BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 17 2008, pages 1935–1941
doi:10.1093/bioinformatics/btn318

Data and text mining

PuReD-MCL: a graph-based PubMed document clustering methodology

T. Theodosiou[1,2,*], N. Darzentas[2], L. Angelis[1] and C. A. Ouzounis[2,3]

[1]Department of Informatics, Aristotle University of Thessalonica, P.O. Box 54124, Thessalonica, Greece, [2]Computational Genomics Unit, Institute of Agrobiotechnology, Centre for Research and Technology Hellas (CERTH), P.O. Box 361, GR–57001, Thessalonica, Greece and [3]Centre for Bioinformatics, School of Physical Sciences & Engineering, King's College London, Strand, London WC2R 2LS, UK

Received on November 12, 2007; revised on May 18, 2008; accepted on June 18, 2008

Advance Access publication July 1, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Biomedical literature is the principal repository of biomedical knowledge, with PubMed being the most complete database collecting, organizing and analyzing such textual knowledge. There are numerous efforts that attempt to exploit this information by using text mining and machine learning techniques. We developed a novel approach, called PuReD-MCL (Pubmed Related Documents-MCL), which is based on the graph clustering algorithm MCL and relevant resources from PubMed.

Methods: PuReD-MCL avoids using natural language processing (NLP) techniques directly; instead, it takes advantage of existing resources, available from PubMed. PuReD-MCL then clusters documents efficiently using the MCL graph clustering algorithm, which is based on graph flow simulation. This process allows users to analyse the results by highlighting important clues, and finally to visualize the clusters and all relevant information using an interactive graph layout algorithm, for instance BioLayout Express 3D.

Results: The methodology was applied to two different datasets, previously used for the validation of the document clustering tool TextQuest. The first dataset involves the organisms *Escherichia coli* and yeast, whereas the second is related to *Drosophila* development. PuReD-MCL successfully reproduces the annotated results obtained from TextQuest, while at the same time provides additional insights into the clusters and the corresponding documents.

Availability: Source code in perl and R are available from http://tartara.csd.auth.gr/∼theodos/

Contact : theodos@csd.auth.gr

1 INTRODUCTION

There is an overwhelming amount of textual knowledge recorded in the biomedical literature, with the number of articles published each year increasing exponentially, following the advances in high-throughput experimental and computational methods. The PubMed database, which is considered one of the most complete repositories of biomedical articles, contains more than 11 million abstracts and receives more than 70 million queries each month.

*To whom correspondence should be addressed.

The vast amount of documents available in the biomedical literature makes the manual handling, analysis and interpretation of textual information a daunting task. Automated methods that assist users to sift through this unstructured heap of valuable archives are becoming increasingly important in scientific research. There have been numerous attempts to develop systems that analyse textual resources and contribute towards the discovery of key facts with minimum user interaction and guidance. These approaches need to be scalable, as automatic as possible, as well as user-friendly (Ananiadou *et al*., 2006). Most text mining systems use natural language processing (NLP), machine learning and data mining techniques in order to process text, usually in the form of document abstracts. Examples include maximum entropy (Raychaudhuri *et al*., 2002), support vector machines (Izumitani *et al*., 2004) and linear discriminant analysis (LDA) (Theodosiou *et al*., 2007), all representing different classification methods based on a training set of documents. These training sets assist the creation of models, which are subsequently used for the categorization of new documents from so-called test sets. Avoiding the training/test set paradigm is advantageous, especially in cases where the partitioning of the document space is not known a priori. For instance, the TextQuest document clustering system uses lists of keywords (Iliopoulos *et al*., 2001) and thus is independent of a training set or a specific algorithm. More complicated text mining techniques involve information extraction from documents. These make extensive use of syntactic and semantic analysis in order to recognize the biological entities described in the text, like in Hu *et al*. (2005) and Nenadic *et al*. (2003). The drawback is that they are computationally demanding, and they typically require a predefined ontology for the domain of discourse (Ananiadou *et al*., 2006; Iliopoulos *et al*., 2001). In summary, the main feature of biomedical text mining methods is the use of NLP techniques, in order to process textual information. Usually, documents are represented using the Vector Space Model (VSM) (Manning and Schutze, 1999; Salton , 1970), where each document is encoded as a vector of weighted words. In information extraction, the syntax and the semantics of text are also taken into account for each document, using more complex structures than VSM (Ananiadou *et al*., 2006).

The motivation behind the current work was to create an approach that is able to cluster arbitrarily large quantities of biomedical text from the PubMed database and exploit useful information by

extracting, filtering and organizing it, without directly relying on sophisticated NLP techniques. Another important aspect was the visualization of the extracted information alongside the relationships between documents thus allowing users to interact with results, leading to better understanding and enhanced knowledge.

The methodology we put forward has a number of desirable advantages:

(1) It relies on robust and precomputed information provided by PubMed curators and computational infrastructure, thus avoiding cumbersome calculations.

(2) It uses an efficient and scalable graph clustering algorithm, previously applied to very large and diverse datasets other than text.

(3) It uses a statistical methodology, based on χ^2-testing and bootstrap procedures, for assessing the quality of the document clusters.

(4) It only uses the document title and medical subject headings (MeSH) terms, selected by a simple yet effective scoring and filtering scheme, to describe the knowledge in documents and clusters.

(5) It embeds the results and relationships interactively, in 2D and 3D space.

