BIOINFORMATICS

Vol. 20 no. 4 2004, pages 557–568
DOI: 10.1093/bioinformatics/btg449

Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors

Florence Horn[1,*], Anthony L. Lau[1] and Fred E. Cohen[1,2]

[1]Department of Cellular and Molecular Pharmacology and [2]Department of Biochemistry and Biophysics, University of California of San Francisco, Genentech Hall, Box 2240, 600 16th Street, San Francisco, CA 94143, USA

Received on August 8, 2003; accepted on September 29, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: The amount of genomic and proteomic data that is published daily in the scientific literature is outstripping the ability of experimental scientists to stay current. Reviews, the traditional medium for collating published observations, are also unable to keep pace. For some specific classes of information (e.g. sequences and protein structures), obligatory data deposition policies have helped. However, a great deal of other valuable information is spread throughout the literature hindering coherent access. We are involved in the Molecular Class-Specific Information System (MCSIS) project, a collaborative effort to design and automate the maintenance of protein family databases. The first two databases, the GPCRDB and NucleaRDB, are focused on G protein-coupled receptors (GPCRs) and nuclear hormone receptors (NRs), respectively. The main aim of the MCSIS project is to gather heterogeneous data from across a variety of electronic and literature sources in order to draw new inferences about the target protein families.

Results: We present a computational method that identifies and extracts mutation data from the scientific literature. We focused on the extraction of single point mutations for the GPCR and NR superfamilies. After validation by plausibility filters, the mutation data is integrated into the corresponding MCSIS where it is combined with structural and sequence information already stored in these databases. We extracted and validated 2736 true point mutations from 914 articles on GPCRs and 785 true point mutations from 1094 articles on NRs. The current version of our automated extraction algorithm identifies 49.3% of the GPCR point mutations with a specificity of 87.9%, and 64.5% of the NR point mutations with a specificity of 85.8%. MuteXt routinely analyzes 100 electronic articles in approximately 1 h.

Availability: Extracted results are available via the GPCRDB and NucleaRDB at http://www.gpcr.org/7tm/mutation/ and http://www.receptors.org/NR/mutation/, respectively. The algorithm is available upon request.

Contact: horn@cmpharm.ucsf.edu

INTRODUCTION

Biologists have come to appreciate that comprehensive databases can accelerate their research efforts. Historically, populating these databases has been challenging since contributions to the traditional scientific literature, and not specialized databases, have been the benchmarks by which funding support and career advancement decisions are made. Increasingly, fast and ready access to biological data has become a research priority. Much of the scientific literature is available online in abstract format, and several journal publishers provide complete articles in HTML and/or PDF. While this allows for the rapid and voluminous distribution of published data, a significant problem remains. How is a scientist to find specific information without having to retrieve and integrate a plethora of data from a large number of papers? Comprehensive databases offer a solution to this problem, but it is difficult to organize scientific curators to create and maintain these resources.

Experimental data such as gene and protein sequences and protein structures are easily available via comprehensive databases. By contrast, mutation and ligand binding information remains less accessible. Site-directed mutagenesis is a commonly used experimental tool. Researchers often need to identify mutations that have already been reported not only to avoid duplication of efforts but also to plan insightful structure–function studies. Frequently, experimental results, such as mutation data, are found in both tabular and textual material and thus elude traditional keyword searches of the primary literature. Instead, a fastidious bibliographic

*To whom correspondence should be addressed.

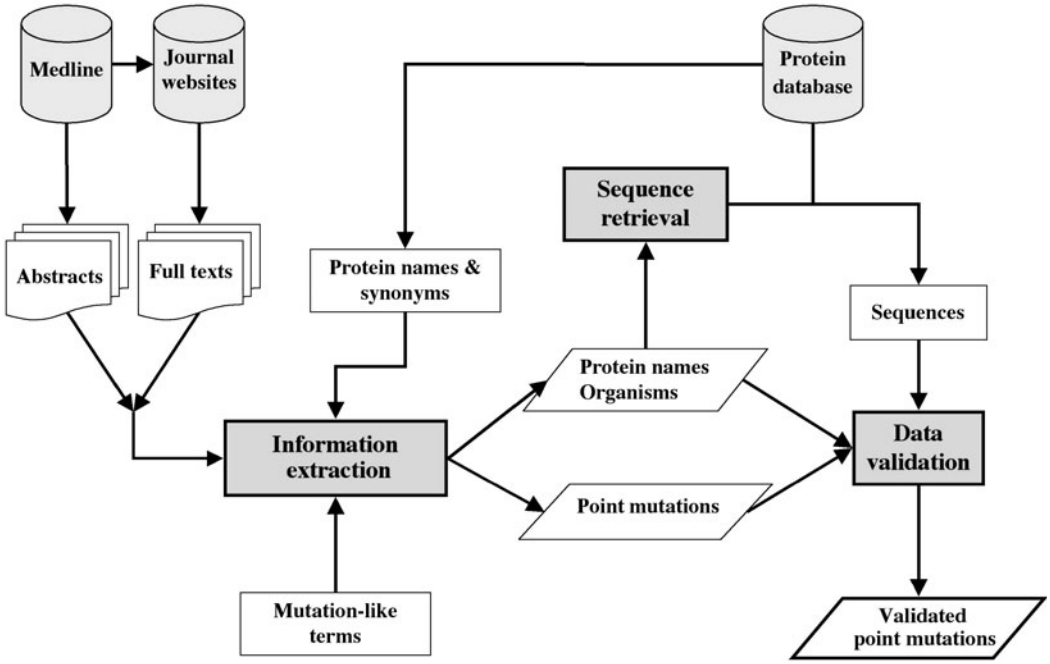

Fig. 1. The flow chart summarizes the main tasks for the extraction of mutation data by the MuteXt program. Citations are selected from Medline and the full texts are imported from journal websites. Each paper is processed to identify and extract protein names, organism types and point mutations. Point mutations are then validated by plausibility filters.

search is required, and then the articles identified must be scrutinized. Several solutions can be envisaged to address this problem. First, would be to require the authors to submit mutation data, or other experimental data to a database, prior to publication. This is done routinely for DNA sequences, and protein sequences and structures. The approach is feasible, but requires that authors, editors and publishing groups accept and adhere to submission guidelines. A more communal approach would be to invite authors to submit published data in electronic form. However, previous attempts to solicit voluntary submissions into a database on G protein-coupled receptor (GPCR) mutations, tinyGRAP (Edvardsen *et al*., 2002), received very limited response. To date, the only effective way to collect mutation data is to extract it by hand from published manuscripts.

