Published online January 2, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 1 ***135–142***
DOI: 10.1093/nar/gkh162

Automatic extraction of mutations from Medline and cross-validation with OMIM

Dietrich Rebholz-Schuhmann*, Stephane Marcel[1], Sylvie Albert[1], Ralf Tolle[2], Georg Casari[3] and Harald Kirsch

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK, [1]LION Bioscience AG, Waldhoferstrasse 98, D-69123 Heidelberg, Germany, [2]PheneX Pharmaceuticals AG, Im Neuenheimer Feld 515, D-69120 Heidelberg, Germany and [3]Cellzome AG, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Received August 13, 2003 ; Revised September 17, 2003; Accepted November 12, 2003

ABSTRACT

Mutations help us to understand the molecular origins of diseases. Researchers, therefore, both publish and seek disease-relevant mutations in public databases and in scientific literature, e.g. Medline. The retrieval tends to be time-consuming and incomplete. Automated screening of the literature is more efficient. We developed extraction methods (called MEMA) that scan Medline abstracts for mutations. MEMA identified 24 351 singleton mutations in conjunction with a HUGO gene name out of 16 728 abstracts. From a sample of 100 abstracts we estimated the recall for the identification of mutation–gene pairs to 35% at a precision of 93%. Recall for the mutation detection alone was >67% with a precision rate of >96%. This shows that our system produces reliable data. The subset consisting of protein sequence mutations (PSMs) from MEMA was compared to the entries in OMIM (20 503 entries versus 6699, respectively). We found 1826 PSM–gene pairs to be in common to both datasets (cross-validated). This is 27% of all PSM–gene pairs in OMIM and 91% of those pairs from OMIM which co-occur in at least one Medline abstract. We conclude that Medline covers a large portion of the mutations known to OMIM. Another large portion could be artificially produced mutations from mutagenesis experiments. Access to the database of extracted mutation–gene pairs is available through the web pages of the EBI (refer to http://www.ebi.ac.uk/rebholz/index.html).

INTRODUCTION

Importance of point mutations in medicine

For more than two decades molecular biologists, together with genetic epidemiologists and medical doctors, have searched for the genetic predisposition of diseases. A major contribution to this predisposition stems from the single base variability of gene sequences, e.g. base deletions, insertions and substitutions, which are called single nucleotide polymorphisms (SNPs) or mutations (for terminology use refer to Materials and Methods). They are the cause of altered gene regulation or changes in amino acid sequence. This is the case in, e.g., sickle cell anaemia (1), where a simple amino acid exchange in the haemoglobin leads to altered crystallization properties, and to a reduced uptake of oxygen due to the resulting deformation of the erythrocytes.

Researchers investigating such disease-relevant mutations have to check public data sources to ascertain the novelty, importance and usefulness of their findings (2). Two important data sources are (i) Medline, a database of indexed abstracts from scientific biomedical literature (3; http://www.ncbi.nlm.nih.gov/PubMed/) and (ii) OMIM (Online Mendelian Inheritance in Man), a database which provides public access to curated data gathered from public scientific literature as well as other sources (4,5; http://www.ncbi.nlm.nih.gov/Omim/). Apart from these two sources dbSNP has to be considered as well (6,7; http://www.ncbi.nlm.nih.gov/SNP/).

Although both OMIM and Medline provide information on mutation–gene pairs, neither offer complete information on mutations (8). OMIM does not contain mutations produced in mutagenesis experiments, while Medline does not include information from the body of the complete publication.

Automatic extraction of mutation–gene pairs from Medline

Since the content of Medline consists of abstracts provided as natural language text, it is not at all easy to access the information via Boolean queries, which are known from, for example, relational databases. Word search is therefore the most important search technique.

Word search tends to be time-consuming and difficult, if not impossible, if all mutations known for one gene from the complete set of abstracts have to be extracted. The retrieval of all abstracts relating to one gene is already a non-trivial task. This is due to the fact that terms often do not refer to exactly one concept, e.g. a gene or a protein. One reason is that terms tend to be ambiguous and then refer to different concepts at the same time, and another reason is that different terms (variants) refer to the same concept. This is the case for typographic variants, acronyms (abbreviations) and synonyms of a term

*To whom correspondence should be addressed. Tel: +44 1223 492594; Fax: +44 1223 444468; Email: rebholz@ebi.ac.uk

(9,10). In the case of ambiguity the retrieval is too large (low precision), and in the case of term variants the retrieval is too small (low recall), if not all variants are considered.

The retrieval of abstracts, which inform about mutations, again generates difficulties. The reason is that authors use different types of descriptions for the representation of a mutation. Although there is a nomenclature, e.g. Trp64Arg, other phrases, e.g. 64 Trp→Arg, are frequently used. Furthermore, mutations are as well encoded in natural language text, e.g. 'tryptophan to arginine substitution at residue 64'. As a result quite a few patterns have to be considered to find all abstracts referring to a specific mutation.

