

Deliverables Report
QLRI-2002-02770 BioMinT
<December 2004>



DELIVERABLE REPORT D10.5b (Year 2)
Yearly technology watch report

AUTHORS: Compiled on behalf of the consortium by Luc Dehaspe (PharmaDM),

STATUS: Final

CHECKERS:

Deliverables Report
QLRI-2002-02770 BioMinT
<December 2004>

PROJECT MANAGER

Name: Professor Terri Attwood
Address: Bioinformatics Group, School of Biological Sciences,
Stopford Building, University of Manchester, Oxford Road,, Manchester, M13 9PT
Phone Number: +44 161 275 5766
Fax Number: +44 161 275 5082
E-mail: attwood@bioinf.man.ac.uk

TABLE OF CONTENTS

1. EXECUTIVE OVERVIEW	3
2. NATURAL LANGUAGE PROCESSING	3
2.1. Named entity recognition	3
2.2. Part of speech tagging.....	5
2.3. Shallow parsing.....	5
3. INFORMATION EXTRACTION	6
4. APPLICATIONS AND (COMMERCIAL) TOOLS	8
5. RESOURCES	11
5.1. Ontologies and vocabularies	11
5.2. Annotated corpora	11
6. CONCLUSION.....	11

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

1. EXECUTIVE OVERVIEW

During the BioMinT project, we evidently want to stay informed about the state of the art and any new developments in text mining, information extraction, information retrieval and natural language processing. The yearly reports give a brief overview of the state of the art and of recent developments in the field.

In the 2003 report, we presented an overview of the state of the art, with somewhat more stress on 2003 work. In this report and in the 2005 one, we focus exclusively on new developments of the past year. This survey includes literature references to algorithms, techniques and novel complete systems, but also to useful resources that can be used to boost any text mining system. For each literature reference, we include a (clipped) abstract in italics.

2. NATURAL LANGUAGE PROCESSING

2.1. Named entity recognition

Jeffrey T. Chang, Hinrich Schütze, and Russ B. Altman

GAPSCORE: finding gene and protein names one word at a time

Bioinformatics 2004 20: 216-225

We have developed a new method, GAPSCORE, to identify gene and protein names in text. GAPSCORE scores words based on a statistical model of gene names that quantifies their appearance, morphology and context.

Results: F-score of 82.5% (83.3% recall, 81.5% precision) for partial matches and 57.6% (58.5% recall, 56.7% precision) for exact matches.

Availability: GAPSCORE is available at <http://bionlp.stanford.edu/gapscore/>

GuoDong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan

Recognizing names in biomedical texts: a machine learning approach

Bioinformatics 2004 20: 1178-1190.

Automatically recognizing biomedical entity names becomes critical and is important for information retrieval, information extraction and automated knowledge acquisition.

We present a named entity recognition system in the biomedical domain, called PowerBioNE. In order to deal with the special phenomena of naming conventions in the biomedical domain, we propose various evidential features: (1) word formation pattern; (2) morphological pattern, such as prefix and suffix; (3) part-of-speech; (4) head noun trigger; (5) special verb trigger and (6) name alias feature. All the features are integrated effectively and efficiently through a hidden Markov model (HMM) and a HMM-based named entity recognizer. In addition, a k-Nearest Neighbor (k-NN) algorithm is proposed to resolve the data sparseness problem in our system. Finally, we present a pattern-based post-processing to automatically extract rules from the training data to deal with the cascaded entity name phenomenon. From our best knowledge, PowerBioNE is the first system which deals with the cascaded entity name phenomenon. Evaluation shows that our system achieves the F-measure of 66.6 and 62.2 on the 23 classes of GENIA V3.0 and V1.1, respectively. In particular, our system achieves the F-measure of 75.8 on the 'protein' class of GENIA V3.0. For comparison, our system