The solution we are currently working towards is the automated extraction of relevant experimental data from electronic literature sources using computational methods. Over the past several years, technologies in information extraction have emerged to process unstructured textual information and to extract specific information. These techniques include regular expression matching, co-occurrence of terms, statistical methods, advanced parsing and hidden Markov models. Several works have been published that explore a variety of innovative techniques to extract biological knowledge from the literature, in particular Medline abstracts. Information extraction has been utilized to identify gene and protein names (Fukuda *et al*., 1998; Proux *et al*., 1998; Leonard *et al*., 2002; Tanabe and Wilbur, 2002), molecular interactions or relationships between substances (Blaschke *et al*., 1999; Rindflesch *et al*., 2000; Proux *et al*., 2000; Humphreys *et al*., 2000; Thomas *et al*., 2000; Yoshida *et al*., 2000; Marcotte *et al*., 2001; Ono *et al*., 2001), specific keywords (Andrade and Valencia, 1997; Ohta *et al*., 1997; Andrade and Bork, 2000), protein location (Craven and Kumlien, 1999) and, recently, the roles of residues in protein molecules (Gaizauskas *et al*., 2003). Natural language processing (NLP) is also used to analyze and parse the text content (for a review, see Hirschman *et al*., 2002). To our knowledge, the application of information extraction to mutation data has yet to be reported.

We have implemented a method of regular expression matching, called MuteXt, which identifies and extracts point mutations from the scientific literature. The extracted data is validated via plausibility filters, stored in a database and made accessible on the World Wide Web (WWW). The program was first trained on GPCR point mutations and then applied to nuclear hormone receptors (NRs).

SYSTEMS AND METHODS

There are three separate tasks involved in automating information retrieval from the scientific literature: document retrieval, information extraction and data validation (Fig. 1).

Document retrieval

Documents are retrieved from the Medline database using the PubMed query system (Schuler *et al*., 1996). Protein family names and the keywords 'mutagenesis', 'mutant' and 'mutation' were used as search criteria. The algorithm then searches for the availability of a full text version for each matching citation using the PubMed E-link system (http://www.ncbi.nlm.nih.gov/entrez/query/static/elink_help.html). This system indicates whether a full text version is provided by the journal publisher and gives, if appropriate, the URL(s) of where the article can be downloaded in HTML and/or PDF. Medline abstracts are used when full texts are not available. HTML and PDF files are converted into text files using the programs 'html2text' (T. Ibbs, http://starship.python.net/crew/tibs/python/html2text) and 'pdftotext' (Glyph and Cog, LLC, http://www.foolabs.com/xpdf/), respectively.

Information extraction

Input data

The input data consists of abstracts and full text articles in plain text, Swiss-Prot entries (Boeckmann *et al*., 2003) for the target family, a dictionary of protein names and a list of mutation-like terms.

Sequence information In the two examples we describe here, the Swiss-Prot entries are already present in two protein family databases that we previously implemented: the GPCRDB (Horn *et al*., 2003) and the NucleaRDB (Horn *et al*., 2001). For a new protein family, one can also use MuteXt to import Swiss-Prot entries.

Name dictionary The dictionary is generated by an automated compilation of the description (DE) and gene name (GN) lines found in the set of Swiss-Prot entries for the protein family of interest. When several names (synonyms or old nomenclature) are given for a single protein, each different term is stored as a separate entry. To avoid confusion, non-specific entries such as 'protein', 'orphan receptor', 'fragment', are automatically removed. In addition, we have manually created a dictionary of synonyms for several GPCRs and NRs because commonly used names of proteins are not always captured in the Swiss-Prot entries. For example, the calcium-sensing receptor is sometimes designated 'Ca+ receptor' or 'calcium receptor' by authors.

Point mutation-like terms We have looked for generic expressions used to define point mutations and manually created a list of exceptions to avoid mislabeling other phrases as mutants. Typically, a point mutation in a protein is described by a term that indicates the wild-type amino acid (in one- or three-letter code), the position in the sequence and the mutant amino acid (also in one- or three-letter code) (den Dunnen and Antonarakis, 2001). However, terms with the one-letter/number/one-letter structure are also used in reference to other objects such as cell lines, drugs and materials used in experiments (e.g. 'T47D' is a cell line and M24R a filter). We have created a list of known point mutation-like terms with their true contextual meanings. For example, the term 'T47D' will be ignored if it is found close to the terms 'cell line', 'tumour', 'tumor' or 'cancer'. This list will have to be updated manually as other point mutation-like terms are detected in the documents subsequently processed. However, the same list can be used for different protein families.

Extraction of point mutations

MuteXt searches for the presence of mutation data in each document (abstract and converted HTML and PDF documents). We use the method of pattern matching with regular expressions to identify point mutations. The pattern must start with one amino acid in the one- or three-letter code followed by a number, and optimally by another amino acid encoded with the same letter code format as the first one. The regular expression we use is:

$$([A–Z][1–9][0–9] + \$)|([A–Z][1–9][0–9] * [A–Z]\$)$$
$$|([A–Z][a–z][a–z][1–9][0–9] * \$)$$
$$|([A–Z][a–z][a–z][1–9][0–9] * [A–Z][a–z][a–z]\$)$$

where [A–Z] and [A–Z][a–z][a–z] must belong to set of 20 amino acids in one- or three-letter code, respectively.

The documents are first split into sentences and words, and non-alphanumeric characters are removed from the words, as point mutations may include non-alphanumeric characters [e.g. 'A234-T', 'A(234)T']. Terms that match the regular expression are then retained for further analysis. When the second amino acid is not indicated (e.g. Ala234), the 10 words that follow are scanned for the presence of the full name or three-letter code of another amino acid. Each identified point mutation is checked against the list of point mutation-like terms. In addition, when both letters in mutations cited in one-letter code belong to the nucleotide code set [A, C, G, T], words surrounding the point mutations are analyzed to avoid extracting DNA or RNA mutants.

Extraction of protein names

MuteXt looks for protein names in the abstract of each article, using the name dictionary and the list of synonyms. The whole article is not considered, because other proteins that are not being studied are often cited in the Introduction and Discussion sections. The regular expression we use is:

$$`\backslash W' + \mathrm{sub}(\text{“}[\text{-}]\text{”},\text{“}[\text{-}]^{*}\text{”}, \mathrm{DE}) + `s * \backslash W'$$

where DE is an entry of the name dictionary.

When no name is found in the abstract, MuteXt looks for protein names in the MeSH terms of the Medline record.

