Computer Corner

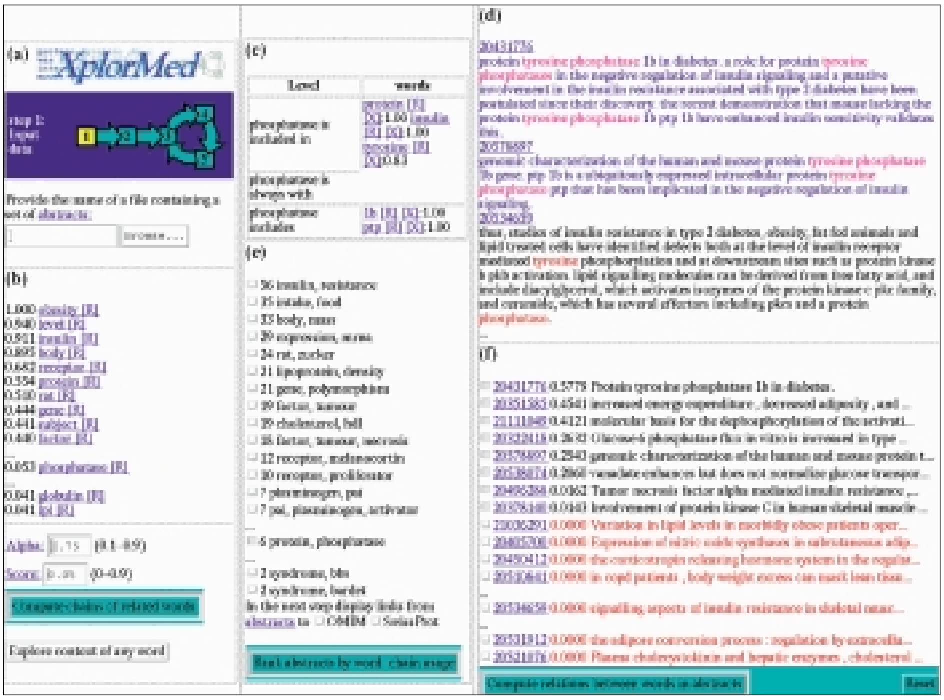
XplorMed: a tool for exploring MEDLINE abstracts

Carolina Perez-Iratxeta, Peer Bork and Miguel A. Andrade

The most frequent access to the MEDLINE database of scientific abstracts is by keyword search. However, this is often not sufficient because although the user might find all the useful abstracts, these are buried in hundreds that are irrelevant. The exploratory tool XplorMed has been developed to analyse the result of any MEDLINE query. It suggests main groups of related topics and documents, sparing the user the need of reading all abstracts.

A wealth of biological scientific information is available in the MEDLINE database (National Library of Medicine), which now holds more than ten million references to papers from biomedical journals dating back to 1966. To recognize research trends, integrate their own research with general knowledge and generate hypotheses to be tested in experiments, the individual researcher needs to be aware of only a small fraction of this information. At least three factors are making this increasingly difficult in the field of molecular biology: the introduction of high-throughput technologies that produce heterogeneous data, the softening of borders between research areas and a general increase of the number of scientific papers published.

As a result, exploratory literature searches in unfamiliar research areas are increasingly required and the researcher could have to analyse the whole corpus of references within MEDLINE. In this task, the classical problems of information retrieval (IR)[1] are met. The selection of relevant documents from a large collection resulting from a user's query has to be optimized to increase both the precision (fraction of retrieved documents that are of potential use for the user) and the recall (the fraction of the possibly useful documents that are retrieved).