In addition to the different variants of mutations, ambiguities also have to be resolved in the identification of mutations. For example C13T can refer either to a nucleotide sequence mutation (NSM) in position 13 where cytosine is replaced by a thymine or to a protein sequence mutation (PSM) where a cysteine is replaced by a threonine. In addition, C13T denotes a neuroblastoma cell line. Such terms have to be disambiguated in a post-processing step (11). Altogether, a correct retrieval of abstracts reporting one or several mutations to a gene has to consider quite a few keywords for the gene and a number of patterns for the detection of the mutation.

Even a successful query leads to the retrieval of a set of documents from which the researcher has to extract the mutation–gene pairs by reading. The association between the mutation and the gene or protein is easy if only one gene is mentioned throughout the abstract. Where several genes are mentioned, the correct association between mutation and gene has to be identified.

Different solutions have been proposed to automatically extract information from scientific literature. Initially they were applied to annotate biological sequences (12), to extract protein–protein interactions (13) and to identify drugs and genes from the scientific literature (14). Up to now no system has been designed to identify mutations in conjunction with a gene (2).

In the next section we describe methods to automate the detection of mutations and the extraction of mutation–gene pairs. The result is a database of such pairs. In the Results section, we assess recall and precision of the extraction methods, e.g. for the mutation extraction and the mutation–gene pair extraction. In addition, we have extracted from OMIM a set of PSM–gene pairs which we compare to our findings from Medline. The presented data lead to assumptions on how the two data sources differ. Furthermore, the links to Medline abstracts allow us to examine the related bibliography. In the last section we discuss limits to our methods, e.g. low recall on mutation–gene pairs, and the question of how far it can be expected that Medline and OMIM cover the same set of mutation–gene pairs.

MATERIALS AND METHODS

Terminology

The methods described in this publication are designed to detect base substitutions as well as short nucleotide insertions and deletions. Within this publication we refer to these events jointly as 'mutations'. Our methodology currently does not co-extract allele frequency data, which means that the numerous SNPs, which can be extracted from the literature, will also be referred to as 'mutations'. The term 'protein sequence mutation' (PSM) identifies those mutations that are reported in amino acid nomenclature and thus clearly lead to altered amino acid sequences. The term nucleotide sequence mutation (NSM) refers to those mutations that are described in DNA nomenclature. As the current DNA nomenclature does not allow a description of the NSM consequences, some of the detected NSMs could at the same time represent PSMs.

Matching technology

The abstracts analysed were downloaded from the public server of Medline according to the rules indicated there. Only those abstracts provided through the public server before September 10th, 2001 were considered. For the complete analysis all abstracts containing a HUGO gene name were scanned.

Our automatic analysis method consists of different components (Fig. 1): (i) an identification module for the gene names, (ii) an identification module for the patterns describing polymorphisms and (iii) a disambiguation module. Additional modules transfer the data into the database and generate Web pages to a query to the database.

The identification modules for the gene names and the polymorphism patterns are based on regular expressions (RegExp). Regular expressions can be implemented in Python, Perl, with the help of Flex in C and C++ and in other techniques (15). The technology to compile RegExp and any instance of compiled regular expressions is also referred to as Finite State Automatons (FSAs). We applied our own implementation of FSAs, which has been developed for linguistics-based sentence analysis using FSAs (16). The automata were optimized to run in 1 GB of main memory. A complete run takes ~3 days on a Linux system (one CPU, 2 GHz, 1 GB main memory).

Any gene name is matched in its uppercase and lowercase variant, e.g. COL1A1 versus col1a1, and in the lowercase variant with a leading uppercase letter. This leads to the regular expression (COL1A1|[cC]ol1a1). All gene names are encoded the same way and all regular expressions are compiled to a single module to perform complete gene name identification in one run. For the identification of gene concepts we extracted 16 142 names from HUGO, which includes the synonyms mentioned there. For the comparison of our data to OMIM we counterchecked that every OMIM entry refers to a named entity represented in HUGO.

The identification of the polymorphisms is again based on regular expressions. The example 'C282Y' (Table 1) is identified from the regular expression [AC-IK-NP-TVWYZ][0-9]+[AC-IK-NP-TVWYZ]. The examples can be recognized with similar expressions, with some minor modifications. The differentiation between a PSM and a NSM in the case of [ACGT] is done on a later stage. Again, all patterns are compiled to a single module to analyse the text in one run. A set of 30 patterns was identified in Medline to extract the mutations. A subset is shown in Table 1. Our patterns consider the one-letter codes for amino acids and nucleotides as well as the three-letter codes and complete names for amino acids only.

