Towards a terminological resource for biomedical text mining

Goran Nenadic[1], Naoki Okazaki[2], Sophia Ananiadou[1]

[1]School of Informatics, University of Manchester, UK
{G.Nenadic, Sophia.Ananiadou}@manchester.ac.uk
[2] Graduate School of Information Science and Technology, University of Tokyo, Japan
okazaki@mi.ci.i.u-tokyo.ac.jp

Abstract

One of the main challenges in biomedical text mining is the identification of terminology, which is a key factor for accessing and integrating the information stored in literature. Manual creation of biomedical terminologies cannot keep pace with the data that becomes available. Still, many of them have been used in attempts to recognise terms in literature, but their suitability for text mining has been questioned as substantial re-engineering is needed to tailor the resources for automatic processing. Several approaches have been suggested to automatically integrate and map between resources, but the problems of extensive variability of lexical representations and ambiguity have been revealed. In this paper we present a methodology to automatically maintain a biomedical terminological database, which contains automatically extracted terms, their mutual relationships, features and possible annotations that can be useful in text processing. In addition to TermDB, a database used for terminology management and storage, we present the following modules that are used to populate the database: TerMine (recognition, extraction and normalisation of terms from literature), AcroTerMine (extraction and clustering of acronyms and their long forms), AnnoTerm (annotation and classification of terms), and ClusTerm (extraction of term associations and clustering of terms).

1. Introduction

The amount of biomedical literature is growing constantly, and there are no signs that this trend will change its direction. The Medline database alone, for example, contains 14 million references. For biomedical research, the literature is still one of the main sources for knowledge access and acquisition. However, the huge number of available publications and diverse and dynamic terminology are the main bottlenecks for efficient mining to the literature.

The identification of terminology is a key factor for accessing and integrating the information stored in literature (Ananiadou & Nenadic, 2006; Krauthammer & Nenadic, 2004). Several biomedical curation teams (e.g. UniProt[1], SGD[2], FlyBase[3], etc.) maintain huge terminological resources for their purposes, which, however, are not based on systematic extraction and collection of terminology from literature. Such manual creation of biomedical terminologies cannot keep pace with the data that becomes available, and are typically restricted to the sub-domains of interest. The majority of such resources are designed for use by human specialists for knowledge integration. This is especially true for bio-ontologies and knowledge bases (e.g. GO[4], UMLS[5], UniProt) which are not primarily designed for automated processing. Still, many of them have been used in attempts to recognise terms in literature, but their suitability for automatic text mining has been questioned (Bodenreider et al., 2002; Liu & Friedman, 2003; McCray et al., 2001). Several approaches have been suggested to automatically integrate and map between resources, but the problems of extensive variability of lexical representations and ambiguity have been revealed.

A similar problem exists in the domain of biomedicine/bioinformatics with regard to gene annotation databases (e.g. GOA[6], FlyBase, etc.). The availability of such annotations is limited by the fact that they are maintained entirely manually, and hence cannot keep pace with the speed of new data that has been provided through various experiments. A solution to this problem has been suggested by providing computer-annotated supplements to such resources (by using clustering, classification and inference), which contain not always perfect but but still extremely useful annotations (e.g. UniProt and its automatic supplement TrEMBL[7]; similarly, the PRINTS and prePRINTS databases[8] (Attwood et al., 2003)).

In this paper we present a methodology to automatically produce and maintain a wide-coverage biomedical terminological database, which contains automatically extracted terms, their mutual relationships, features and possible annotations that can be useful for automatic text processing. We envisage this resource as a supplement to manually curated terminological databases.

The paper is organised as follows. In Section 2, we briefly present the main problems with existing biomedical terminological resources and discuss related work on automatic terminology management. Section 3 presents the overall architecture and design of our system, while some challenges and problems are discussed in Section 4.