(a)

Provide the name of a file containing a set of abstracts:

(b)

1.000 obesity [R]
0.940 level [R]
0.911 insulin [R]
0.895 body [R]
0.682 receptor [R]
0.554 protein [R]
0.510 rat [R]
0.444 gene [R]
0.441 subject [R]
0.440 factor [R]
...
0.053 phosphatase [R]
...
0.041 globulin [R]
0.041 lpl [R]

Alpha: 0.75 (0.1–0.9)
Score: 0.20 (0–0.9)

(c)

Level	words
phosphatase is included in	protein [R] [X] 1.00 insulin [R] [X] 1.00 tyrosine [R] [X] 0.83
phosphatase is always with	
phosphatase includes	1b [R] [X] 1.00 ptp [R] [X] 1.00

(e)

□ 56 insulin, resistance
□ 35 intake, food
□ 33 body, mass
□ 29 expression, mrna
□ 24 rat, zucker
□ 21 lipoprotein, density
□ 21 gene, polymorphism
□ 19 factor, tumour
□ 19 cholesterol, hdl
□ 18 factor, tumour, necrosis
□ 12 receptor, melanocortin
□ 10 receptor, proliferator
□ 7 plasminogen, pai
□ 7 pai, plasminogen, activator
...
□ 6 protein, phosphatase
...
□ 2 syndrome, bbs
□ 2 syndrome, bardet

In the next step display links from abstracts to □ OMIM □ SwissProt

(d)

20431776
protein tyrosine phosphatase 1b in diabetes. a role for protein tyrosine phosphatases in the negative regulation of insulin signaling and a putative involvement in the insulin resistance associated with type 2 diabetes have been postulated since their discovery. the recent demonstration that mouse lacking the protein tyrosine phosphatase 1b ptp 1b have enhanced insulin sensitivity validates this.
20578697
genomic characterization of the human and mouse protein tyrosine phosphatase 1b gene. ptp 1b is a ubiquitously expressed intracellular protein tyrosine phosphatase ptp that has been implicated in the negative regulation of insulin signaling.
20534659
thus, studies of insulin resistance in type 2 diabetes, obesity, fat fed animals and lipid treated cells have identified defects both at the level of insulin receptor mediated tyrosine phosphorylation and at downstream sites such as protein kinase b pkb activation. lipid signalling molecules can be derived from free fatty acid, and include diacylglycerol, which activates isozymes of the protein kinase c pkc family, and ceramide, which has several effectors including pkcs and a protein phosphatase.
...

(f)

□ 20431776 0.5779 Protein tyrosine phosphatase 1b in diabetes.
□ 20351383 0.4541 increased energy expenditure , decreased adiposity , and ...
□ 21131848 0.4121 molecular basis for the dephosphorylation of the activati...
□ 20322418 0.2632 Glucose-6 phosphatase flux in vitro is increased in type ...
□ 20578697 0.2540 genomic characterization of the human and mouse protein t...
□ 20538874 0.2060 vanadate enhances but does not normalize glucose transpor...
□ 20496288 0.0362 Tumor necrosis factor alpha mediated insulin resistance ,...
□ 20378140 0.0163 Involvement of protein kinase C in human skeletal muscle ...
□ 21006291 0.0000 Variation in lipid levels in morbidly obese patients oper...
□ 20465706 0.0000 Expression of nitric oxide synthases in subcutaneous adip...
□ 20490412 0.0000 the corticotropin releasing hormone system in the regulat...
□ 20520661 0.0000 in copd patients , body weight excess can mask lean tissu...
...
□ 20534659 0.0000 signalling aspects of insulin resistance in skeletal musc...
...
□ 20531912 0.0000 the adipose conversion process : regulation by extracellu...
□ 20521876 0.0000 Plasma cholecystokinin and hepatic enzymes , cholesterol ...