Extraction of organism types

The species names of all the organisms indicated in the Swiss-Prot entries of the target family are searched in all sections of

Table 1. An example of data validation using text distances between point mutations, protein names and organisms

(a) Information extracted from a full text article	
PubMed identifier	10891484
Point mutations extracted	E403K, E453K, E456K, E457K, E460K, F450A
Molecule names identified	Peroxisome proliferator activated receptor, retinoid X receptor, RXR, thyroid hormone receptor, VDR, vitamin D3 receptor
Synonyms identified	TR, T3 receptor, T3R
Organisms identified	human, mouse ('mice' listed in Medline MeSH terms)
Relevant sentences	The **human retinoid X receptor** alpha (**hRXR** alpha) AF-2 mutant (**F-450-A**, **E-453-K**, **E-456-K**) was generated by introducing site-directed mutations into the open reading frame of pFLAG-**RXR** alpha(17), using PCR (. . .)
	Specific AF2 mutants used in this experiment include **TR** alpha-AF2mt (**hTR** alpha, **E-403-K**), **TR** alpha-Delta 4 (C-terminal deletion of helix 12, Delta 401) and **hRXR** alpha-AF2mt (**F-450-A**, **E-453-K** and **E-456-K**)
	Similarly, the AF2 mutants of **hTR** alpha (**E-403-K**) and **hTR** beta (**E-457-K**, **E-460-K**) were generated by introducing site-directed mutations into the open reading frames of pFLAG-**hTR** alpha and pFLAG-**TR** beta, using PCR (. . .)
Shortest word distances for the pairs of terms that co-occur in the sentences.	E403K/RXR: 13, E403K/TR: 2, F450A/RXR: 3, F450A/TR: 14, E453K/RXR: 5, E453K/TR: 16, E456K/RXR: 6, E456K/TR: 16, E457K/TR: 2, E460K/TR: 3, RXR/human: 0, TR/human: 0 F450A/human: 9, E453K/human: 10, E456K/human: 11

Application of the plausibility filters		Validation status	Output of the distance filter (matching term pairs)	Final validation status
(b) Output of the sequence filter: mutation/Swiss-Prot entry				
E403K[a]		**NO**		
F450A	RXRA_HUMAN	**OK**		
E453K	NRH4_HUMAN	MA	(E453K,human)	
	RXRA_HUMAN	MA	(RXR,human), (E453K,human)	**OK**
E456K	RXRA_HUMAN	**OK**		
E457K	RXRG_HUMAN	MA	(RXR,human)	
	RXRG_MOUSE	MA		
	THB1_HUMAN	MA	(TR,human), (E457K,TR)	**OK**
	THB1_MOUSE	MA	(E457K,TR)	
E460K	THB1_HUMAN	MA	(TR,human), (E460K,TR)	**OK**[b]
	THB1_MOUSE	MA	(E460K,TR)	
	THB2_HUMAN	MA	(TR,human), (E460K,TR)	
E488K[c]	PPAT_HUMAN	MA		

[a] E403 only exists in the human thyroid hormone receptor isoform alpha-1, not displayed in the corresponding Swiss-Prot entry.
[b] THB2_HUMAN is ignored because it matches only one mutation.
[c] E488K corresponds to E460K with an offset of +28 in PPAT_HUMAN.

the documents and in the MeSH terms of the Medline records. MuteXt also considers different ways to describe the same organism. For example, 'human' can be designated using different terms, such as 'patient, Caucasians, woman, children'. In addition, organism names found near the words 'serum' and 'cells' are ignored (e.g. bovine serum, Chinese hamster ovary cells). If no organisms are found in the document, a list of most often studied organisms is used for subsequent data validation.

Data validation

We apply two different plausibility filters to validate the extracted point mutations.

Sequence filter The first filter consists of checking whether the wild-type amino acids (e.g. Ala in 'A234T') in the extracted point mutations are found at the indicated positions in the corresponding sequences. For each article, Swiss-Prot entries are selected using the extracted organism type(s) and protein name(s). Point mutations are checked for their wild-type residue position in the selected sequence(s). If the position does not match any of the selected entries, the point mutation is labeled 'NO'. When several sequences match the residue position, the point mutation is labeled 'MA' for 'maybe'. If the residue is found in one unique Swiss-Prot entry, the mutant is validated with the label 'OK' (Table 1). We have allowed for alternative residue numbering for some proteins

that present several isoforms or a signal sequence. Swiss-Prot entries often display the longest sequence and the sequence variations are indicated in the feature section. This often leads to a difference of sequence lengths and residue numbering between proteins used in experimental studies and proteins described in Swiss-Prot. For example, point mutations for the rat lutropin receptor were correctly validated after using an offset of +26 to the residue numbering used in several publications.

Distance filter When several sequences remain after the first filter, word distances between the different extracted data (i.e. point mutations, protein names and organism types) are estimated two by two when they occur in the same sentence. All the words are counted. The algorithm recognizes terms like 'hTR' where the initial of the organism is merged with the protein name. The shortest distances are kept in order to obtain pairs of terms [e.g. '(TR, human)', '(E457K, TR)', '(E453K, human)'] that often reduce the number of possible sequences for each 'MA' point mutation. If several Swiss-Prot entries are still plausible, the number of point mutations for each Swiss-Prot entry is then considered. If only one sequence remains after the application of the filter, the label 'OK' is attributed to the point mutation (Table 1).

Evaluation of the performance of MuteXt

There is no gold standard for the list of all mutations made for a given protein or protein family. Thus, we compared our algorithm to the accuracy of human curators over a reasonable subset of the articles identified with our Medline queries. Validated and non-validated point mutations were examined manually for all the selected articles in order to measure the performance of the method and to refine the information extraction step and the plausibility filters. Each point mutation was labeled true positive (TP), false positive (FP), true negative (TN) or false negative (FN). The effectiveness of MuteXt is measured in terms of sensitivity (or recall), specificity (or precision), error rate and accuracy:

- –Sensitivity is the percentage of relevant point mutations identified by MuteXt: $TP/(TP + FN)$.
- –Specificity is the percentage of correct point mutations validated by the system: $TP/(TP + FP)$.
- –Error rate is the percentage of wrong decisions made by MuteXt: $(FP + FN)$/total extracted.
- –Accuracy is the percentage of correct decisions made by MuteXt: $(TP + TN)$/total extracted.

The current version of the program is written in Python v2.2. The content of the Medline entries, full text versions (PDF and HTML) and all the information extracted from the documents are stored in a MySQL database.