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

outperforms the best published result by 7.8 on GENIA V1.1, without help of any dictionaries. It also shows that our HMM and the k-NN algorithm outperform other models, such as back-off HMM, linear interpolated HMM, support vector machines, C4.5, C4.5 rules and RIPPER, by effectively capturing the local context dependency and resolving the data sparseness problem. Moreover, evaluation on GENIA V3.0 shows that the post-processing for the cascaded entity name phenomenon improves the F-measure by 3.9. Finally, error analysis shows that about half of the errors are caused by the strict annotation scheme and the annotation inconsistency in the GENIA corpus. This suggests that our system achieves an acceptable F-measure of 83.6 on the 23 classes of GENIA V3.0 and in particular 86.2 on the 'protein' class, without help of any dictionaries. We think that a F-measure of 90 on the 23 classes of GENIA V3.0 and in particular 92 on the 'protein' class, can be achieved through refining of the annotation scheme in the GENIA corpus, such as flexible annotation scheme and annotation consistency, and inclusion of a reasonable biomedical dictionary.

Availability: A demo system is available at <http://textmining.i2r.a-star.edu.sg/NLS/demo.htm>.

Technology license is available upon the bilateral agreement.

Krauthammer M, Nenadic G.

Term identification in the biomedical literature.

J Biomed Inform. 2004 Dec;37(6):512-26.

This article overviews state-of-the-art approaches in term identification. The process of identifying terms is analysed through three steps: term recognition, term classification, and term mapping. For each step, main approaches and general trends, along with the major problems, are discussed. By assessing previous work in context of the overall term identification process, the review also tries to delineate needs for future work in the field.

Collier N, Takeuchi K.

Comparison of character-level and part of speech features for name recognition in biomedical texts.

J Biomed Inform. 2004 Dec; 37(6):423-35.

[...] there has been an intensive investigation into the named entity (NE) task as a core technology in all of these tasks which has been driven by the availability of high volume training sets such as the GENIA v3.02 corpus. Despite such large training sets accuracy for biology NE has proven to be consistently far below the high levels of performance in the news domain where F scores above 90 are commonly reported which can be considered near to human performance. We argue that it is crucial that more rigorous analysis of the factors that contribute to the model's performance be applied to discover where the underlying limitations are and what our future research direction should be. Our investigation in this paper reports on variations of two widely used feature types, part of speech (POS) tags and character-level orthographic features, and makes a comparison of how these variations influence performance. We base our experiments on a proven state-of-the-art model, support vector machines using a high quality subset of 100 annotated MEDLINE abstracts. Experiments reveal that the best performing features are orthographic features with F score of 72.6. Although the Brill tagger trained in-domain on the GENIA v3.02p POS corpus gives the best overall performance of any POS tagger, at an F score of 68.6, this is still significantly below the orthographic features. In combination these two features types appear to interfere with each other and degrade performance slightly to an F score of 72.3.

Sven Mika and Burkhard Rost

Protein names precisely peeled off free text

Bioinformatics 2004 20: i241-i247

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

[.] we introduced a novel system that combines a pre-processing dictionary- and rule-based filtering step with several separately trained support vector machines (SVMs) to identify protein names in the MEDLINE abstracts.

Results: Our new tagging-system NLProt is capable of extracting protein names with a precision (accuracy) of 75% at a recall (coverage) of 76% after training on a corpus, which was used before by other groups and contains 200 annotated abstracts. For our estimate of sustained performance, we considered partially identified names as false positives. One important issue frequently ignored in the literature is the redundancy in evaluation sets. We suggested some guidelines for removing overly inadequate overlaps between training and testing sets. Applying these new guidelines, our program appeared to significantly out-perform other methods tagging protein names. NLProt was so successful due to the SVM-building blocks that succeeded in utilizing the local context of protein names in the scientific literature. We challenge that our system may constitute the most general and precise method for tagging protein names.

Availability: <http://cubic.bioc.columbia.edu/services/nlprot/>

2.2. Part of speech tagging

L. Smith, T. Rindflesch, and W. J. Wilbur

MedPost: a part-of-speech tagger for bioMedical text

Bioinformatics 2004 20: 2320-2321

a part-of-speech tagger that achieves over 97% accuracy on MEDLINE citations.