Fig. 1. Extracts from a typical run in the XplorMed server exploring proteins associated with obesity. After separating unwanted abstracts (e.g. those on nutrition) several candidate proteins are revealed and phosphatases are explored further. (a) Step 1: the results from a keyword query in PubMed with 'obesity AND protein' restricted, for simplicity, to the second half of the year 2000, were used as input to the system; that is, the resulting 463 abstracts are analysed now by XplorMed. The box graph depicts the steps of the analysis process. Steps 1 and 2 prepare the input, and Steps 3–5 can be iterated. In Step 2, one can select abstracts from the input based on the major MeSH categories (not applied in this example). (b) Step 3: words in the abstracts ranked by their degree of relatedness to other words. Names of proteins appear in the list, for example, phosphatase. Now one can narrow the focus and explore, for example, the association of phosphatases with obesity. (c) Exploring the relations of the word 'phosphatase' to other words. The table shows dependent terms (e.g. in the 463 abstracts 'ptp' only occurs with 'phosphatase') and terms on which 'phosphatase' depends (e.g. 'phosphatase' only occurs with 'tyrosine'). (d) Retrieval of sentences within the 463 abstracts containing either the word 'tyrosine' or 'phosphatase' with a colour code: red, target words (unless they appear consecutively – magenta); blue, sentences containing both words. MEDLINE identifiers (e.g. 20431776) are linked to the corresponding MEDLINE entry. Through these sentences it becomes obvious that the terms are related through 'protein tyrosine phosphatase'. Note that 'ptp' is the abbreviation, and '1b' the name of a gene encoding human and mouse 'ptp'. The third abstract (MEDLINE ID: 20534659) contains both terms but they are related in a more casual way. (e) Step 4: word chains of related terms ordered by the number of abstracts containing the corresponding words. 'Protein' and 'phosphatase' co-occur in six abstracts. (f) Step 5: selection of abstracts based on the 'protein, phosphatase' word chain. The selection is not hard because an abstract might be selected even if it does not contain one term of the word chain providing that it contains other terms strongly related to the missing one. For example, the seventh abstract (MEDLINE ID: 20496288) does not contain the word 'phosphatase' but was selected because it deals with 'tyrosine phosphorylation'. The top two abstracts from the previous set were selected but not the third one. From here one could iterate to Step 3. A deeper description of this example is given in a tutorial at http://www.bork.embl-heidelberg.de/xplormed/example/

Current limitations

The MEDLINE public servers already allow primary forms of document retrieval, the most popular of which is the selection of entries by the presence of a word or series of words (that can be combined by logical rules) in the titles, abstracts or MeSH terms (controlled set of keywords manually assigned by the compilers of MEDLINE, http://www.nlm.nih.gov/mesh/meshhome.html). Such 'hard' selection rules are not optimal for IR of documents written by humans (in natural language)[2]. Obviously, making a more unspecific keyword search can maximize recall. However, the number of entries retrieved might then be huge (low precision result) making the search useless. The user still has to filter the results in some way.

An additional problem of an exploratory search is that it is indeterminate. The user might not be able to state precisely the subjects of interest. This information is only known after examination of the literature. Furthermore, the user might find additional words that could be used to

0968-0004/01/$ – see front matter © 2001 Elsevier Science Ltd. All rights reserved. PII: S0968-0004(01)01926-0

Table 1. Improvement of precision of a MEDLINE query by XplorMed selection[a]

Review	MEDLINE query (1999–2000)	Precision in the MEDLINE query	Selection by MeSH term	Precision after MeSH selection	Selection by XplorMed word chain	Precision by XplorMed	Ref.
Recognition of antigens by single-domain antibody fragment	antibody AND (camel OR llama)	4/32 = 0.130	None	4/32 = 0.130	antibody, chain, vhh	4/10 = 0.400	9
Actin-based motility: stop and go with Ena/VASP proteins	vasp	25/68 = 0.370	Biological sciences	18/35 = 0.510	vasp, phosphoprotein	17/29 = 0.590	10
The histone fold is a key structural motif of transcription factor TFIID	transcription AND factor AND histone	11/322 = 0.034	Biological sciences	8/256 = 0.031	factor, transcription, complex	6/65 = 0.092	11
Transmembrane signaling in bacterial chemoreceptors	chemotaxis AND receptor	12/598 = 0.015	Biological sciences	8/390 = 0.021	receptor, cell, kinase	6/76 = 0.079	12