Table 2. Summary of the results obtained by application of MuteXt to GPCRs

	tinyGRAP full texts	Information retrieval MuteXt		Sensitivity all formats
		Full texts	Abstracts	
Retrieved articles	914	914 382	532	100% (41.8%)[a]
Relevant articles	914	732 372	360	80.1% (40.7%)[a]
Extracted point mutations	5451	3996 (2685)[b] 3074 (1935)[b]	922 (750)[b]	49.3% (69.5%)[a]

[a]Full texts only.
[b]Mutations present in the tinyGRAP control dataset.

RESULTS

G protein-coupled receptors

Our first application of MuteXt was to GPCRs, a pharmacologically important and evolutionary diverse family of proteins. The GPCR superfamily is an opportune set because of the existence of a manually curated collection of mutation data stored in the tinyGRAP database (Edvardsen *et al.*, 2002). This provides us with an independent control on the sensitivity and specificity of our algorithm. The single point mutations of tinyGRAP v6.0 were used as a control dataset (data kindly provided by O. Edvardsen). This set consists of 5451 point mutations manually collected from 914 papers published between 1987 and 2000. Of the 914 citations, MuteXt retrieved only 382 unique full texts: 368 in HTML, 337 in PDF and 323 articles in both formats (Table 2).

We extracted a total of 3996 point mutations from the selected documents. Only 2685 are present in the control dataset. Thus, we retrieved 49.3% of the 5451 point mutations of tinyGRAP. A large fraction of the missing mutations relates to the fact that many full text articles were not electronically accessible (Fig. 2). Point mutations are rarely cited in the abstract section. In fact, we obtain a sensitivity of 69.5% if we consider only articles available as full texts (Table 2). The manual check also showed us that some mutations missed by MuteXt were in fact described in figures. HTML provides access to the text while PDF allows access to tables and text. Neither method provides routes to read the figures. Other mutants are described in natural language with no explicit citation recognizable by MuteXt. For example, the phrase 'we performed an alanine scanning of all the threonine or serine residues between positions 150 and 350 in the adenosine type 1 receptor' is easy for a human to interpret but this is much more challenging for a computer application. In other cases, the numbering used did not correspond to the sequence numbering used in the Swiss-Prot entries. Finally, a few point

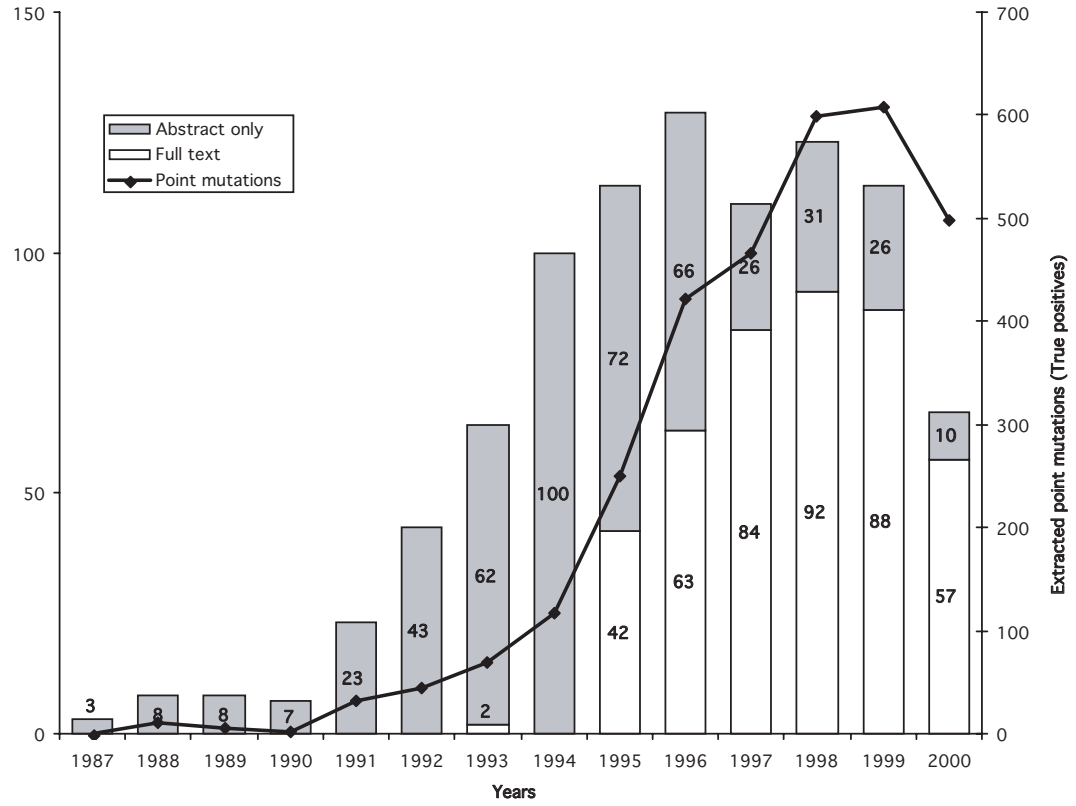

Fig. 2. Distribution of articles in the tinyGRAP dataset that are electronically available as full text or as abstract only, per year of publication. The line plots the number of TP point mutations extracted by MuteXt.

mutants were missed because they were written with a nomenclature that we voluntarily ignore, such as 'AT234' or 'T234' instead of 'A234T'.

The scientific literature is both self correcting and redundant. Thus, it is not surprising that some identical point mutations are described in different articles. This redundancy allows us to 'rescue' some point mutations that MuteXt fails to extract from the original papers. Among the 5451 point mutations in the control dataset, MuteXt extracted 2254 point mutations correctly and 484 'partial' point mutations for which either the sequence or the mutant amino acid could not be identified (Fig. 3). Out of the 2713 missed point mutations, 542 were found in other articles present in the tinyGRAP dataset. The processing of newly imported articles allowed us to rescue an additional set of 258 point mutations. If we consider that point mutations with no mutant amino acid identified (i.e. A234 instead of A234T) may rescue point mutations that MuteXt failed to validate, a total of 915 point mutations have been rescued from 271 articles (Fig. 3). Some point mutations have been rescued several times. For example, the mutation D578G in the human Lutropin receptor has been found in 11 articles.

The point mutations identified by MuteXt but not included in the tinyGRAP dataset occur for the following reasons: (i) they are present in the tinyGRAP database but not in the control dataset because they are considered either as part of multiple substitutions, or as chimeric mutations; (ii) they are cited in the processed texts but were originally described in (and thus associated with) other publications and (iii) they are FP and TN point mutations.