2. Related work

Numerous existing biomedical resources have been used in attempts to recognise terms in the literature in order to support access to biological information and text mining. Sources commonly used for this purpose are the

[1] http://www.ebi.uniprot.org/
[2] http://www.yeastgenome.org/
[3] http://flybase.org/
[4] http://geneontology.org
[5] http://www.nlm.nih.gov/research/umls/
[6] http://www.ebi.ac.uk/GOA/
[7] http://www.ebi.ac.uk/trembl/
[8] http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/

Unified Medical Language System (UMLS), UniProt (Universal Protein Resource), Gene Ontology (GO), etc. These resources contain an abundance of terms (e.g. UMLS integrates around 5 million concept names from more than 100 controlled vocabularies (Bodenreider, 2006). However, many studies have suggested that information available in existing biomedical resources is not sufficient for text mining applications, and that substantial re-engineering is needed to tailor the resources for automatic processing (McCray et al., 2001; Liu et al., 2001), Bodenreider et al. 2002; McCray et al., 2002). In particular, the following major problems have been identified (see also (Ananiadou & Nenadic 2006)):

a) The resources are focused on facilitating knowledge integration by human specialists and not on automated processing.

b) It is difficult and time-consuming to update and curate the resources, since this is being done manually. Despite a huge number of terms, the resources are still limited in terminological coverage of the domain, either because a particular term-form and/or concept is missing, or available resources do not represent a specific type of entities (e.g. terms that refer to families or group of proteins (Blaschke and Valencia, 2002)).

c) Terminological variation is not systematically represented. Biological names are very complex and include an enormous amount of synonyms and different variant term-forms (acronyms, morphological and derivational variations, etc. (Nenadic et al., 2002))

d) Many biological terms and their variants are ambiguous. Terms share their lexical representation with both common English words (e.g. gene names such as *an*, *by*, *can* and *for*) and with other terms. However, terminological ambiguity is in the best case only recorded.

e) Although some cross-references between different resources exist, many problems still remain related to communication between and integration of different terminological databases.

f) Bio-lexical resources do not use language engineering standards (Heid & McNaught, 1991) and therefore are hardly reusable: new applications typically build their own resource model.

Several approaches have been suggested to address some of these problems. The results revealed the difficulties mainly in terms of extensive variability of lexical representations, and the problem of ambiguity with respect to mapping into a referent data source (Liu & Friedman, 2003). For example, attempts to integrate gene names (from UniProt) into UMLS were not successful since they increased ambiguity. The TERMINO system (Harkema et al., 2004) suggested a database schema for storing and integrating large-scale and diverse "biomedical word lists" collected from various resources (including UMLS, GOA, and in-house ontologies). It attempted to establish and maintain links between resources of various types (containing ontology concept names, terms, controlled vocabularies, nomenclatures, classification descriptors, etc.). Its specific focus is to provide storage efficiency and flexibility, and to provide a dictionary-based look-up mechanism. However, these integration approaches still do not solve the problems mentioned above. In particular, they do not improve the terminological coverage and variation and ambiguity resolution, and typically do not follow any language engineering standards.

3. System architecture

Our methodology aims at the integration of existing manually and automatically collected terminological resources. The system that provides automatic terminology results consists of five modules:

- TerMine, a module for recognition, extraction and normalisation of terms from the literature,
- AcroTerMine a module for the extraction and clustering of acronyms and their long forms,
- AnnoTerm, a module for annotation and classification of terms (e.g. gene/protein names),
- ClusTerm, a module for extraction of term associations and clustering of terms,
- TermDB, a terminology database management system.

Each of the modules is applied against a Medline[9] sub-corpus, containing around 7 million abstracts (total of 52GB including the meta-data provided by the National Library of Medicine). In the following text, we describe each of the modules in detail.

3.1. TerMine

TerMine[10] is based on the C-value method (Frantzi et al., 2000), a hybrid method combining statistics with linguistic knowledge to automatically recognise and extract multi-word terms. Term candidates are suggested by a set of morpho-syntactic filters, while their termhoods are estimated by a corpus-based statistical measure. The measure (called *C-value*) amalgamates four numerical characteristics of a candidate term, namely the frequency of occurrence, the frequency of occurrence as a substring of other candidate terms (in order to tackle nested terms), the number of candidate terms containing the given candidate term as a substring, and the number of words contained in the candidate term. This approach facilitates the recognition of embedded sub-terms, which are particularly frequent in biomedicine. TerMine also incorporates spelling and morphological variation and acronym recognition (see Section 3.2).

