓	✓
Pulmonary stenosis	8q22-24	dup		
Tetralogy of fallot	8q22-24	dup	✓	✓
Truncus arteriosus	22q11	del	✓	✓
Truncus arteriosus	2q22	del		
Ventricular septal defect	22q11	del	✓	✓
Ventricular septal defect	4q31	del		✓
Ventricular septal defect	8q24	dup		✓

All 13 cardiac anomalies discussed by Brewer *et al.* are shown. Check marks indicate the significance with which our method associated band and concept.

Our method for associating biomedical concepts to cytogenetic bands provides diagnostics support to clinicians looking to identify chromosomal regions containing genes involved in disease processes, and to determine clinical entities linked to genomic aberrations in patients. It supports genetic counselling and an educated followup of clinical cases. It also aids cytogeneticists to generate refined accounts on cytogenetical findings they interpret and report to medical professionals (such as gynecologists, pediatricians, psychiatrists or genetic counselors) and to the patient's family.

For researchers, the generation of a phenotypic genome map based on text mining will ease the identification of genes involved in disease processes and could delineate novel clinically recognizable entities. Through our controlled vocabularies, their research can be focused on specific knowledge domains. Additionally, the tool provides non-cytogeneticists an accessible bridge to the cytogenetic literature.

The databases can support curation of chromosomal aberration catalogs. They do not render case report catalogs obsolete, rather, they aim at complementing these resources by offering a publicly available, free, online and searchable resource that is kept up to date through regular automated updates.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

Our research is supported by the following grants. Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants.

Flemish Government: FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

Conflict of interest statement. None declared.

REFERENCES

1. Brewer,C., Holloway,S., Zawalnyski,P., Schinzel,A. and FitzPatrick,D. (1998) A chromosomal deletion map of human malformations. *Am. J. Hum. Genet*., **63**, 1153–1159.
2. Brewer,C., Holloway,C., Zawalnyski,P., Schinzel,A. and FitzPatrick,D. (1999) A chromosomal duplication map of malformations: regions of suspected haplo- and triplolethality—and tolerance of segmental aneuploidy—in humans. *Am. J. Hum. Genet*., **64**, 1702–1708.
3. Schinzel,A. (2001) *Catalogue of Unbalanced Chromosome Aberration in Man*. de Gruyter. Berlin.
4. Perez-Iratxeta,C., Wjst,M., Bork,P. and Andrade,M.A. (2005) G2D: a tool for mining genes associated with disease. *BMC Genet*., **6**, 45.
5. Hoffmann,R., Dopazo,J., Cigudosa,J.C. and Valencia,A. (2005) HCAD, closing the gap between breakpoints and genes. *Nucleic Acids Res*., **33**, 511–513.
6. Korbel,J.O., Doerks,T., Jensen,L.J., Perez-Iratxeta,C., Kaczanowski,S., Hooper,S.D., Andrade,M.A. and Bork,P. (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*., **3**, e134.
7. Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Evaluation Studies. *Nucleic Acids Res*., **33**, 1544–1552.
8. Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**(Suppl 2), ii252–ii258.
9. van Driel,M.A., Bruggeman,J., Vriend,G., Brunner,H.G. and Leunissen,J.A.M. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*., **14**, 535–542.
10. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P.C.W., Leunissen,J.A.M., Brunner,H.G. and Vriend,G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res*., **33**(Web Server issue): 758–761.
11. Masseroli,M., Galati,O. and Pinciroli,F. (2005) GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res*., **33**(Web Server issue), 717–723.
12. Hatcher,E. and Gospodnetić,O. (2004) *Lucene in Action*. Manning Publications Co. Greenwich, Connecticut, USA.
13. Shaffer,L.G. and Tommerup,N. (2005) *ISCN 2005*. Karger Basel.
14. Levan,G. (1974) Nomenclature on G-bands in rat chromosomes. *Hereditas*, **77**, 37–52.
15. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.-C., De Moor,B., Marynen,P. *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol*., **24**, 537–544.
16. Glenisson,P., Coessens,B., Van Vooren,S., Mathys,J., Moreau,Y. and De Moor,B. (2004) TXTGate: profiling gene groups with text-based information. *Genome Biol*., **5**, R43.
17. Mohnish,S. (2002) Oxford Medical Databases: London Dysmorphology Database Version 3.0. *J. Med. Genet*., **39**, 782–783.
18. The Gene Ontology Consortium (2000) Gene Ontology; tool for the unification of biology. *Nat. Genet*., **25**, 25–29.
19. Hunter,A., Kaufman,M.H., McKay,A., Baldock,R., Simmen,M.W. and Bard,J.B.L. (2003) An ontology of human developmental anatomy. *J. Anatomy*, **203**, 347–355.
20. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
21. Falcon,S. and Gentleman,R. (2006) Using Gostats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
22. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
23. Lee,J.S.M., Katari,G. and Sachidanandam,R. (2005) GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, **6**, 189.
24. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane, D.W., Reinhold,W.C. *et al.* (2003) Gominer: a resource for biological interpretation of genomic and proteomic data. *Geome. Biol*., **4**, R28.
25. Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based Gene Ontology. *Genome Biol*., **5**, R101.
26. Castillo-Davis,C.I. and Hartl,D.L. (2003) GeneMerge-post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.
27. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
28. Young,A., Whitehouse,N., Cho,J. and Shaw,C. (2005) Ontology-traverser: an R package for GO analysis. *Bioinformatics*, **21**, 275–276.
29. Wrobel,G., Chalmel,F. and Primig,M. (2005) GoCluster integrates statistical analysis and functional interpretion of microarrary expression data. Evaluation Studies. *Bioinformatics*, **21**, 3575–3577.
30. Doerge,R.W. (2006) Bioinformatics and computational biology solutions using R and bioconductor edited by Gentleman,R., Carey,V., Huber,W., Irizarry,R. and Dudoit,S. *Biometrics*, **62**, 1270–1271.
31. Barriot,R., Poix, J., Groppi,A., Goffard,N., Sherman,D., Dutour,I., de Daruvar,A. (2004) New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res*., **32**, 3581–3589.
32. Yakut,T., Kilic,S.S., Cil,E., Yapici,E. and Egeli,U. (2006) FISH investigation of 22q11.2 deletion in patients with immunodeficiency and/or cardiac abnormalities. *Pediatr Surg. Int*., **22**, 1–4.
33. Krantz,I.D., Smith,R., Colliton,R.P., Tinkel,H., Zackai,E.H., Piccoli,D.A., Goldmuntz,E. and Spinner,N.B. (1999) Jagged1 mutations in patients ascertained with isolated congenital heart defects. *Am. J. Med. Genet*., **84**, 56–60.
34. Kosaki,R., Kosaki,K., Matsushima,K., Mitsui,N., Matsumoto,N. and Ohashi,H. (2005) Refining chromosomal region critical for Down syndrome-related heart defects with

a case of cryptic 21q22.2 duplication. *Congenit. Anom. (Kyoto)*, **45**, 62–64.

35. Robinson,W.P., Waslynka,J., Bernasconi,F., Wang,M., Clark,S., Kotzot,D. and Schinzel,A. (1996) Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. *Genomics*, **34**, 17–23.
36. Wren,J.D., Hildebrand,W.H., Chandrasekaran,S. and Ulrich Melcher,U. (2005). Markov model recognition and classification of DNA/protein sequences within large text databases. *Bioinformatics*, **21**, 4046–4053.
37. Shah,P.K., Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2003) Information extraction from full text scientific articles: where are the keywords? Evaluation Studies. *BMC Bioinformatics*, **4**, 20.