Availability: Software, documentation and a corpus of 5700 manually tagged sentences are available at <ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz>

Ahmed Amrani, Yves Kodratoff and Oriane Matte-Tailliez

A Semi-automatic System for Tagging Specialized Corpora

Advances in Knowledge Discovery and Data Mining 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings

In this paper, we treat the problem of the grammatical tagging of non-annotated corpora of specialty. The existing taggers are trained on general language corpora, and give inconsistent results on the specialized texts, as technical and scientific ones. In order to learn rules adapted to a specialized field, the usual approach labels manually a large corpus of this field. This is extremely time-consuming. We propose here a semi-automatic approach for tagging corpora of specialty. ETIQ, the new tagger we are building, make it possible to correct the base of rules obtained by Brills tagger and to adapt it to a corpus of specialty. The user visualizes an initial and basic tagging and corrects it either by extending Brills lexicon or by the insertion of specialized lexical and contextual rules. The inserted rules are richer and more flexible than Brills ones. To help the expert in this task, we designed an inductive algorithm biased by the correct knowledge he acquired beforehand.

2.3. Shallow parsing

Morphological and Syntactic Processing for Text Retrieval - Vilares, Alonso (2004)

This article describes the application of lemmatization and shallow parsing as a linguistically-based alternative to stemming in Text Retrieval, with the aim of managing linguistic variation at both word level and phrase level. Several alternatives for selecting the index terms among the syntactic

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

dependencies detected by the parser are evaluated.

3. INFORMATION EXTRACTION

Florence Horn, Anthony L. Lau, and Fred E. Cohen

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

Bioinformatics 2004 20: 557-568

Motivation: 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. 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.

Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, and Ilya Mazo

Extracting human protein interactions from MEDLINE using a full-sentence parser

Bioinformatics 2004 20: 604-611

Automated literature processing tools could be employed to extract and organize biological data into a knowledge base, making it amenable to computational analysis and data mining.

Results: We present MedScan, a completely automated natural language processing-based information extraction system. We have used MedScan to extract 2976 interactions between human proteins from MEDLINE abstracts dated after 1988. The precision of the extracted information was found to be 91%. Comparison with the existing protein interaction databases BIND and DIP revealed that 96% of extracted information is novel. The recall rate of MedScan was found to be 21%.

Additional experiments with MedScan suggest that MEDLINE is a unique source of diverse protein function information, which can be extracted in a completely automated way with a reasonably high precision. Further directions of the MedScan technology improvement are discussed.

Availability: MedScan is available for commercial licensing from Ariadne Genomics, Inc

Daniel M. McDonald, Hsinchun Chen, Hua Su, and Byron B. Marshall

Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser

Bioinformatics 2004 20: 3370-3378

We present the Arizona Relation Parser that differs from other parsers in its use of a broad coverage syntax-semantic hybrid grammar. While syntax grammars have generally been tested over more documents, semantic grammars have outperformed them in precision and recall. We combined access to syntax and semantic information from a single grammar. The parser was trained using 40 PubMed abstracts and then tested using 100 unseen abstracts, half for precision and half for recall.

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

Expert evaluation showed that the parser extracted biologically relevant relations with 89% precision. Recall of expert identified relations with semantic filtering was 35 and 61% before semantic filtering. Such results approach the higher-performing semantic parsers. However, the AZ parser was tested over a greater variety of writing styles and semantic content.

Availability: Relations extracted from over 600 000 PubMed abstracts are available for retrieval and visualization at <http://econport.arizona.edu:8080/NetVis/index.html>

Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, and Ming Li

Discovering patterns to extract protein–protein interactions from full texts

Bioinformatics 2004 20: 3604-3612

We present a novel and robust approach for extracting protein–protein interactions from literature. Our method uses a dynamic programming algorithm to compute distinguishing patterns by aligning relevant sentences and key verbs that describe protein interactions. A matching algorithm is designed to extract the interactions between proteins. Equipped only with a dictionary of protein names, our system achieves a recall rate of 80.0% and precision rate of 80.5%.