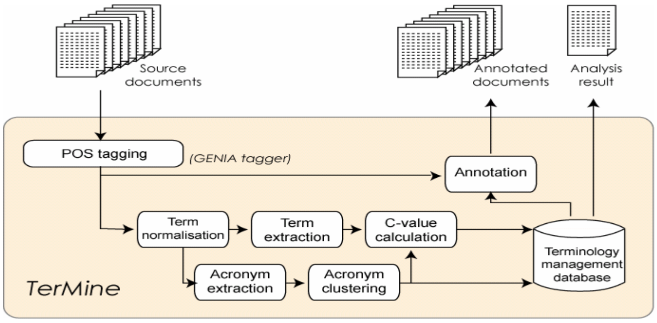


Figure 1: The term extraction process in TerMine

The TerMine system architecture is presented on Figure 1. Currently, the enumeration of all term candidates and the C calculation are run via a fast ANSI C

[9] http://www.ncbi.nlm.nih.gov/entrez
[10] http://www.nactem.ac.uk

implementation, with the processing time linear to the size of the input data. The input data is part-of-speech (POS) tagged by either the GENIA tagger[11] version 2.1 (Tsuruoka et al., 2005) or TreeTagger[12] version 3.1. The configurable linguistic filter is based on definite finite automaton (DFA).

The internal data (such as statistical information, nested/embedded terms) needed for the calculation of C-values is stored in a database. This solution provides both opportunities to parallelise the processing and to facilitate updates of C-values incrementally (see Section 4). The finial results are stored in the TermDB (see Section 3.6).

3.2. AcroTerMine

Acronyms are a very frequent term variation phenomenon, typically used as synonyms for the corresponding full form terms (Chang & Schuetze, 2006). For example, a recent study (Wren et al., 2005) reported that only 25% of documents relevant to the concept *c-jun N-terminal kinase* could be retrieved by using the full form, as in more than 33% of the documents the concept is referred to by using its acronym *JNK*. Thus, discovering acronyms and relating them to their long forms is an essential aspect of terminology management.

There are no formal rules or guidelines for the coinage of new biomedical acronyms, and, therefore, acronyms are subjected to variation and ambiguity. Some acronyms are synonymous: the same term may have several acronyms (*NF kappa B* and *NF kB* for *nuclear factor kappa B*). Other acronyms are polysemous (ambiguous): the same acronym may correspond to different terms (*GR* is used as an acronym for *glucocorticoid receptor* and for *glutathione reductase*).

AcroTerMine[13] implements recognition, extraction and clustering of acronyms (Naoaki & Ananiadou, 2006). The AcroTerMine results are integrated with TerMine, as acronyms typically indicate highly important terminological entities. Acronym recognition is based on a term extraction technique (the C-value method). We focus on terms appearing frequently on the proximity of an acronym. If a word sequence co-occurs frequently with a specific acronym and not with other surrounding words, we assume that there is a relationship between the acronym and the word sequence. This word sequence is likely to be a long form of the acronym. Figure 2 shows an overall architecture of the acronym recognition system.

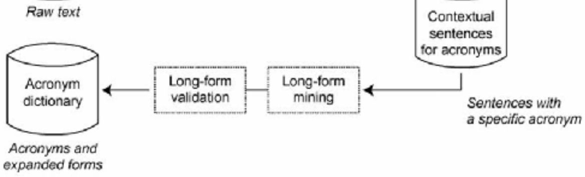

A large collection of text
Raw text
Short-form mining
All sentences with any acronyms
Contextual sentences for acronyms
Sentences with a specific acronym
Long-form mining
Long-form validation
Acronym dictionary
Acronyms and expanded forms

Figure 2: AcroTerMine architecture

3.3. AnnoTerm

The biomedical community has developed several widely accepted domain-specific ontologies that are used for knowledge sharing and integration. For example, the Gene ontology (GO) has been developed to *annotate* gene products with a shared set of concepts (Camon et al., 2004). It is used to describe genes in several organisms (humans, yeast, fly, rat, etc.), and huge databases of gene annotations have been produced and made available (including SGD, FlyBase, GOA). The Gene Ontology contains three branches, which aim to provide controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products. Each gene product is manually assigned one or more terms from each of the three branches. The relationships between a gene product and its molecular function, biological process and cellular component are all many to many, and are assigned independently.

Despite huge human efforts provided for the annotation task, many gene products are still not described (see Section 1). In addition to data mining techniques applied on experimental data, text mining has been also used to provide help for human annotators with the time-consuming curation of biological databases and suggest possible GO annotations for a given gene product (Camon et al., 2005; Blaschke et al., 2005).