[a] To evaluate the capability of XplorMed to increase the precision of a query result in MEDLINE, four reviews[9–12] were analysed. Accordingly to each review, one MEDLINE query in PubMed (http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) was performed (limited to years 1999–2000 for simplicity) with carefully chosen keywords. The precision of such a query is defined by the fraction of papers that are also quoted in the review. For example, in the analysis of the review on chemoreceptors[12], the corresponding MEDLINE query produced 598 papers of which 12 were cited in the review (precision = 0.015). The precision of this MEDLINE query can sometimes be improved by application of MeSH term selection; in this example, the selection of papers related to biological sciences reduced the total set to 390, although only eight papers quoted in the review remained there (precision 0.021). A word chain suggested by XplorMed produced a much smaller set of 76 papers of which six were referenced in the review (precision = 0.079); thereby, a considerable improvement in precision was achieved.

expand the query in MEDLINE (e.g. unexpected abbreviations of a protein name or synonyms of a disease) or to explore unexpected findings laterally related to the original subject. For example, assume that a researcher requires an overview of the proteins associated with obesity. A query in MEDLINE with 'obesity AND protein' (resulting in >12 000 abstracts!) contains a heterogeneous collection of papers about the molecular biology of obesity hidden among papers about nutrition.

Analysis with XplorMed

To aid the kind of manual analysis that one would need in the cases described above, we have developed a system (XplorMed, http://www.bork.embl-heidelberg.de/xplormed/) that provides an intermediate level of analysis (Fig. 1). The system starts with the results of a query in MEDLINE (Fig. 1a). The relations between words (nouns) present in the same abstract are described using two fuzzy binary relations[3] that are more suitable for the modelling of natural language relationships than other approaches[4]. (A fuzzy binary relation is binary because it describes a relation between two elements 'a' and 'b', and it is fuzzy because it describes this relation with values between 0 and 1.) The degree of relatedness between the words 'a' and 'b' is estimated by the number of abstracts in which 'a' and 'b' co-occur divided by the total number of abstracts containing 'a' or 'b' (e.g. in papers talking about the baker's yeast, 'Saccharomyces' is related to 'cerevisiae'). The degree of inclusion of word 'a' into word 'b' is estimated by the number of abstracts in which 'a' co-occurs with 'b' divided by the number of abstracts containing 'a'. This relation expresses the fact that words related to general concepts include other words (e.g. 'magnetic' can include 'resonance', 'field' and 'force'). Important words can be identified by their high association score (the sum of the inclusion degree of all other words in that word, normalized to the maximum) (Fig. 1b). The user can explore the relation of a word to another (Fig. 1c) and the sentences containing them in the original abstracts (Fig. 1d). The system selects words with a high association score and the strongest relations between them are displayed as chains of related words (Fig. 1e). Finally, the user can choose one or more of the suggested word chains to select abstracts from the original query (Fig. 1f). The confined set of abstracts can be fed into the system in an iterative way for computation of new word relationships (e.g. on the abstracts related to 'protein tyrosine kinase' found in the query 'obesity AND protein') and further selection.

XplorMed overcomes some of the limitations of current approaches to abstract analysis. On the one hand, XplorMed detects keywords by examination of the query without the need of previously analysing any corpus (a large collection of abstracts; for example, the full MEDLINE) as opposed to detecting keywords using the difference between the frequency of usage of a word in a corpus and in the query text[5,6]. On the other hand, XplorMed relates abstracts by their usage of sets of keywords connected by dependency instead of finding related abstracts by the cosine score between vectors of word composition[7]. The cosine score is a commonly used measure of similarity between texts (e.g. implemented as clickable 'MEDLINE neighbours' in the PubMed server[8]). However, one major drawback of this measure is that it is very sensitive to biases in the word usage and other particularities inherent to human (natural) language.

We foresee that this approach can be successfully used for automation of data mining and data organization in molecular biology. XplorMed is, to the best of our knowledge, the first practical application towards this goal that can save time, for example by gaining precision of a query (Table 1) and by providing an interface to analyse the content of large samples of abstracts.