Availability: The program is available on request from the authors.

Narayanasamy V, Mukhopadhyay S, Palakal M, Potter DA.

TransMiner: Mining Transitive Associations among Biological Objects from Text.

J Biomed Sci. 2004 Nov-Dec;11(6):864-73 .

Associations among biological objects such as genes, proteins, and drugs can be discovered automatically from the scientific literature. TransMiner is a system for finding associations among objects by mining the Medline database of the scientific literature. The direct associations among the objects are discovered based on the principle of co-occurrence in the form of an association graph. The principle of transitive closure is applied to the association graph to find potential transitive associations. The potential transitive associations that are indeed direct are discovered by iterative retrieval and mining of the Medline documents. Those associations that are not found explicitly in the entire Medline database are transitive associations and are the candidates for hypothesis generation. The transitive associations were ranked based on the sum of weight of terms that co-occur with both the objects. The direct and transitive associations are visualized using a graph visualization applet.

Liu Y, Brandon M, Navathe S, Dingledine R, Ciliax BJ.

Text mining functional keywords associated with genes.

Medinfo. 2004;2004:292-6.

We extended Andrade and Valencia's method [1] to statistically mine functional keywords associated with genes from MEDLINE abstracts. The MEDLINE abstracts are analyzed statistically to score and rank keywords for each gene using a background set of words for baseline frequencies. We generally got very good functional keyword information about the genes we tested, which was confirmed by searching for the individual keywords in context. The keywords extracted by our algorithm reveal a wealth of potential functional concepts, which were not represented in existing public databases. We feel that this approach is general enough to apply to medical and biological literature to find other relationships: drugs vs. genes, risk-factors vs. genes, etc.

Padmini Srinivasan and Bisharah Libbus

Mining MEDLINE for implicit links between dietary substances and diseases

Bioinformatics 2004 20: i290-i296

This work presents our text mining algorithm and demonstrates its use to uncover information that could form the basis of new hypotheses. In particular, we use it to discover novel uses for Curcuma longa, a dietary substance, which is highly regarded for its therapeutic properties in Asia.

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

Results: Several disease were identified that offer novel research contexts for curcumin. We analyze select suggestions, such as retinal diseases, Crohn's disease and disorders related to the spinal cord. Our analysis suggests that there is strong evidence in favor of a beneficial role for curcumin in these diseases. The evidence is based on curcumin's influence on several genes, such as COX-2, TNF-alpha, JNK, p38 MAPK and TGF-beta. This research suggests that our discovery algorithm may be used to suggest novel uses for dietary and pharmacological substances. More generally, our text mining algorithm may be used to uncover information that potentially sheds new light on a given topic of interest.

Corney DP, Buxton BF, Langdon WB, Jones DT.

BioRAT: extracting biological information from full-length papers.

Bioinformatics. 2004 Nov 22;20(17):3206-13

Recently, several information extraction systems have been developed that attempt to simplify the retrieval and analysis of biological and medical data. Most of this work has used the abstract alone, owing to the convenience of access and the quality of data. Abstracts are generally available through central collections with easy direct access (e.g. PubMed). The full-text papers contain more information, but are distributed across many locations (e.g. publishers' web sites, journal web sites and local repositories), making access more difficult. In this paper, we present BioRAT, a new information extraction (IE) tool, specifically designed to perform biomedical IE, and which is able to locate and analyse both abstracts and full-length papers. BioRAT is a Biological Research Assistant for Text mining, and incorporates a document search ability with domain-specific IE. RESULTS: We show first, that BioRAT performs as well as existing systems, when applied to abstracts; and second, that significantly more information is available to BioRAT through the full-length papers than via the abstracts alone. Typically, less than half of the available information is extracted from the abstract, with the majority coming from the body of each paper. Overall, BioRAT recalled 20.31% of the target facts from the abstracts with 55.07% precision, and achieved 43.6% recall with 51.25% precision on full-length papers.