The AnnoTerm module uses a supervised machine learning approach to assign GO terms to genes (Rice et al., 2005). We construe the gene annotation task as a modified form of a classification problem, and our approach is based on support vector machines (SVMs), which have been demonstrated to perform well at the document classification task. The approach is mainly based on the idea that biological entities (represented by domain terms) that co-occur in text with a gene of interest are indicative of its function, and that genes with similar co-occurrences of terms have related roles. Consequently, learning relevant and informative co-occurring terms for a given GO term should give clues for assignment of that GO term to genes that have similar distributional patterns.

Assignments of GO terms (both for learning and predicting) are based on collecting “weak” co-occurrence evidence within documents, rather than on explicit statement(s) of protein function. Actually, the GO terms appear explicitly in the biomedical documents very rarely. For example, in a collection of 53,000 abstracts (containing 8M words) only 8,000 occurrences corresponding to 739 different GO terms have been spotted, with only 392 terms appearing in two or more abstracts. Similar observations have been made by McCray et al. (2001) who reported that only 40% of GO terms were found in a corpus of 400,000 abstracts.

Therefore, an important facet of our approach is that GO assignments are not derived from a single, “relevant” passage or sentence, but from a set of document(s) relevant to a given protein. As features used for gene classification (i.e. association of GO terms), we use terms automatically extracted by TerMine. For each GO term, we have trained a SVM classifier, which is stored in an internal database for efficient access for classification of a new gene. The prediction of GO terms for a given gene is then performed in two steps. First we create a feature vector for the gene by collecting all terms from documents in which the gene appears. In the second step, we test the

[11] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[12] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[13] http://www.nactem.ac.uk

feature vector against respective GO classifiers, and select the GO term(s) associated with the top-ranked classifiers (i.e. the classifiers whose decision function values are the highest). These GO terms, along with the most frequent co-occurring terms, are then stored in the TermDB.

Note that target terms (e.g. genes) and the annotation scheme (e.g. GO) used in AnnoTerm are flexible, as the users can specify both resources (including the training sets) as well as the depth of the ontology they want to use in the annotation process.

3.4. ClusTerm

Biomedical entities are related in many ways: they have functional, structural, causal, hyponymous or other links. Relationships include diverse types of general (such as generalisation, specialisation, meronymy) and domain-specific relations (such as binding, phosphorilation, inhibition, etc.). For example, the term *NF-kappa B* is a hyponym of the term *transcription factor*, while the binding relationship links *amino acid* and *amino acid receptor*, as well as *CREP* and *CREP binding protein*; further examples of diverse relationships are co-location of proteins *ScNFU1* and *Nfs1p* in yeast mitochondria (Leon et al., 2003), or structural and functional similarities between proteins *red blood cell protein 4.1* and *synapsin I* (Krebs et al., 1987).

Mining such term relationships is performed by ClusTerm. Its aim is to extract such (pairs of) terms that are (potentially) semantically linked. We do not aim at identifying the type of the relationship(s) that exists among them, but rather at discovering links regardless of the type of the relationship (such terms are considered as semantically related). For example, a similar type of information is provided by some sources integrated in UMLS; they store limited information on *co-occurring concepts* (COC): “concepts that occur together in the same ‘entries’ in some information source”.

The method built in ClusTerm combines various text-based aspects of terms (Nenadic & Ananiadou, 2006), and provides three profiles that are attached to each term. *Lexical* profile gathers all lexically similar terms. Lexical similarities are based on the level of sharing of constituents. Syntactic similarities rely on expressions (such as term enumerations and conjunctions) in which a sequence of terms appears as a single syntactic unit. Thus, a *syntactic* profile provides a set of terms that appear together within a single syntactic unit, sharing the same verb or preposition. Finally, contextual similarities are based on automatic discovery of relevant contexts shared among terms. “Contexts” are described in a more generic way, and terms that tend do appear in similar contexts are linked (along with the contexts in question). These terms and associated contexts constitute a *contextual* profile of a given term.

Apart from storing term profiles, these similarities can be further used as features for clustering terms and visualisation of their relationships.