Acknowledgements

Thanks to our group members for fruitful discussions, to Alberto Pascual (CNB, Madrid) for useful insights, and to the developers and maintainers of the MEDLINE database. A publicly available standard grammatical tagger (TreeTagger, from Helmut Schmid of the Institut für Maschinelle Sprachverarbeitung, Stuttgart University, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html) was used for sentence boundary detection, word stemming, word lemmatization and grammatical parsing.

References

1 Lewis, D. and Jones, K. (1996) Natural language processing for information retrieval. *Communications of the ACM* 39, 92–101

2 Manning, C.D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press

3 Miyamoto, S. (1990) *Fuzzy Sets in Information Retrieval and Cluster Analysis. Theory and Decision Library*, Kluwer Academic Publishers

4 Zimmermann, J.H. (1985) *Fuzzy set Theory and its Applications*, Kluwer Academic Publishers

5 Andrade, M.A. and Valencia, A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14, 600–607

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

7 Salton, G. (1989) *Automatic Text Processing*. Addison–Wesley, Reading, MA, USA

8 Wilbur, W.J. and Yang, Y. (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.* 26, 209–222

9 Muyldermans, S. *et al.* (2001) Recognition of antigens by single-domain antibody fragments: the superfluous luxury of paired domains. *Trends Biochem. Sci.* 26, 230–235

10 Reinhard, M. *et al.* (2001) Actin-based motility: stop and go with Ena/VASP proteins. *Trends Biochem. Sci.* 26, 243–249

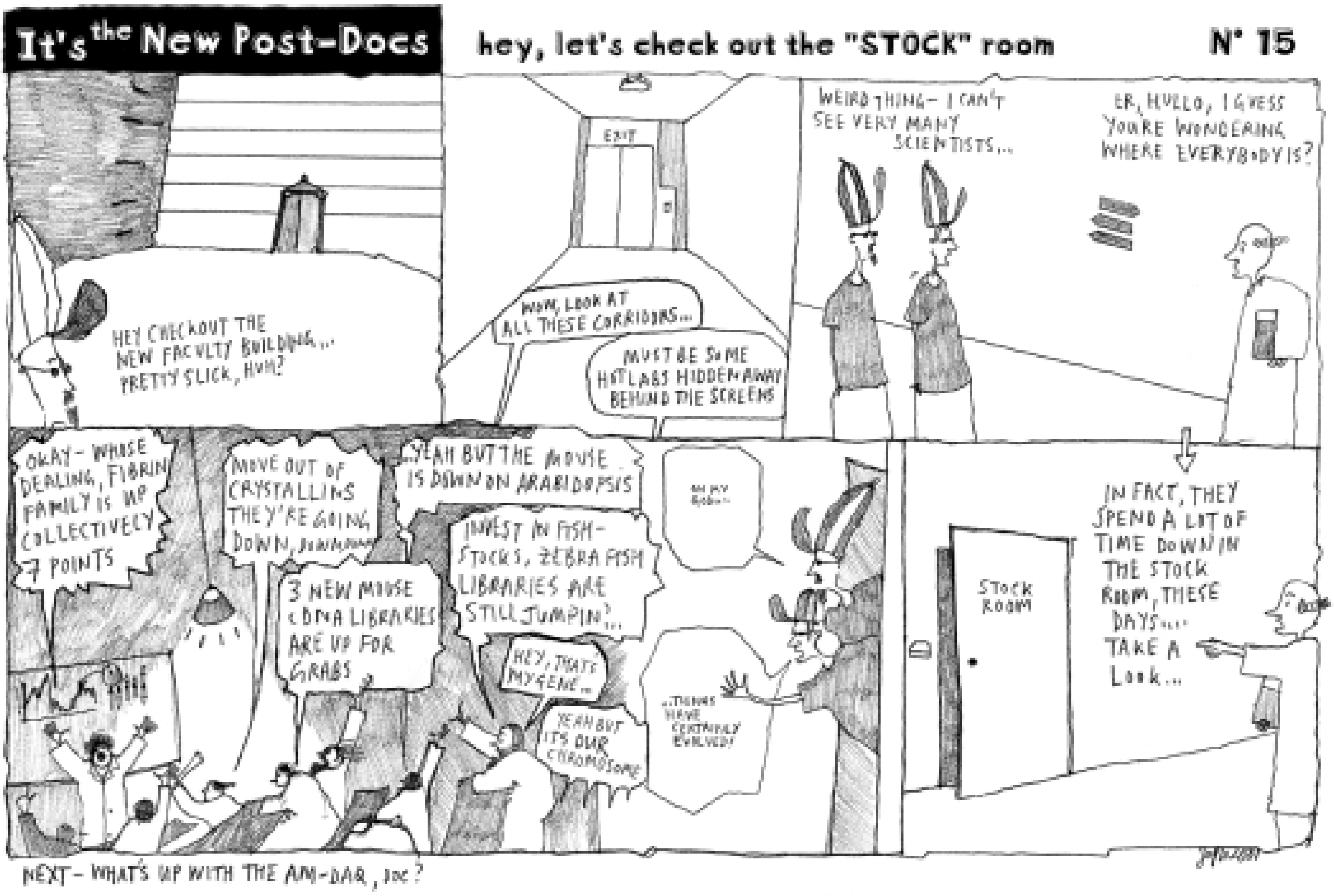
11 Gangloff, Y. *et al.* (2001) The histone fold is a key structural motif of transcription factor TFIID. *Trends Biochem. Sci.* 26, 250–257