In order to find the gene name mentioned in conjunction with a mutation, the patterns have been applied to the

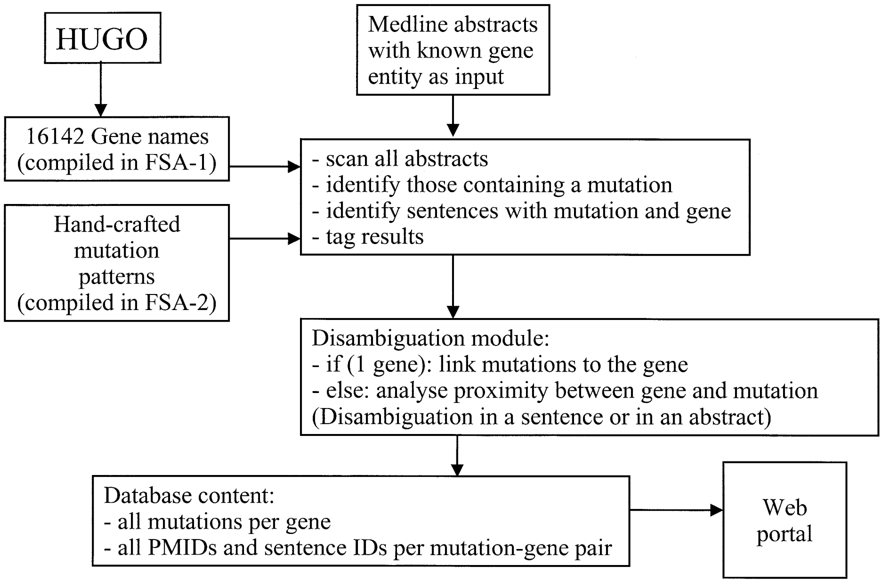
HUGO
16142 Gene names
(compiled in FSA-1)
Hand-crafted
mutation
patterns
(compiled in FSA-2)
Medline abstracts
with known gene
entity as input
- scan all abstracts
- identify those containing a mutation
- identify sentences with mutation and gene
- tag results
Disambiguation module:
- if (1 gene): link mutations to the gene
- else: analyse proximity between gene and mutation
(Disambiguation in a sentence or in an abstract)
Database content:
- all mutations per gene
- all PMIDs and sentence IDs per mutation-gene pair
Web
portal

Figure 1. Workflow overview. 16 142 HUGO gene names were integrated as patterns into a finite state automaton. This is also true for mutation patterns, which encoded a mutation as regular expression. All Medline abstracts were scanned and the different FSAs extracted the phrases and tagged the result.

Table 1. Several examples of phrases describing mutations found in Medline

Arg506 to Gln
valine 804→leucine
Ile15 to Thr15
Pro12Ala
arginine(3500)–glutamine
C282Y
A1166→C
677C→T
1166A/C,
359 (Ile/Leu)
Nucleotide 383T→C
codon 113 and His→Arg
Cys/Val343
Val→Ala at codon 113
IVS1–2A→G
codon 241 and codon 247, where the single base changes from C to T
Methionine to threonine substitution at residue 235
Methionine for valine at position 30
Ser→Leu change at amino acid 217
Heterozygosity for the IVS-I-5 (G→C) mutation
A fourth mutation, 433–2(A→G) transition, was identified at the splice-acceptor site in intron 2

sentences. If an abstract contains only one gene name (or synonym), the detected mutation phrase was associated to this gene, independently of the localization of the mutation in the abstract. If the abstract contains several gene names, the phrase is kept if at least one gene name appears in the same sentence. If several gene names have been encountered in one sentence, then syntactical rules and proximity parameters were used as decision criteria.

1117 (3.85%) of our findings have to be classified as ambiguous, because they use a one-letter code (A, T, C, G) to describe the substitution. Automatic disambiguation with the help of contextual information was not able to tell whether the letter referred to a nucleotide or an amino acid, and curation of the items from a sample led to the result that a portion of the sample cannot be disambiguated at all (results not shown).

After the automatic extraction the outcome is evaluated regarding recall (relevant facts found/relevant facts available) and precision (relevant facts found/facts found).

To estimate recall and precision we retrieved those abstracts from PubMed that contained the keyword either ‘mutation’ or ‘polymorphism’. Each of these keywords is frequently found in conjunction with mutations. From this set of abstracts a random sample of 100 abstracts was selected, each of which mentions at least one mutation. This subset of abstracts was analysed manually and with the help of our automatic methods. The results were compared to each other.

Finally, we downloaded a version of OMIM dating to December 2001, which was publicly available, and extracted all mutation–gene pairs as they were explicitly provided through this database (4). We selected the set of PSMs and NSMs (substitution type) to be able to compare our extracted information to OMIM (cross-validation), and determined the intersection between both data sets.

RESULTS

Evaluation of the extraction method

We propose a method which allows automatic extraction of mutation–gene pairs from Medline. It is primarily focused on nucleotide and protein sequence mutations of the substitution type (Table 1). We take the gene names from the HUGO nomenclature (17,18; http://www.gene.ucl.ac.uk/nomenclature/). Protein names are also extracted, if they are synonymous to a gene name. Our method identified 24 351 unique pairs from 16 728 Medline abstracts.