Of the extracted mutations for GPCRs, 77.9% were validated. Manual checking revealed that 87.9% of these point mutations were TPs (Table 3 and Fig. 4). Most of the correct mutations that MuteXt could not validate (FNs) correspond to point mutations for which none or several Swiss-Prot entries were retrieved. This is mainly due to molecule names that are not recognized in the text because of their spelling. Indeed, several receptor families are composed of different subtypes such as alpha, beta and gamma. In HTML and PDF format, these subtypes are often written using Greek symbols that are encoded by images. During the conversion into text, this subtype information is often lost, especially from PDF documents. The program 'pdftotext' replaces images by a blank space and consequently, MuteXt may fail to identify the complete molecule name. In HTML, the type of Greek letter is sometimes indicated in plain text within the HTML tag and is kept during the conversion into text. For example, the HTML phrase 'adrenergic

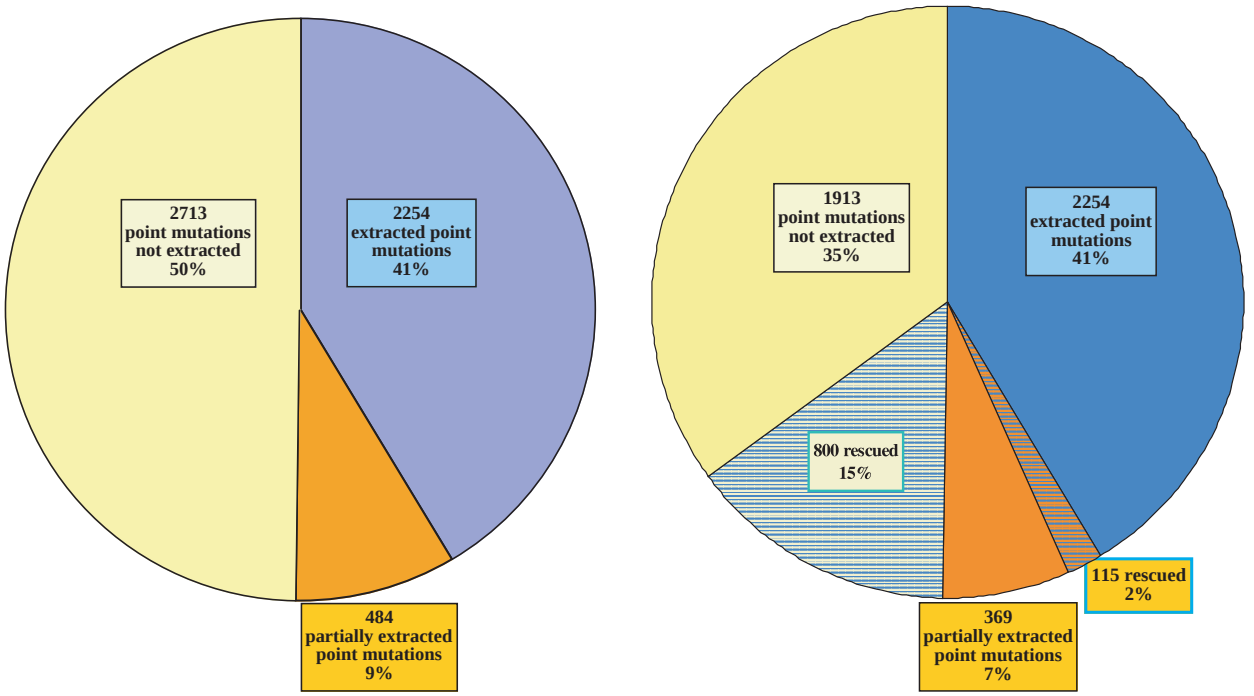

Fig. 3. The left chart shows the distribution of points mutations present in the tinyGRAP control dataset that MuteXt extracted or missed. The right chart indicates the number of point mutations that have been extracted from other articles (i.e. 'rescued' mutations, blue stripes and frames).

Table 3. Results obtained for GPCRs point mutations after validation by plausibility filters and manual checks

	Information extraction			
	Abstracts	HTML full texts	PDF full texts	Total[a]
Articles	360	368	337	732
Point mutations	922	2448	2829	3996
TP	624 67.7%	1717 70.1%	1905 67.3%	2736 68.5%
FP	82 8.9%	175 7.1%	277 9.8%	375 9.4%
TN	87 9.4%	305 12.5%	366 12.9%	480 12.0%
FN	129 14.0%	251 10.3%	281 9.9%	405 10.1%
Sensitivity[b]	82.9%	87.2%	87.1%	87.1%
Specificity	88.4%	90.8%	87.3%	87.9%
Error rate	22.9%	17.4%	19.7%	19.5%
Accuracy	77.1%	82.6%	80.3%	80.5%

[a]Results for the three text sources combined altogether.
[b]Sensitivity values do not include the FN point mutations that were not extracted from the articles.

<img src = alpha.gif alt = 'alpha'>1 receptor' will be converted correctly by the program 'html2text'. However, if the attribute 'alt' is not indicated, MuteXt will extract the phrase 'adrenergic 1 receptor' with no indication of the receptor subtype. Therefore, the set of selected Swiss-Prot entries includes all possible subtypes, and if the proteins share a high degree of sequence homology, point mutations could match several or all the selected sequence entries. If a hybrid manual/algorithmic curation strategy were used, these ambiguous situations would be targeted for manual curation. Alternatively, the 'pdftotext' program could be improved. Other FNs result from the fact that the organism studied is not indicated in the text. Authors sometimes refer to previously published work where details on material, including the organism used, and methods can be found.

While it took more than three weeks for four people to curate manually 190 articles and add the data to tinyGRAP (M.Beukers, personal communication), algorithmic extraction and validation of the 914 articles required <9 h. It should be noted, however, that manual curation can also collect information on the experimental procedures and effects of the mutations.