4. APPLICATIONS AND (COMMERCIAL) TOOLS

Jung-Hsien Chiang, Hsu-Chun Yu, and Huai-Jen Hsu

GIS: a biomedical text-mining system for gene information discovery

Bioinformatics 2004 20: 120-121

A biomedical text-mining system focused on four types of gene-related information: biological functions, associated diseases, related genes and gene-gene relations. The aim of this system is to provide researchers an easy-to-use bio-information service that will rapidly survey the rapidly burgeoning biomedical literature.

Availability: <http://iir.csie.ncku.edu.tw/~yuhc/gis/>

H. Pan, L. Zuo, V. Choudhary, Z. Zhang, S. H. Leow, F. T. Chong, Y. Huang, V. W. S. Ong, B. Mohanty, S. L. Tan, S. P. T. Krishnan, and V. B. Bajic

Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining

Nucleic Acids Res., July 1, 2004; 32(suppl_2)

A system for text-mining of PubMed documents for potential functional association of transcription factors (TFs) with terms from Gene Ontology (GO) and with diseases. DTFAM has been trained and

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

tested in the selection of relevant documents on a manually curated dataset containing >3000 PubMed abstracts relevant to transcription control.

Availability: <http://research.i2r.a-star.edu.sg/DRAGON/TFAM/>.

R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, A. Coden, J. Cooper, A. Inokuchi, B. Iyer, Y. Mass, H. Matsuzawa, and L. V. Subramaniam

Text analytics for life science using the Unstructured Information Management Architecture

IBM Systems journal Volume 43, Number 3, 2004 Unstructured Information Management, pgs. 490-515

Several groups in the IBM Research Division are collaborating on the development of a prototype system for text analysis, search, and text-mining methods to support problem solving in life science. The system is called "BioTeKS" ("Biological Text Knowledge Services"), and it integrates research technologies from multiple IBM Research labs. BioTeKS is also the first major application of the UIMA (Unstructured Information Management Architecture) initiative also emerging from IBM Research. BioTeKS is intended to analyze biomedical text such as MEDLINE™ abstracts, medical records, and patents; text is analyzed by automatically identifying terms or names corresponding to key biomedical entities (e.g., "genes," "proteins," "compounds," or "drugs") and concepts or facts related to them.

N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda.

A text-mining system for knowledge discovery from biomedical documents

IBM Systems journal Volume 43, Number 3, 2004 Unstructured Information Management, pgs. 516-533.

This paper describes the application of IBM TAKMI® for Biomedical Documents to facilitate knowledge discovery from the very large text databases characteristic of life science and healthcare applications. This set of tools, designated MedTAKMI, is an extension of the TAKMI (Text Analysis and Knowledge Mining) system originally developed for text mining in customer-relationship-management applications. MedTAKMI dynamically and interactively mines a collection of documents to obtain characteristic features within them. By using multifaceted mining of these documents together with biomedically motivated categories for term extraction and a series of drill-down queries, users can obtain knowledge about a specific topic after seeing only a few key documents. In addition, the use of natural language techniques makes it possible to extract deeper relationships among biomedical concepts. The MedTAKMI system is capable of mining the entire MEDLINE® database of 11 million biomedical journal abstracts. It is currently running at a customer site.

Chen H, Sharp BM.

Content-rich biological network constructed by mining PubMed abstracts.

BMC Bioinformatics. 2004 Oct 08;5(1):147.

We present a NLP-based text-mining approach, Chilibot, which constructs content-rich relationship networks among biological concepts, genes, proteins, or drugs. Amongst its features, suggestions for new hypotheses can be generated. Lastly, we provide evidence that the connectivity of molecular networks extracted from the biological literature follows the power-law distribution, indicating scale-free topologies consistent with the results of previous experimental analyses.. Chilibot <http://www.chilibot.net> can be accessed free of charge to academic users.