For the evaluation of our methods we distinguished between the overall conclusion that an abstract reports a mutation–gene pair (contained mutation–gene pair) and the more specific set of individual facts listed in the abstract (cited mutation–gene pair). The same distinction is used for contained named entities versus cited named entities. Precision and recall referring to cited mutation–gene pairs consider several findings in an abstract as individual results. For a contained mutation–gene pair it is sufficient if it is correctly identified throughout the abstract at least once. For the identification of genes we use the HUGO nomenclature and therefore only consider cited HUGO genes as contained genes.

We determined recall and precision from a random sample of 100 abstracts containing mutations (refer to Materials and Methods), which were retrieved by keyword search and analysed by hand. These abstracts had been automatically processed with our methods.

In addition to the mutations, the associated gene was extracted from the sentence containing the mutation citation or from the remaining parts of the abstract (contained mutation–gene pairs). In only 91 abstracts, referring to 233 cited mutations (162 contained mutations), could a contained HUGO gene be verified by hand. Thirty-five abstracts contained several gene entities, which led to ambiguities in the identification of the correct mutation–gene pair. The precision for the association of the contained gene to the contained mutation is 93.4%. Recall was estimated to 35.2% (57 out of 162).

For the cited mutations our extraction methods have 99% precision at a recall of 74% (Table 2). This proves that our method is precise in the detection of mutations. The recall was higher where PSMs were specified with the one-letter amino acid code (79.1%), e.g. R for arginine, in comparison to the other types (66.7%), e.g. Arg for arginine, which is due to the fact that language patterns for the one-letter code have lower variability. The recall in either case can be further improved with additional patterns describing as yet missed representations of mutations.

Cross-validation of PSM–gene pairs with OMIM

The manual validation of the complete set of extracted mutation–gene pairs would be extremely time consuming and requires background knowledge in the respective disease area. An alternative is the comparison of the extracted data to the OMIM database. It has been generated from public literature and other sources with the help of curators and provides information on the genetic cause of diseases in humans as well as references to known mutation–gene pairs. For our comparison the public version from December 2001 was used.

First we identified the largest sub-selection of mutation–gene pairs from OMIM that could be compared to our extracted data (Mutation Extraction from Medline Abstracts, MEMA). The sub-selection consists of 6699 PSMs (substitution type). This group is the largest fraction in OMIM. NSMs were not considered, since their one-letter code always has to be disambiguated. The remaining polymorphism entries in OMIM (3384) refer to other types of mutations, e.g. deletions and insertions. Table 3 gives an overview of the content in OMIM and in MEMA. Out of 6699 PSM–gene pairs from OMIM, 1826 were identified automatically (27%).

For all PSM–gene pairs in OMIM we selected the abstracts which contained the gene as well as the PSM (PSM and gene co-occurrence). We could identify such an abstract for 2002 PSM–gene pairs. We analysed more closely why this number is low in comparison to the complete set of mutation–gene pairs in OMIM (6699 pairs for 1041 genes). 242 gene names do not appear in co-occurrence with any kind of mutation in Medline, and another 92 genes occur together with a PSM, but not with any of the PSMs mentioned in OMIM. In the first case a larger gene name set will improve the recall, and in the second case a citation might be found in a different source, e.g. in the complete publication or in a journal that is not listed in Medline.

Next we counterchecked how many of these PSM–gene pairs were extracted by our methods. We identified 1826 pairs representing 91%. This is the recall of contained PSM–gene pairs in the complete set of pre-selected documents. The PSM–gene pairs contained in both data sets are called cross-validated pairs or BOTH pairs.

Distribution of PSM–gene pairs in OMIM and Medline

In the last step we compared OMIM to data extracted by MEMA for those genes where cross-validated PSM–gene pairs have been found. The expectation is that a gene with a large number of PSMs kept in OMIM will have also a large number of PSMs available from Medline. But we found that the correlation coefficient of the number of PSMs per gene in OMIM in comparison to MEMA is only 0.53. Figure 2 lists the genes with the highest number of PSM–gene pair entries in OMIM.

We grouped the genes into three different categories: (i) OMIM-owned genes, (ii) MEMA-owned genes and (iii) BOTH-owned genes. In the case of OMIM ownership the majority (>50%) of PSMs for a given gene is unique to OMIM and, in the case of MEMA-ownership, unique to Medline. In other words, in either case the majority of PSMs for a given gene have not been cross-validated. If a gene is attributed 'BOTH-owned', then the majority of its PSMs have been cross-validated. In general this classification attribute

Table 2. Recall and precision estimates for different parameters

	Recall Total	(%)	Precision Total	(%)
Cited mutation in one-letter code	151/191	(79.1)	151/152	(99.3)
Cited mutation in three-letter code or in complete name	52/78	(66.7)	52/54	(96.3)
Cited mutation	204/273	(74.7)	204/207	(98.6)
Contained mutation	143/190	(75.3)	143/146	(97.9)
Contained mutation–gene pairs	57/162	(35.2)	57/61	(93.4)

The numbers were estimated from a sample of 100 abstracts, which led to the identification of 273 citations: 191 for the one-letter code and 78 for the three-letter code and complete name. Precision is high for the identification of the mutation. On the other hand, the recall for the association between the gene mentioned in the abstract and the mutation needs to be improved. Mutation–gene pairs are mainly missed due to ambiguities from additional genes in the context of the mutation and due to named entities not compliant with the HUGO nomenclature.