Nuclear hormone receptors

When we started this work, there was no database specifically focused on assembling mutations for the NR superfamily. Thus, the selection of articles was done using Medline queries exclusively. This led to the identification of 1094 articles. Only 434 unique full texts were retrieved: 401 in HTML, 389 in PDF and 356 in both formats. Of the 1094 articles, 343 were found to contain point mutations (Table 4). The relatively low percentage of relevant documents retrieved shows that the Medline queries we performed were not specific enough. Keywords such as 'mutagenesis' and 'mutation' do not always

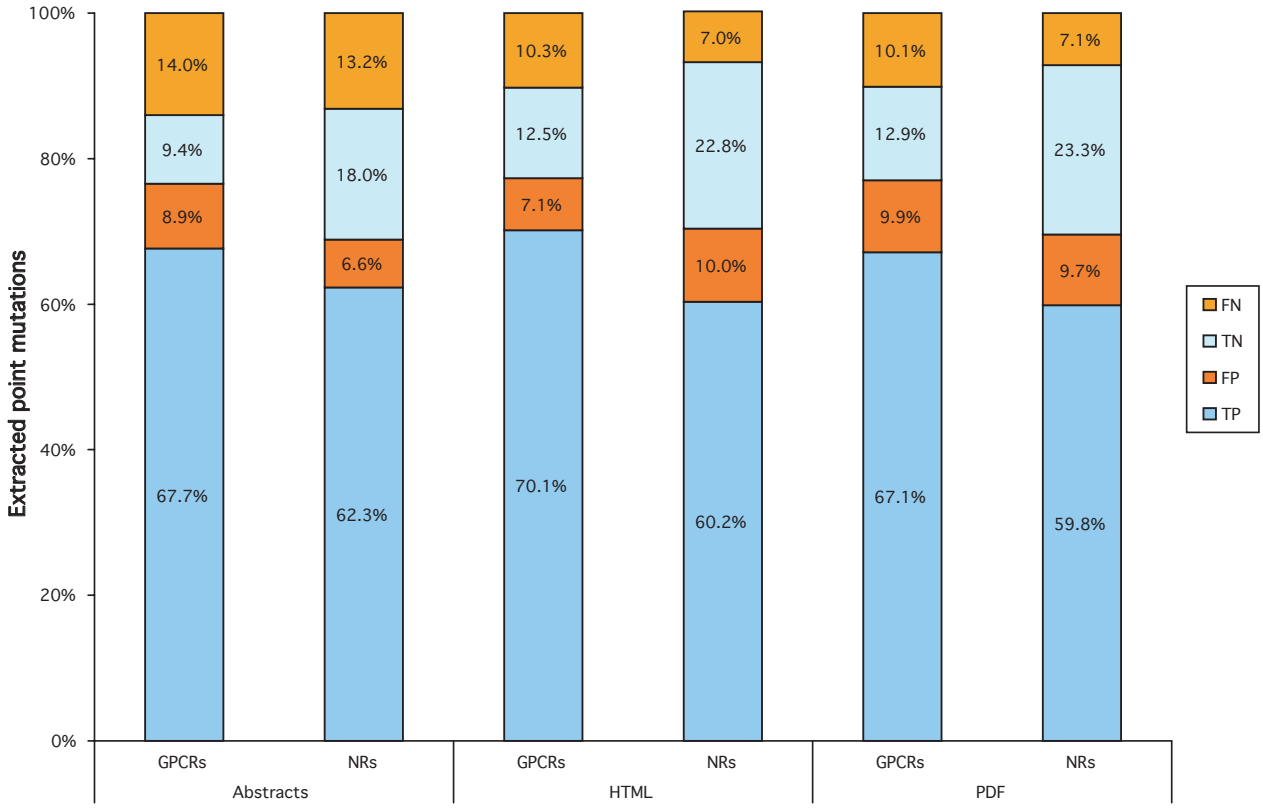

Fig. 4. Percentage of TP and FP, and TN and FN point mutations extracted from abstracts, HTML and PDF full text articles for GPCRs and NRs.

Table 4. Summary of the results obtained by application of MuteXt to NRs

	Information retrieval			
	Full texts	Total	Abstracts	Sensitivity
Retrieved articles	434	1096	662	100% (41.8%)[a]
Relevant articles	270	343	73	n.d.[b]
Extracted point mutations (343 articles)	1171	1338	167	64.5%

[a]Full texts only.
[b]Not determined.

Table 5. Results obtained for NR point mutations after validation by plausibility filters and manual checks

	Information extraction							
	Abstracts		HTML full texts		PDF full texts		Total[a]	
Articles	73		225		258		343	
Point mutations	167		947		1108		1338	
TP	104	62.3%	570	60.2%	663	59.8%	785	58.7%
FP	11	6.6%	95	10.0%	108	9.7%	130	9.7%
TN	30	18.0%	216	22.8%	258	23.3%	323	24.1%
FN	22	13.2%	66	7.0%	79	7.1%	100	7.5%
Sensitivity[b]	82.5%		89.6%		89.4%		88.7%	
Specificity	90.4%		85.7%		86.0%		85.8%	
Error rate	19.8%		17.0%		16.9%		17.2%	
Accuracy	80.2%		83.0%		83.1%		82.8%	

[a]Results for the three text sources combined altogether.
[b]Sensitivity values do not include the FN point mutations not extracted from the articles.

select articles describing mutation data. In addition, a significant number of documents retrieved, while mentioning nuclear receptors, were in fact focused on other molecules, such as ligands or modulators of NRs.

We extracted a total of 1338 point mutations from the 343 articles (Table 5). A manual examination of the errors revealed that difficulties in recognizing the receptor names lead to algorithmic failures. The naming conventions for some nuclear receptors are still in flux and efforts to make these names 'backwards' compatible have fallen short of our goal. The retrieval of receptor names can be improved by a more complete dictionary of synonyms. As with the GPCR family,

the other point mutations that were not identified are described in figures or in natural language, or are indicated with a non-standard numbering system.

MuteXt validated 68.4% of the extracted mutations for NRs. The manual check determined that 85.8% of the validated point mutations are TPs (Table 5 and Fig. 4). Most of the FNs result from the high sequence homology between NRs from the same subfamily. Therefore, as with the GPCRs, several sequences are selected because organism types or receptor subtypes are not found in the text. The percentage of TNs is significantly higher for NRs than GPCRs. It appears that a portion of the point mutations extracted reflect mutations of the binding partners of the NRs, and not the NRs' directly. In most of these cases, the partners are co-regulators of NRs, heat-shock proteins or other transcription factors. The implementation of another filter could reduce the number of TNs by looking at the distances between point mutations and known partners of NRs. However, as more than 90% of these point mutations are not validated by MuteXt, such a filter is not absolutely required for our method to be effective.

Data integration

True positive point mutations are integrated into the corresponding protein database (GPCRDB: http://www.gpcr.org/7tm/mutation/ and NucleaRDB: http://www.receptors.org/NR/mutation/), where they are combined with sequence and structural information. Mutation data for GPCRs can be searched via a simple query system. One can search data by receptor family or subfamily, domain (e.g. transmembrane domain 3) or residue numbering (Swiss-Prot, GPCRDB and Ballesteros–Weinstein numbering systems). Extracted point mutations for NRs can be searched via the same simple query system, as well as through the NRMD (Van Durme *et al*., 2003), a newly created database of NR mutations, which collects mutation data from different resources and includes our extracted data.