12 Falke, J.J. and Hazelbauer, G.L. (2001) Transmembrane signaling in bacterial chemoreceptors. *Trends Biochem. Sci.* 26, 257–265

Carolina Perez-Iratxeta*
Peer Bork
Miguel A. Andrade
European Molecular Biology Laboratory,
Meyerhofstr. 1, Heidelberg 69012, Germany;
and Max Delbrück Center for Molecular
Medicine, Dept of Bioinformatics,
PO Box 740238, D-13092, Berlin-Buch,
Germany.
*e-mail: cperez@embl-heidelberg.de

Corrigendum

In the February issue, we published an article ‘The complexities of dystroglycan’ by Steven J. Winder (*TiBS* 26, 118–124). It is stated in the text and shown in Fig. 2 that biglycan associates with the N terminus of α-dystroglycan. This is, in fact, incorrect; biglycan associates specifically with the C terminus of α-dystroglycan. (Further details can be obtained from Bowe *et al.* (2000) *J. Cell Biol.* 148, 801–810.)


It's the New Post-Docs
hey, let's check out the "STOCK" room
N° 15
HEY CHECKOUT THE NEW FACULTY BUILDING... PRETTY SLICK, HUH?
EXIT
WOW, LOOK AT ALL THESE CORRIDORS...
MUST BE SOME HOT LABS HIDDEN AWAY BEHIND THE SCREENS
WEIRD THING – I CAN'T SEE VERY MANY SCIENTISTS...
ER, HULLO, I GUESS YOU'RE WONDERING WHERE EVERYBODY IS?
OKAY – WHOSE DEALING, FIBRIN FAMILY IS UP COLLECTIVELY 7 POINTS
MOVE OUT OF CRYSTALLINS THEY'RE GOING DOWN, DOWN...
...YEAH BUT THE MOUSE IS DOWN ON ARABIDOPSIS
3 NEW MOUSE cDNA LIBRARIES ARE UP FOR GRABS
INVEST IN FISH-STOCKS, ZEBRA FISH LIBRARIES ARE STILL JUMPIN'...
HEY, THAT'S MY GENE...
YEAH BUT IT'S OUR CHROMOSOME
...THINGS HAVE CERTAINLY EVOLVED!
IN FACT, THEY SPEND A LOT OF TIME DOWN IN THE STOCK ROOM, THESE DAYS... TAKE A LOOK...
STOCK ROOM
NEXT – WHAT'S UP WITH THE AM-DAR, DOC?

Pete Jeffs is a freelancer working in Paris, France.