Table 3. Number of mutation–gene pairs in OMIM and in MEMA

	Genes	Mutations	PSMs	NSMs	Ambiguous mutations	Total
Extracted from OMIM	1215	10 083	6699	207	0	6906
Extracted by MEMA	2115	24 351	20 503	2376	1117	23 996
Common to OMIM and MEMA	782	1887	1826	38	0	1864
Unique to OMIM	433	8196	4873	169	0	5042
Unique to MEMA	1333	22 464	18 677	2338	1117	22 132

The OMIM version of December 2001 provided 10083 entries, of which 6699 refer to PSMs and 207 refer to NSMs. The remaining entries refer to deletions, insertions and other types of polymorphisms. MEMA provides access to 24 351 mutations, of which 20 503 refer to PSMs, 2376 are NSMs and 1117 could not be classified automatically. 1826 PSM–gene pairs were cross-validated through OMIM and MEMA.

Table 4. Total number of PSM–gene pairs for different gene categories

Gene categories	No. of genes	PSMs in OMIM	(Average)	PSMs in OMIM + MEMA	(Average)	PSMs in MEMA	(Average)
OMIM-owned	91	1005	(11.0)	221	(2.4)	284	(3.1)
BOTH-owned	206	595	(2.9)	747	(3.6)	712	(3.5)
MEMA-owned	185	642	(3.5)	858	(4.6)	4532	(24.5)
Total	482	2242		1826		5528	

Genes with cross-validated PSM–gene pairs have been categorized according to the source where the majority of pairs have been found, e.g. for OMIM-owned genes the majority of pairs can be retrieved from OMIM. The numbers show that Medline provides the largest portion of data. For OMIM-owned genes the number of commonly known pairs is low. This can be due to differences in the naming conventions or due to the fact that OMIM uses information sources other than Medline. (Average refers to the number of PSMs per gene.)

indicates which source provides the largest portion of findings (>50%) for PSMs associated to the gene.

Table 4 gives an overview on the PSM–gene pair findings across the different categories. The average numbers for commonly known pairs range from 2.4 to 4.6 pairs per gene for all three categories. As expected, the average number of pairs known to OMIM for OMIM-owned genes is high (11.0 PSM per gene), and the average number of pairs in MEMA for MEMA-owned genes is even higher (24.5 PSM per gene). As a result Medline provides a large portion of PSM–gene pairs and only 18.8% of the genes are owned by OMIM. One interpretation of these data is that a large number of PSMs for a given gene is not of interest to OMIM, since it is not relevant to human disease, they could include, e.g., experimentally induced PSMs or PSMs that are only synonymous variants of existing OMIM entries.

Finally we extracted the date of first publication linked to the extracted PSM–gene pairs and analysed the distribution over time (Fig. 3). We expected to see an increase of published PSM–gene pairs over the past 10 years in Medline, which is the case and which reflects the increase in research work done in this field. In 2001 the number is smaller than that for 2000, since we used for Medline and OMIM the versions published in September and December 2001, respectively, which did not yet contain the full data for 2001. The earliest cited pair is registered in 1971, and the first citation for a cross-validated PSM–gene pair stems from 1984. The largest number of PSM–gene pairs belongs to MEMA-owned genes. Furthermore, the increase of cited PSM–gene pairs is mainly due to PSMs of MEMA-owned genes and to PSMs of BOTH-owned genes.

DISCUSSION

Scientific literature is the most important information source for researchers to publish their findings and to stay informed about the scientific work done in other groups. Electronic access to the information source, e.g. via PubMed for abstracts from scientific literature, and efficient query interfaces, e.g.