In both databases, mutation data is accessible via a suite of different HTML pages. These include the list of proteins for which point mutations are available or the list of all the point mutations found at the same position in a family or subfamily of receptors. Sentences that contained the point mutations are also displayed (Fig. 5). Each HTML page provides links toward the Swiss-Prot entry, the Medline citation, the family page in the GPCRDB or NucleaRDB, and other cross-references to internal and remote information resources. Where appropriate, point mutations are numbered using the general numbering system used in the GPCRDB and in the NucleaRDB. The general numbering facilitates the retrieval of all point mutations that are located in the same structural domain or at a given position, and therefore, allows for better harvesting of the mutation information. Point mutations are also indicated in multiple sequence alignments of each cognate protein family and subfamily (see http://www.receptors.org/NR/seq/003_003/003_003.MSF.mutant.html, for an example). This allows users to compare point mutations between different related proteins and organisms.

DISCUSSION

The retrieval of point mutations is mostly limited by the availability of full text articles. This is an intrinsic limitation of the MuteXt approach, but this relative deficiency should shrink as more publications are made available online in HTML or PDF versions. For example, in the GPCR dataset, only two full text articles were retrieved for the 256 articles published prior 1995, and full text versions are available for 85% of the articles in the dataset published in 2000 (Fig. 2). Currently, most articles published after July 1995 are made available online in full text. Several publishers are providing full text versions of older articles. Unfortunately, these are distributed as images embedded in PDF that MuteXt cannot read. However, given the trend toward electronic publication, our ability to access the entire contents of articles will increase.

MuteXt extracts all the point mutations it finds in the text without distinction between the ones described by the authors and those only cited and previously published. It is possible to make this distinction but we believe that these 'extra' point mutations provide a mechanism for recovering point mutations that might have been missed by MuteXt in the original citation(s) (Fig. 3). The automated identification and processing of the corresponding original citations is possible and may be effective in the next release of MuteXt.

PDF files are an excellent source of point mutations. Indeed, a large number of publications gather point mutations into tables. Tables are generally in image format in the HTML full text versions and therefore not readable by MuteXt. Consequently, PDF files are the only way to access the contents of these tables. However, the conversion of PDF files into text leads to two main problems. Sentences are often truncated due to the insertion of figures, boxes or footers. This decreases the efficiency of the validation filters for the estimation of word distances between terms. The second problem owes to the conversion of Greek letters (see the Results section) and arrows. Arrows are sometimes used to describe a point mutation but the pdftotext program converts them into a '3'. For example, the point mutation 'Ala23→Thr' will be converted into 'Ala233Thr'. Fortunately, most of these fictitious point mutations are eliminated by the validation filters. The problems due to the conversion of PDF files can be compensated for by processing the corresponding HTML full texts, if both are available. Therefore, the combination of both HTML and PDF formats allows for an optimal identification of point mutations.

This work points out an important problem related to the establishment and use of nomenclature, for both point mutations and molecule names. The different ways by which some mutations and proteins are named can be confusing to both humans and computer programs. A nomenclature has been established for the description of point mutations