3.5. TermDB

TermDB is a relational database management system that stores terminological data mined from the literature. For each term-form extracted from the literature, the database stores the associated variants and acronyms, as well as a set of references to documents in which the term-form has been identified as one of the most important. For non-ambiguous terms that already exist in an external database (e.g. UMLS or UniProt), TermDB provides the link to the respective entries. If an acronym is ambiguous, the database stores all the associated long forms with the links to the related documents.

Currently, only UniProt/TrEMBL gene annotations are addressed by storing sets of assigned GO terms and co-occurring entities. For other term associations, all three term profiles are stored, along with associated textual contexts (in the case of syntactic and contextual profiles).

The database can be used to provide an interface for interactive access and integration of literature (see Appendix).

4. Discussion and challenges

The proposed model integrates automatic terminology management and existing resources. In particular, the model is focused on facilitating automated processing and time-consuming updates of the resources. Also, the database incorporates terminological variants and acronyms that have been systematically collected from the literature, with additional information (i.e. textual contexts) that may help in resolving terminological ambiguities and establishing additional terminological relationships. Finally, the model tries to integrate various resources by providing links to different terminological databases. This way, the existing resources are augmented with information mined from the literature.

In order to provide an efficient terminological resource, several challenges needed to be addressed:

a) **scalability**: having in mind the size of biomedical literature and the number of associated terms and named entities, our aim is to provide a highly scalable and efficient model, capable of processing huge amounts of documents (e.g. the whole of Medline). We have already compiled information on several million terminological entities (see Table 1) from several million documents. All terminological data (including various statistics, features, co-occurring terms etc.) is stored in an internal database for efficient access and future processing.

50165.3	amino acid	17056.8	tumor cell
38639.7	blood pressure	16995.8	coronary artery
38162.5	molecular weight	16746.1	cyclic amp
30539.9	blood flow	16622.2	nervous system
29822.5	fatty acid	16367.4	plasma membrane
28021.9	cell line	15969.6	peripheral blood
23856.0	electron microscopy	15442.9	side effect
23082.1	heart rate	15379.8	rat liver
22668.4	body weight	15206.0	myocardial infarct
21630.3	control group	15056.1	normal subject
20686.2	bone marrow	14752.7	red blood cell
19694.9	enzymatic activity	14423.1	protein synthesis
19155.5	guinea pig	14416.7	spinal cord
18261.6	lymph node	14264.7	monoclonal antibody
18255.5	gel electrophoresis	13502.3	blood cell

Table 1: An excerpt of top 30 terms (out of 9.8 million term candidates), extracted from a corpus of 1.3 million Medline abstracts (2GB)

b) **incremental processing**: since the Medline database alone, for example, grows by around 2000 documents each day and with many of the documents introducing new terms and variants, there is a need to systematically update the resources. By using the internal database for storing data needed for terminological processing, updating of the resource can be done incrementally when new documents emerge, without the need to re-process the previously analysed documents.

One of the further improvements of the methodology presented here is to integrate linguistically-based normalisation of term candidates, so that terms that share a canonical representation are linked in a *synterm* (a set of synonymous terms (Nenadic et al., 2004)).

The provided model can be embedded in various text mining solutions to support several services. Also, it can be implemented as a Web service to provide term annotation of documents (see Appendix). This way, the model would allow the users to retrieve, highlight and/or visualise terms, their annotations, associated entities, as well as annotate documents using the resource.

Apart from tagging terms in documents, the model can be used as a resource for terminological research and structuring, e.g. by mining various associations and features from the data already collected from documents. Thus, the model is not only a useful processing collection, but also a resource for terminological research.

5. Conclusion

One of the main challenges in biomedical text mining is the identification of terminology. Although numerous, the existing terminological resources typically lack the coverage and information that is needed for automated processing of the literature. In this paper, we have suggested a methodology that integrates automatic term and acronym recognition, variation conflation and association extraction with existing resources. We believe that the suggested model can be used as large-scale terminological support for processing and accessing huge amounts of documents, e.g. for providing text mining services[14] to the academic community. Also, it provides information that can be further used for sophisticated and complex investigation of biomedical terminology.

The system is designed to be scalable and to support incremental update of the resources. Basically, the solution is built around an internal database that stores necessary statistics, nested terms, term co-occurrences etc. and can be updated as new documents are processed.

The model also aims to address the following challenges: the focus on automated processing and updates, representation of terminological variants and information for resolving terminological ambiguities, and storage of terminological relationships. Finally, the model tries to integrate automatically extracted terminological data with existing resources by providing links to referential databases. This way, the existing resources are augmented with information mined from the literature.