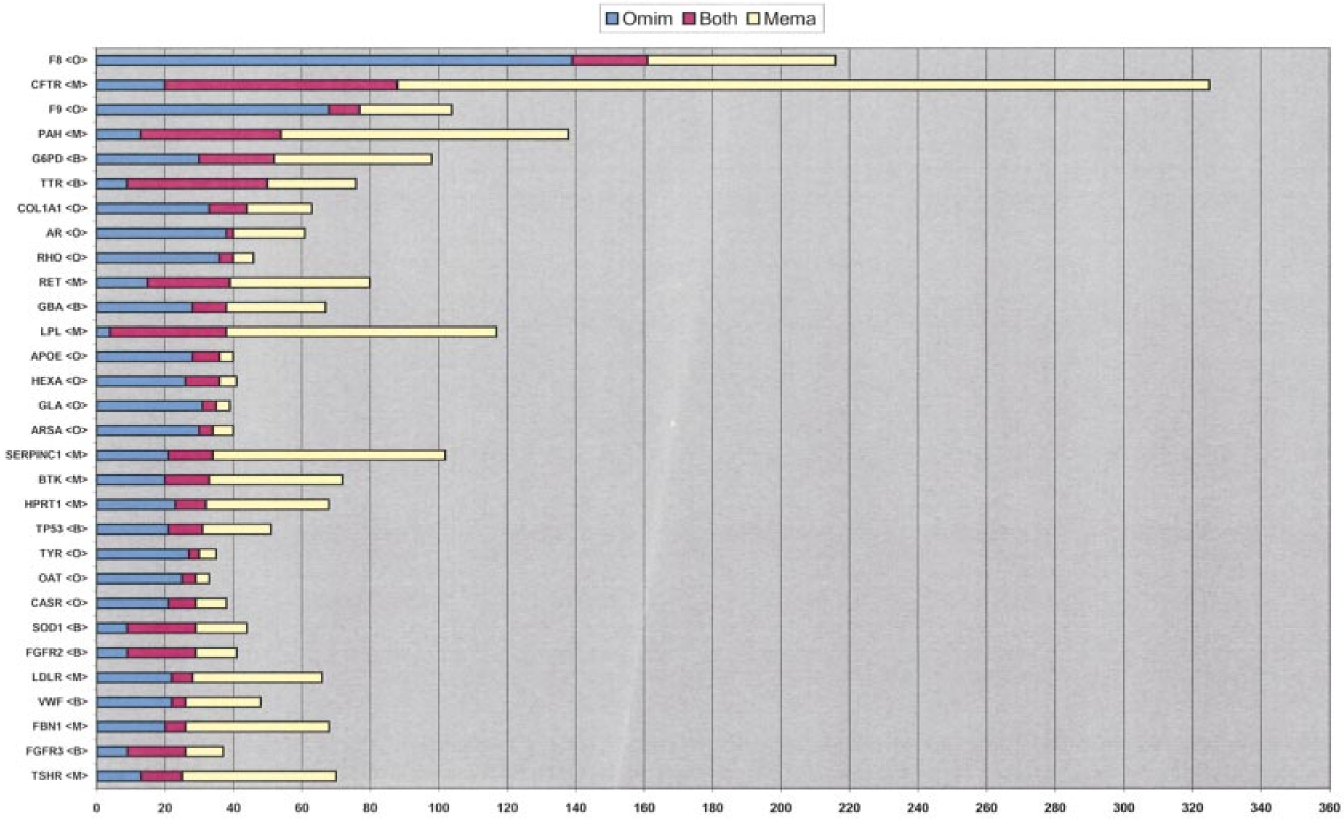

Figure 2. Number of PSM–gene pairs per gene sorted according to OMIM. From top to the bottom different genes are listed, and the blocks to the right represent the number of PSM–gene pairs: first the number of PSM–gene pairs unique to OMIM, then those contained in both databases and finally those listed in MEMA only (Medline). The genes have been sorted according to the number of pairs in OMIM, which is the sum of the first two sections in the block. <O> refers to OMIM-owned genes, and <M> and <B> to MEMA-owned or BOTH-owned genes, respectively (see text). OMIM provides more PSM–gene pairs than MEMA for only 12 out of 30 genes, although these are the top 30 of OMIM. The correlation coefficient for the distribution of mutations per gene found in OMIM and in Medline is 0.53.

the query engine of PubMed, are enabling the rapid retrieval of information, but are not sufficient if a large amount of data has to be compiled and made available for further processing and use. As a consequence, researchers have investigated methods to automatically analyse scientific text and to provide facts from the set of documents in a condensed form (12,13,19,20). In our approach we present the extraction of mutation–gene pairs from Medline abstracts.

We used HUGO nomenclature to detect gene names, which is an important standardization for comparing the extracted data with other sources, e.g. with OMIM. If a mutation was detected and if, in addition, a gene name according to HUGO nomenclature was extracted, then the gene name was used as a synonym for the protein. In addition to our approach, automatic extraction of terminology can be applied to further improve the recall of gene entities (9,11,20).

Precision for the identification of cited mutations as well as contained mutations is high (99 and 98%, respectively), while the recall can be further improved with the help of additional patterns capturing missed representations. The detection of the contained mutation–gene pair relies on the correct identification of at least one gene and one mutation citation, and on the correct association of the two parts. Our methods have proven to be precise (94%), but the recall is rather low (35.2%), which is mainly due to the fact that a large number of abstracts from our sample (35 of 91) contained several gene entities. This is an example of the known fact that the identification of the relationship between concepts is a complex task (21). Our solution is tuned to provide high precision at the expense of lower recall.