Muller HM, Kenny EE, Sternberg PW.

Textpresso: an ontology-based information retrieval and extraction system for biological literature.

PLoS Biol. 2004 Nov;2(11):e309.

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

We have developed *Textpresso*, a new text-mining system for scientific literature whose capabilities go far beyond those of a simple keyword search engine. *Textpresso*'s two major elements are a collection of the full text of scientific articles split into individual sentences, and the implementation of categories of terms for which a database of articles and individual sentences can be searched. The categories are classes of biological concepts (e.g., gene, allele, cell or cell group, phenotype, etc.) and classes that relate two objects (e.g., association, regulation, etc.) or describe one (e.g., biological process, etc.). Together they form a catalog of types of objects and concepts called an ontology. After this ontology is populated with terms, the whole corpus of articles and abstracts is marked up to identify terms of these categories. The current ontology comprises 33 categories of terms. A search engine enables the user to search for one or a combination of these tags and/or keywords within a sentence or document, and as the ontology allows word meaning to be queried, it is possible to formulate semantic queries. Full text access increases recall of biological data types from 45% to 95%. Extraction of particular biological facts, such as gene-gene interactions, can be accelerated significantly by ontologies, with *Textpresso* automatically performing nearly as well as expert curators to identify sentences; in searches for two uniquely named genes and an interaction term, the ontology confers a 3-fold increase of search efficiency. *Textpresso* currently focuses on *Caenorhabditis elegans* literature, with 3,800 full text articles and 16,000 abstracts. The lexicon of the ontology contains 14,500 entries, each of which includes all versions of a specific word or phrase, and it includes all categories of the Gene Ontology database. *Textpresso* is a useful curation tool, as well as search engine for researchers, and can readily be extended to other organism-specific corpora of text. *Textpresso* can be accessed at <http://www.textpresso.org> or via WormBase at <http://www.wormbase.org>.

Karopka T, Scheel T, Bansemer S, Glass A.

Automatic construction of gene relation networks using text mining and gene expression data.

Med Inform Internet Med. 2004 Jun;29(2):169-83.

Microarray gene expression analysis is a powerful high-throughput technique that enables researchers to monitor the expression of thousands of genes simultaneously. Using this methodology huge amounts of data are produced which have to be analysed. Clustering algorithms are used to group genes together based on a predefined distance measure. However, clustering algorithms do not necessarily group the genes in a biological meaningful way. Additional information is needed to improve the identification of disease relevant genes. The primary objective of our project is to support the analysis of microarray gene expression data by construction of gene relation networks (GRNs). Required information can not be found in a structured representation like a database. In contrast, a large number of relations are described in biomedical literature. The main outcome of this project is the implementation of a software system that provides clinicians and researchers with a tool that supports the analysis of microarray gene expression data by mapping known relationships from the biomedical literature to local gene expression experiments.

Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B.

TXTGate: profiling gene groups with text-based information.

Genome Biol. 2004;5(6):R43. Epub 2004 May 28.

We implemented a framework called TXTGate that combines literature indices of selected public biological resources in a flexible text-mining system designed towards the analysis of groups of genes. By means of tailored vocabularies, term- as well as gene-centric views are offered on selected textual fields and MEDLINE abstracts used in LocusLink and the Saccharomyces Genome Database. Subclustering and links to external resources allow for in-depth analysis of the resulting term profiles.

Deliverables Report

QLRI-2002-02770 BioMinT

<December 2004>

Eskin E, Agichtein E.

Combining text mining and sequence analysis to discover protein functional regions.

Pac Symp Biocomput. 2004;:288-99.