For future work, we plan to provide a real-time version of the resource that will implement a Web service for tagging terms in documents and for term association mining. Also, we plan to represent the resource using language engineering standards to ensure its reusability.

6. Acknowledgments

This research was partially supported by the UK BBSRC grant “Mining Term Associations from Literature to Support Knowledge Discovery in Biology” (BB/C007360/1). The UK National Centre for Text Mining (NaCTeM) is funded by the Joint Information Systems Committee (JISC), the Biotechnology and Biological Sciences Research Council (BBSRC), and the Engineering and Physical Sciences Research Council (EPSRC).

7. References

Ananiadou, S., Nenadic, G. (2006). Automatic Terminology Management in Biomedicine. In S. Ananiadou, J. McNaught (Eds.), *Text Mining for Biology and Biomedicine*, Artech House Books, pp. 67-98.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003) "PRINTS and its automatic supplement, prePRINTS." *Nucleic Acids Research*, 31(1), 400-402.

Blaschke, C., Valencia, A. (2002). Molecular Biology Nomenclature thwarts Information-Extraction Progress”. *IEEE Intelligent Systems*, Vol.17, No.3, 2002, pp.73-76.

Blaschke, C, Andres, E, Krallinger, M, Valencia, A. (2005) Evaluation of the BioCreAtIvE assessment task 2. *BMC Bioinformatics*, 2005, 6(Suppl 1):S16

Bodenreider, O. (2006). Lexical, Terminological, and Ontological Resources for Biological Text Mining. In S. Ananiadou, J. McNaught (Eds.), *Text Mining for Biology and Biomedicine*, Artech House Books, pp. 43-66.

Bodenreider, O., Mitchell, J.A., McCray, A.T. (2002). Evaluation of the UMLS as a Terminology and Knowledge Resource for Biomedical Informatics. In *Proceedings of AMIA 2002*. pp. 61-65

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R., (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*. 2004 32(Database issue): D262–D266

Camon E., Barrell D., Dimmer E., Lee V., Magrane M., Maslen J., Binns D., Apweiler R. (2005). An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6(1): S17 (2005)

Chang, JT., Schuetze, H. (2006). Abbreviations in biomedical text. In S. Ananiadou and J. McNaught (Eds), *Text Mining for Biology and Biomedicine*, Artech House Books, pp. 99–119.

Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multi-word terms. *International Journal of Digital Libraries* 3(2) 117–132, 2000

Harkema, H., Gaizauskas, R., Hepple, M., Roberts, A., Roberts, I., Davis, N., Guo, Y. (2004). A large-scale terminology resource for biomedical text processing. In *Proceedings of BioLINK 2004*. pp. 53-60.

Heid, U., McNaught, J. (eds) (1991). EUROTRA-7 Study: Feasibility and project definition study of the reusability

[14] For example, this type of services is provided by the UK National Centre for Text Mining (NaCTeM).

of lexical and terminological resources in computerised applications. IMS, University of Stuttgart, Stuttgart.

Krauthammer, M., Nenadic, G. (2004). Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics* (Special Issue on Named Entity Recognition in Biomedicine), Vol.37(6), 2004, 512-526

Krebs, KE, Prouty, SM, Zagon, IS, Goodman SR (1987). Structural and functional relationship of red blood cell protein 4.1 to synapsin I. *Am J Physiol.* 1987 Oct;253(4 Pt 1):C500-5

Leon, S, Touraine, B, Ribot, C, Briat, JF, Lobreaux, S, (2003). Iron–sulphur cluster assembly in plants: distinct NFU proteins in mitochondria and plastids from Arabidopsis thaliana. *Biochem. J.* (2003) 371, 823–830

Liu, H., Lussier, Y., Friedman, C. (2001). A Study of Abbreviations in UMLS. In *Proceedings of AMIA 2001*. pp. 393-397.

Liu H, Friedman C. (2003). Mining terminological knowledge in large biomedical corpora. In *Proceedings of PSB 2003*. pp. 415-426.

McCray AT, Bodenreider O, Malley JD, Browne AC. (2001). Evaluating UMLS strings for natural language processing. In *Proceedings of AMIA 200*. pp. 448-452

McCray, A., Browne, A.C., Bodenreider, O. (2002). The Lexical Properties of the Gene Ontology. In *Proceedings of AMIA 2002*. 504-508.