The largest portion of mutation–gene pairs from MEMA and in OMIM refers to PSMs (substitution type). Out of the complete set of pairs from OMIM (6699) our method identified 1826 from Medline, which is 27% of recall. It can be assumed that Medline does not cover the full scope of pairs contained in OMIM. Indeed, only for 2002 pairs did we find at least one abstract from Medline containing both the gene and the mutation. The methods applied automatically extracted 1826 PSM–gene pairs, which is 91% of recall. This seems to contradict the recall measured through the sample of abstracts and can be explained by the fact that PSM–gene pairs appear redundantly in different abstracts. In addition, we can expect that the access to and the analysis of complete publications will increase the amount of PSM–gene pairs automatically extracted from the scientific literature.

For 4697 of 6699 PSM–gene pairs from OMIM no abstract can be found in Medline where the PSM and the gene entity co-occur. This leads to the conclusion that information extraction has to deal with the trivial constraint that only contained information can be extracted. Furthermore, OMIM

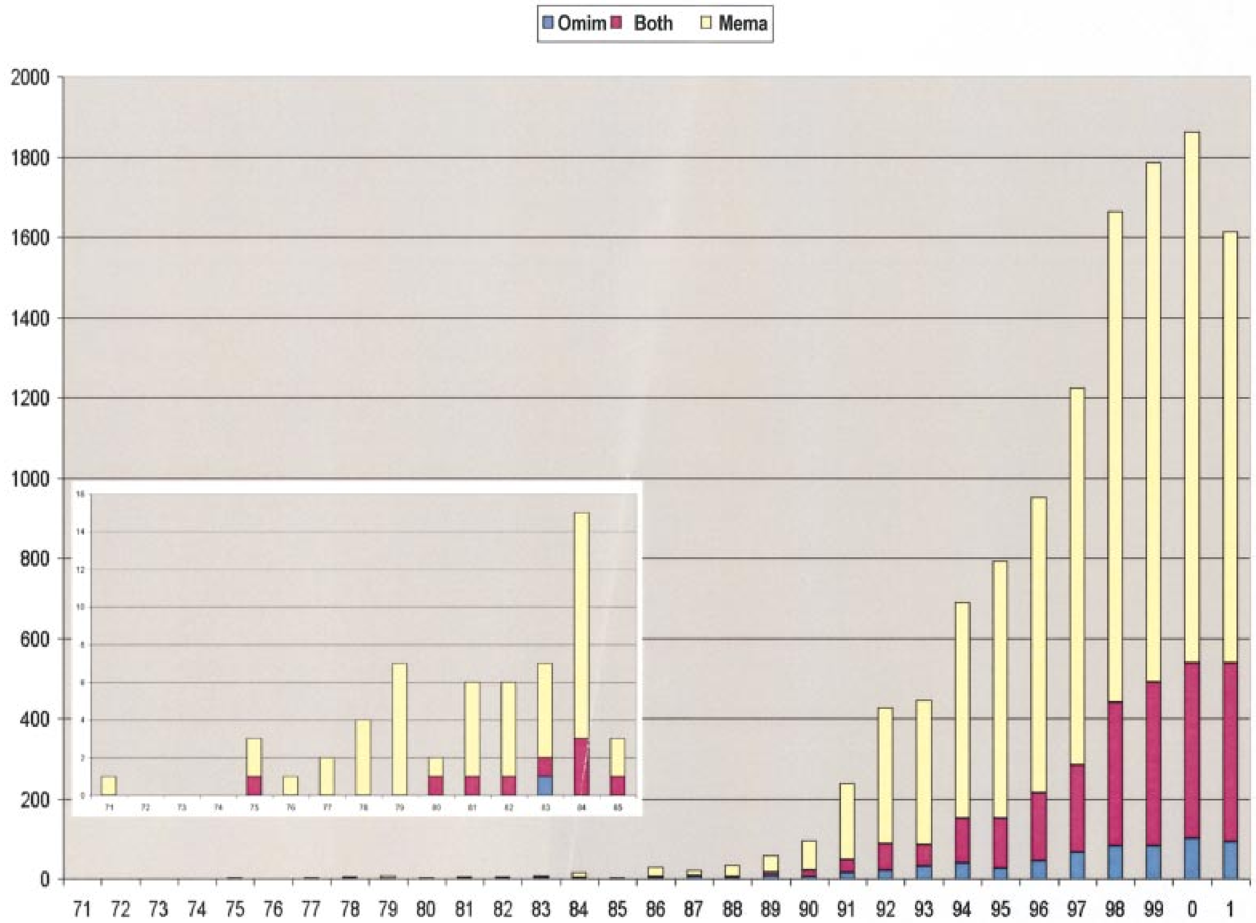

Figure 3. Number of PSMs from 1971 to 2001. The diagram shows the number of PSMs found for genes where at least one PSM–gene pair has been cross-validated. The blocks from bottom to top represent the number of PSM–gene pairs in total for OMIM-owned, for BOTH-owned and for MEMA-owned genes. The inset displays the number of PSMs of the years 1971–1985 at a larger scale (0–16). During the years 1990–2000 a steady increase in PSM–gene pairs takes place. Only a small portion is integrated into OMIM, mainly represented by the PSM–gene pairs of BOTH-owned genes. This is explained by the fact that Medline reports on experimentally induced mutations. Such mutation–gene pairs are not relevant to OMIM, since the evidence for impact to a human disease might not be known or might be unclear (see Discussion).

and Medline differ to a large extent, since each one provides PSM–gene pairs which are not contained in the other source. The top ranked genes for OMIM are listed in Figure 2. Amongst them are F8, F9, Col1A1, AR, RHO, APOE and HEXA. The top-ranked MEMA-owned genes are CFTR, BCHE, PAH, LPL, SERPINA1, PLP1 and C2 (not shown).