Recently presented protein sequence classification models can identify relevant regions of the sequence. This observation has many potential applications to detecting functional regions of proteins. However, identifying such sequence regions automatically is difficult in practice, as relatively few types of information have enough annotated sequences to perform this analysis. Our approach addresses this data scarcity problem by combining text and sequence analysis. First, we train a text classifier over the explicit textual annotations available for some of the sequences in the dataset, and use the trained classifier to predict the class for the rest of the unlabeled sequences. We then train a joint sequence text classifier over the text contained in the functional annotations of the sequences, and the actual sequences in this larger, automatically extended dataset. Finally, we project the classifier onto the original sequences to determine the relevant regions of the sequences. We demonstrate the effectiveness of our approach by predicting protein sub-cellular localization and determining localization specific functional regions of these proteins.

5. RESOURCES

5.1. Ontologies and vocabularies

E. Demir, O. Babur, U. Dogrusoz, A. Gursoy, A. Ayaz, G. Gulesir, G. Nisanci, and R. Cetin-Atalay

An ontology for collaborative construction and analysis of cellular pathways

Bioinformatics 2004 20: 349-356.

We define an ontology for an intuitive, comprehensive and uncomplicated representation of cellular events. The ontology presented here enables integration of fragmented or incomplete pathway information via collaboration, and supports manipulation of the stored data. In addition, it facilitates concurrent modifications to the data while maintaining its validity and consistency. Furthermore, novel structures for representation of multiple levels of abstraction for pathways and homologies is provided. Lastly, our ontology supports efficient querying of large amounts of data.

We have also developed a software tool named pathway analysis tool for integration and knowledge acquisition (PATIKA) providing an integrated, multi-user environment for visualizing and manipulating network of cellular events. PATIKA implements the basics of our ontology.

Availability: PATIKA version 1.0 beta is available upon request at <http://www.patika.org>

5.2. Annotated corpora

Hu ZZ, Mani I, Hermoso V, Liu H, Wu CH.

iProLINK: an integrated protein resource for literature mining.

Comput Biol Chem. 2004 Dec;28(5-6):409-16.

Motivated by the promise of text mining methodologies, but at the same time, the lack of adequate curated data for training and benchmarking, the Protein Information Resource (PIR) has developed a

Deliverables Report
QLRI-2002-02770 BioMinT
<December 2004>

resource for protein literature mining-iProLINK (integrated Protein Literature INformation and Knowledge). As PIR focuses its effort on the curation of the UniProt protein sequence database, the goal of iProLINK is to provide curated data sources that can be utilized for text mining research in the areas of bibliography mapping, annotation extraction, protein named entity recognition, and protein ontology development. The data sources for bibliography mapping and annotation extraction include mapped citations (PubMed ID to protein entry and feature line mapping) and annotation-tagged literature corpora. The latter includes several hundred abstracts and full-text articles tagged with experimentally validated post-translational modifications (PTMs) annotated in the PIR protein sequence database. The data sources for entity recognition and ontology development include a protein name dictionary, word token dictionaries, protein name-tagged literature corpora along with tagging guidelines, as well as a protein ontology based on PIRSF protein family names. iProLINK is freely accessible at , with hypertext links for all downloadable files.

6. CONCLUSION

This description of developments in the field of biological text mining should be a good overview, but can of course not be comprehensive. However, the following pointers should lead to lots of additional information.

- <http://www.cis.upenn.edu/~mamandel/term.html>: contains an enormous overview of resources for biomedical terminology and ontologies;
- <http://www.bionlp.org/>: A global portal site for any news on natural language processing of biology text
- <http://www.ccs.neu.edu/home/futrelle/bionlp/papers/ShatkayFeldmanJCB03.pdf> : February 2nd, 2004: A lengthy review, "Mining the Biomedical Literature in the Genomic Era: An Overview" by H. Shatkay and R. Feldman appeared in the December 2003 issue of the Journal of Computational Biology, 10 (3) 821-855.
- <http://www.tufts.edu/~amorqa02/bcresources.html>: freely available BioNLP resources
- <http://www.ebi.ac.uk/Rebholz/resources.html>: links to text mining tools, companies and resources compiled by the Rebholz group.

This technology watch will be continued and reported on in the third and final year of the project.