Nenadic, G., Spasic, I., Ananiadou, S. (2002). Automatic Acronym Acquisition and Term Variation Management within Domain-Specific Texts. In *Proceedings of LREC 2002*. pp. 2155-2162.

Nenadic, G., Ananiadou, S., McNaught, J. (2004) Enhancing Automatic Term Recognition through Recognition of Variation. In *Proceedings of COLING 2004*. pp 604-610.

Nenadic, G., Ananiadou, S. (2006). Mining Semantically Related Terms from Biomedical Literature. *ACM Transactions on ALIP*, Vol. 5, No. 1, Special issue on Text Mining and Management in Biomedicine, pp 1-22.

Okazaki, N., Ananiadou, S. (2006). *Clustering acronyms in biomedical text for disambiguation*. In Proceedings of LREC 2006.

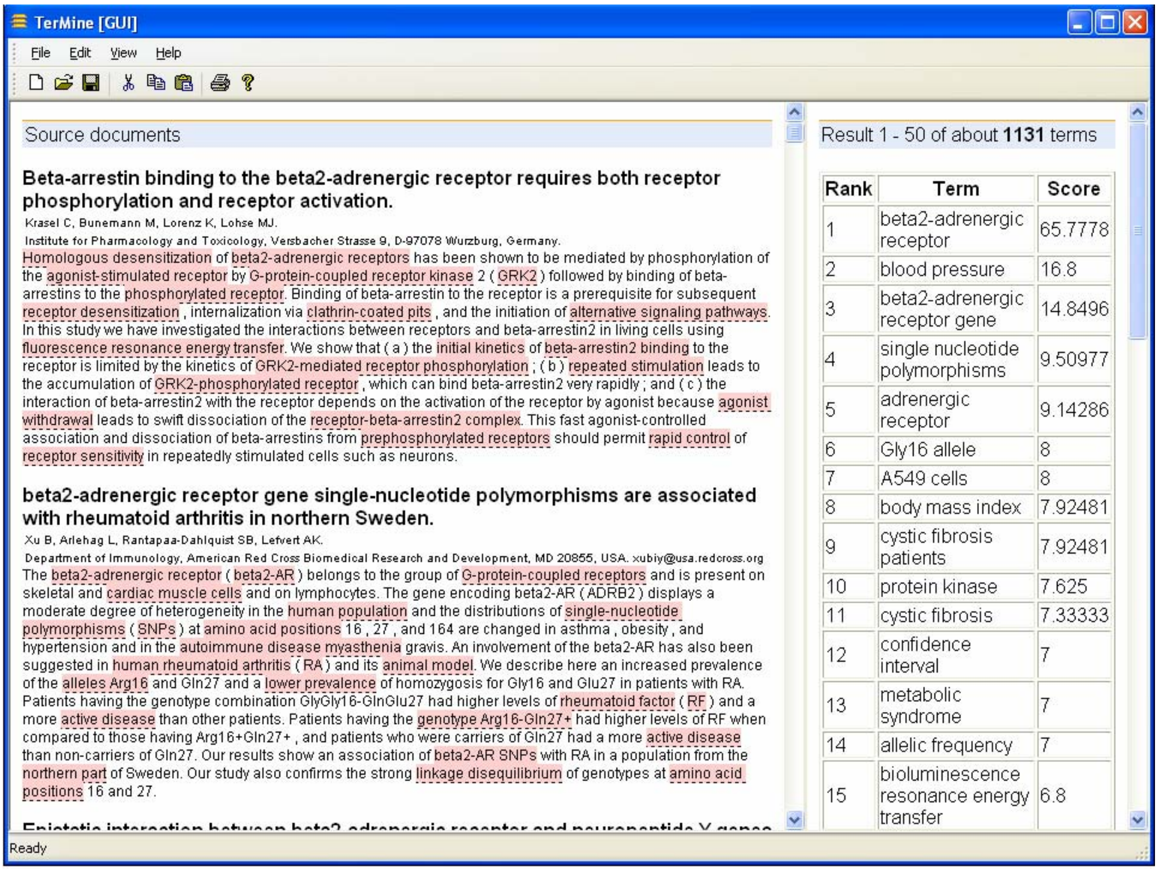
Rice, S., Nenadic, G., Stapley, B. (2005). Mining Protein Function from Text Using Term-based Support Vector Machines. *BMC Bioinformatics 2005*, 6(Suppl 1):S22

Tsuruoka, Y., Tateishi, Y., Kim, JD, Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*. LNCS 3746, pp. 382-392.