The reason for these differences can be found in the interpretation of the purpose of OMIM and Medline. OMIM is focused on human disease and gathers any kind of information which helps to understand the cause of disease. In contrast, Medline contains details of mutagenesis experiments which might not yet be relevant to a type of disease, or which even do not appear in nature. This hypothesis might hold for CFTR.

Nevertheless, Medline abstracts are without doubt an important information source to curators working for OMIM, and the increase in detected and reported mutation–gene pairs over the past 20 years is represented in the increase of PSMs belonging to BOTH PSM–gene pairs (Fig. 3). We conclude that our information extraction methods allow fast access to mutation–gene pairs contained in scientific literature and that such methods complement data-mining methods (22) and efficiently support the work of curators (23), if curation is still wanted.

ACKNOWLEDGEMENTS

This work has been supported by grants from the German Ministry of Research (BMBF-Project 0312385, 'BioPath') and was embedded into the Text Mining Group of the research department at LION Bioscience AG, Heidelberg.

REFERENCES

1. Perutz,M.F. and Lehmann,H. (1968) Molecular pathology of human haemoglobin. *Nature*, **219**, 902–909.
2. Tolle,R. (2001) Information Technology Tools for Efficient SNP Studies. *Am. J. Pharmacogenomics*, **1**, 1–12.
3. Medline database (December 2001) Access through US National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA.
4. Online Mendelian Inheritance in Man, OMIM™ (December 2001) McKusick–Nathans Institute for Genetic Medicine, Johns Hopkins

University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

5. Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **15**, 57–61.
6. Single Nucleotide Polymorphism (dbSNP) Access through US National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA.
7. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
8. Andrade,M.A. and Bork,P. (2000) Automated extraction of information in molecular biology. *FEBS Lett.*, **476**, 12–17.
9. Nenadic,G., Spasic,I. and Ananiadou,S. (2002) Automatic acronym acquisition and management within domain-specific texts. *Proceedings of the Third Conference on Language Resources and Evaluation (LREC-3)*, Las Palmas, Spain, European Language Resources Association (ELRA), France, pp. 2155–2162.
10. Yu,H. and Agichtein,E. (2003) Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, **19** (Suppl. 1), i340–349.
11. Hatzivassiloglou,V., Duboue,P.A. and Rzhetsky,A. (2001) Disambiguating proteins, genes and RNA in text: a machine learning approach. *Bioinformatics*, **17** (Suppl. 1), 97–106.
12. Andrade,M. and Valencia,A. (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *Proc. Intl Conf. ISMB*, **5**, 25–32.
13. Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Intl. Conf. ISMB*, **7**, 60–67.
14. Rindflesch,T.C., Tanabe,L., Weinstein,J.N. and Hunter,L. (2000) EDGAR: extraction of drugs, genes and relations from biomedical literature. *Pac. Symp. Biocomput.*, 517–528.
15. Hopcroft,J.E., Motwani,R. and Ullmann,J.D. (2001) *An Introduction to Automata Theory, Languages and Computation*. Addison-Wesley Publishing Co., Reading MA, ISBN 0-201-44124-1.
16. Roche,E. and Schabes,Y. (1997) *Finite-State Devices for Natural Language Processing*. MIT Press, Cambridge, MA.
17. HUGO Gene Nomenclature Committee (January 2002) Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.
18. White,J.A., McAlpine,P.J., Antonarakis,S., Cann,H., Eppig,J.T., Frazer,K., Frezal,J., Lancet,D., Nahmias,J., Pearson,P. *et al.* (1997) Guidelines for Human Gene Nomenclature. *Genomics*, **45**, 468–471.
19. Marcotte,E.M., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.
20. Proux,D., Rechenmann,F., Julliard,L., Pillet,V. and Jacq,B. (1998) Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 72–80.
21. Blaschke,C., Hirschman,L. and Valencia,A. (2002) Information extraction in molecular biology. *Brief. Bioinformatics*, **3**, 154–165.
22. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genet.*, **31**, 316–319.
23. Albert,S., Gaudan,S., Knigge,H., Raetsch,A., Delgado,A., Huhse,B., Kirsch,H., Albers,M., Rebholz-Schuhmann,D. and Koegl,M. (2003) Computer-assisted generation of a protein-interaction database for nuclear receptors. *J. Mol. Endocrinol.*, **17**, 1555–1567.