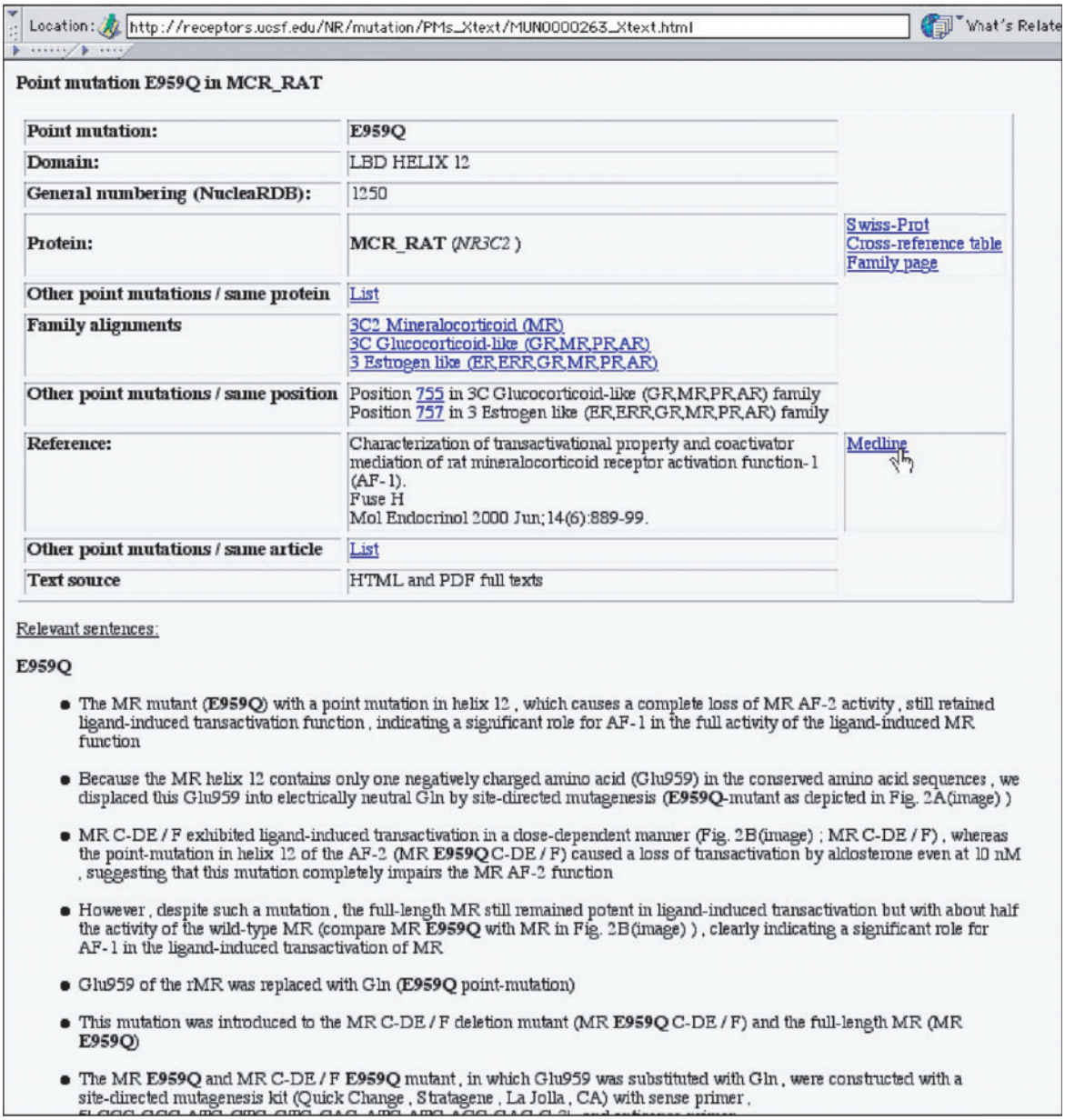

Fig. 5. An example of a NucleaRDB page that displays information on a point mutation extracted by MuteXt. The top of the page describes the point mutation and provides hyperlinks to different mutation-related pages, as well as to the corresponding Swiss-Prot entry, Medline record and NucleaRDB pages. The bottom of the page displays sentences from the article where the point mutation was described.

(den Dunnen and Antonarakis, 2001), but it appears that this nomenclature is not always followed. Frequently, point mutations are described differently by authors from different fields. For example, some biochemists use 'AT234' instead of 'A234T', and some structural biologists use 'T234' to describe a receptor mutant. The first mutant is not detected by MuteXt in its current implementation and the second one cannot be validated because the first letter does not correspond to the wild-type amino acid at the indicated position in the corresponding sequence. These issues arose in about 2% of the articles we analyzed. In addition, we found several examples where MuteXt extracted point mutations that were, in fact, typographical errors. Often these errors resulted from an interchange of letters and numbers of real point mutations, which led to FPs and TNs.

Most researchers utilize the sequence numbering indicated in sequence databases. However, we faced some exceptions where authors used a different numbering scheme. In most of these cases, this is due to the existence of splicing variants or isoforms. We have resolved this numbering problem

by allowing alternative numbering for such cases. Unfortunately, in several articles, the numbering used by the authors was impossible to correlate with the Swiss-Prot numbering. In these cases, we could search for the correct numbering by automatically scanning sequences of Swiss-Prot entries to find the one(s) that match all the indicated amino acids. This approach is currently being implemented but will be applied to articles that describe more than five point mutations to avoid fortuitous matches.

The multiplicity of protein names necessitates the manual compilation of synonym dictionaries, which are specific to the protein family of interest. Experimentalists logically prefer to name proteins according to their function rather than by means established by a nomenclature committee. For example, it is easy to understand that scientists prefer to designate the protein they work on 'LHR-1' for 'Liver Receptor Homologue', 'alpha-1-fetoprotein' or 'CYP7A promoter binding factor' rather than 'NR5A2' as defined by standardized nomenclature (Nuclear Receptors Committee, 1999). It is difficult for someone not familiar with a given receptor to realize that all the names refer to the same protein. We therefore suggest that authors indicate the official protein name at least once in their publication.

CONCLUSION

We have developed a method that extracts point mutations from the literature and have applied the method to GPCRs and NRs. The efficiency of the method is consistent across these two well-studied families and relies on the accessibility of full texts and the use of established nomenclature for protein names and point mutations. MuteXt cannot be as accurate as a human curator. However, our algorithm can analyze 1000 articles in less than 10 h with specificity $>86\%$. Thus, this would seem to be a useful approach to the practical challenges of manual curation.

MuteXt can easily be applied to other protein families. This only requires the retrieval of the corresponding Swiss-Prot entries and the establishment of a dictionary of molecule names. The extraction of more complex mutations, such as deletions, insertions or chimera, is also possible. The challenge that remains is to extract information describing the effects of the mutations. This will probably require advanced NLP techniques in order to perform a deep analysis of the structure of the sentences surrounding the mutations.

ACKNOWLEDGEMENTS

We thank Barnaby C. H. May and Lawrence C. Lee for stimulating discussions, Erik J. Ellestad and Sarit Helman for technical assistance. We acknowledge the NIH for support.

REFERENCES

Andrade,M.A. and Valencia,A. (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 25–32.

Andrade,M.A. and Bork,P. (2000) Automated extraction of information in molecular biology. *FEBS Lett.*, **476**, 12–17.

Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 60–67.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Craven,M. and Kumlien,J. (1999) Constructing biological knowledge bases by extracting information from text sources. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 77–86.

den Dunnen,J.T. and Antonarakis,S.E. (2001) Nomenclature for the description of human sequence variations. *Hum. Genet.*, **109**, 121–124.

Edvardsen,O., Reiersen,A.L., Beukers,M.W. and Kristiansen,K. (2002) tGRAP, the G-protein coupled receptors mutant database. *Nucleic Acids Res.*, **30**, 361–363.

Fukuda,K., Tamura,A., Tsunoda,T. and Takagi,T. (1998) Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.*, 707–718.

Gaizauskas,R., Demetriou,G., Artymiuk,P.J. and Willett,P. (2003) Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, **19**, 135–143.

Hirschman,L., Park,J.C., Tsujii,J., Wong,L. and Wu,C.H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.

Horn,F., Vriend,G. and Cohen,F.E. (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.*, **29**, 346–349.

Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.

Humphreys,K., Demetriou,G. and Gaizauskas,R. (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.*, 505–516.

Leonard,J.E., Colombe,J.B. and Levy,J.L. (2002) Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, **18**, 1515–1522.

Marcotte,E.M., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.

Nuclear Receptors Committee (1999) A unified nomenclature system for the nuclear receptor superfamily. *Cell*, **97**, 161–163.

Ohta,Y., Yamamoto,Y., Okazaki,T., Uchiyama,I. and Takagi,T. (1997) Automatic construction of knowledge base from biological papers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 218–225.

Ono,T., Hishigaki,H., Tanigami,A. and Takagi,T. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Proux,D., Rechenmann,F. and Julliard,L. (2000) A pragmatic information extraction strategy for gathering data on genetic interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 279–285.

Proux,D., Rechenmann,F., Julliard,L., Pillet,V.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.

Rindflesch,T.C., Tanabe,L., Weinstein,J.N. and Hunter,L. (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.*, 517–528.

Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.

Tanabe,L. and Wilbur,W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*, **18**, 1124–1132.

Thomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.*, 541–552.

Van Durme,J.J., Bettler,E., Folkertsma,S., Horn,F. and Vriend,G. (2003) NRMD: nuclear receptor mutation database. *Nucleic Acids Res.*, **31**, 331–333.

Yoshida,M., Fukuda,K. and Takagi,T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169–175.