Wren, J.D., et al. (2005). Biomedical Term Mapping Databases. *Nucleic Acid Research*, 33, 2005, (Database Issue): D289--293.

Appendix

A screen-shot providing an interface for interactive access and integration of biomedical literature (see http://www.nactem.ac.uk).



Source documents

Beta-arrestin binding to the beta2-adrenergic receptor requires both receptor phosphorylation and receptor activation.

Krasel C, Bunemann M, Lorenz K, Lohse MJ.

Institute for Pharmacology and Toxicology, Versbacher Strasse 9, D-97078 Wurzburg, Germany.

Homologous desensitization of beta2-adrenergic receptors has been shown to be mediated by phosphorylation of the agonist-stimulated receptor by G-protein-coupled receptor kinase 2 (GRK2) followed by binding of beta-arrestins to the phosphorylated receptor. Binding of beta-arrestin to the receptor is a prerequisite for subsequent receptor desensitization , internalization via clathrin-coated pits , and the initiation of alternative signaling pathways. In this study we have investigated the interactions between receptors and beta-arrestin2 in living cells using fluorescence resonance energy transfer. We show that (a) the initial kinetics of beta-arrestin2 binding to the receptor is limited by the kinetics of GRK2-mediated receptor phosphorylation ; (b) repeated stimulation leads to the accumulation of GRK2-phosphorylated receptor , which can bind beta-arrestin2 very rapidly ; and (c) the interaction of beta-arrestin2 with the receptor depends on the activation of the receptor by agonist because agonist withdrawal leads to swift dissociation of the receptor-beta-arrestin2 complex. This fast agonist-controlled association and dissociation of beta-arrestins from prephosphorylated receptors should permit rapid control of receptor sensitivity in repeatedly stimulated cells such as neurons.

beta2-adrenergic receptor gene single-nucleotide polymorphisms are associated with rheumatoid arthritis in northern Sweden.

Xu B, Arlehag L, Rantapaa-Dahlquist SB, Lefvert AK.

Department of Immunology, American Red Cross Biomedical Research and Development, MD 20855, USA. xubiy@usa.redcross.org

The beta2-adrenergic receptor (beta2-AR) belongs to the group of G-protein-coupled receptors and is present on skeletal and cardiac muscle cells and on lymphocytes. The gene encoding beta2-AR (ADRB2) displays a moderate degree of heterogeneity in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 16 , 27 , and 164 are changed in asthma , obesity , and hypertension and in the autoimmune disease myasthenia gravis. An involvement of the beta2-AR has also been suggested in human rheumatoid arthritis (RA) and its animal model. We describe here an increased prevalence of the alleles Arg16 and Gln27 and a lower prevalence of homozygosis for Gly16 and Glu27 in patients with RA. Patients having the genotype combination GlyGly16-GlnGlu27 had higher levels of rheumatoid factor (RF) and a more active disease than other patients. Patients having the genotype Arg16-Gln27+ had higher levels of RF when compared to those having Arg16+Gln27+ , and patients who were carriers of Gln27 had a more active disease than non-carriers of Gln27. Our results show an association of beta2-AR SNPs with RA in a population from the northern part of Sweden. Our study also confirms the strong linkage disequilibrium of genotypes at amino acid positions 16 and 27.

Epistatic interaction between beta2-adrenergic receptor and neuropeptide Y genes

Result 1 - 50 of about **1131** terms

Rank	Term	Score
1	beta2-adrenergic receptor	65.7778
2	blood pressure	16.8
3	beta2-adrenergic receptor gene	14.8496
4	single nucleotide polymorphisms	9.50977
5	adrenergic receptor	9.14286
6	Gly16 allele	8
7	A549 cells	8
8	body mass index	7.92481
9	cystic fibrosis patients	7.92481
10	protein kinase	7.625
11	cystic fibrosis	7.33333
12	confidence interval	7
13	metabolic syndrome	7
14	allelic frequency	7
15	bioluminescence resonance energy transfer	6.8