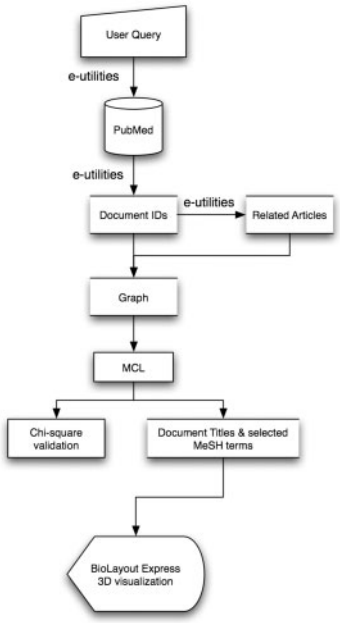

Fig. 1. The different steps of the proposed methodology.

2 METHODS

The general idea of our methodology is to represent the relationships between the documents with a graph, and cluster them using the Markov clustering algorithm (MCL) (van Dongen, 2000). In order to avoid the direct use of NLP techniques, precomputed relationships from PubMed are utilized to build the association graph between documents. The clusters are visualized in 2D or 3D space and are described using the document titles and the MeSH terms.

Specifically, the methodology involves the following main steps (Fig. 1):

(1) The retrieval of a set of documents (abstracts) through the query system of PubMed.

(2) The retrieval of precomputed related documents for each document of the previous set, again from PubMed.

(3) The creation of a graph between the documents of the first step, with vertices representing documents and edges representing document relatedness.

(4) The clustering of the graph using the MCL.

(5) The annotation of the clusters and their documents using an efficient in-house algorithm.

(6) The statistical validation of the clustering results.

(7) The visualization of the results using BioLayout Express 3D.

Steps (1) and (2):

A user query is performed at PubMed, and for each of the returned documents its precomputed related documents are retrieved using the e-utilities module of Entrez (Wheeler *et al*., 2007).

The algorithm for calculating the related documents is based on comparing words from the title, the abstract and the MeSH terms of the documents. It uses NLP techniques and the VSM in order to represent each document as a vector of weighted words and apply vector scoring (Wilbur and Yang, 1996).

Step (3):

A graph between the documents returned by the query is created. Each vertex/node of the graph represents a document. The edges connect relevant documents together based on the 'related articles' information from PubMed. Each edge has a weight, which is the vector score returned by the 'related articles' tool, re-scaled between 0 and 100.

In this step, there is a possibility that some documents do not relate to any of the other. These documents are represented in the graph as stand-alone, disconnected nodes and do not affect the next step of the clustering, since they have no connections to the rest of the documents.

Step (4):

The MCL algorithm is used to uncover clusters in the graph. MCL performs unsupervised clustering (i.e. there is no need to predefine the number of clusters) on weighted graphs. It has previously been used successfully in biology for detecting protein families (Enright *et al*., 2002), and has also been shown to perform better than other algorithms when used in protein–protein interaction networks (Brohee and van Helden, 2006).

MCL is based on the idea that natural clusters in a graph have many edges between the members of each cluster and few across clusters. Once inside a cluster, a hypothetical entity moving around randomly will have little chance to escape that cluster. MCL simulates random walks (flow) within the entire graph and augments flow where it is already strong and weakens it where it is weak. After many iterations of this process, the underlying cluster structure of the graph gradually becomes visible. Regions of the graph with high flow describe clusters that are separated by boundaries with no flow. MCL simulates the random walks within a graph by two algebraic operations, called expansion and inflation that are applied to a stochastic matrix. The matrix representing the graph is used as input, while expansion and inflation are applied for many rounds, until there is little or no change in the matrix. The final matrix then represents the clustering of the graph nodes (documents in this context). Expansion refers to the power of a stochastic matrix, using the normal matrix product. Inflation is the entry-wise Hadamard—Schur product (Radhakrishna and Bhaskara, 1998) combined with diagonal scaling, and is responsible for both the strengthening and the weakening of the flow. The value of the inflation parameter controls cluster granularity (van Dongen, 2000).

The MCL algorithm is considered to be very fast and scalable (van Dongen, 2000), since its worst case time complexity is $O(N*L\hat{}2)$, where N is the number of documents and L is an MCL parameter usually between 500 and 1000. The space complexity is $O(N*L)$ (van Dongen, 2000).

Step (5):

Each cluster and included documents are annotated using the document titles and a selected number of MeSH terms and chemical substances. The

MeSH terms form a controlled structured vocabulary, used for indexing PubMed documents. The chemical substances are supplementary concept records of the MeSH hierarchy that are not used for indexing. MeSH terms have been shown to be good indicators for representing the contents of a cluster (Yamamoto and Takagi, 2007), and are generally considered to be useful for the extraction of contents from whole documents without using complex NLP techniques (Struble and Dharmanolla, 2004).

The titles of the documents are used as they are without any further processing. On the other hand, the MeSH terms and the chemical substances (collectively defined as 'terms' hereafter) to be used for annotation are chosen for each cluster based on a scoring and filtering scheme.

The scoring is based on TF.IDF (Text Frequency–Inverse Document Frequency) (Manning and Schutze, 1999), which expresses the specificity and the coverage that the terms confer to the clusters. Specificity corresponds to in-cluster frequency, i.e. the probability of a term to belong to a cluster of interest. Coverage corresponds to out-cluster frequency, i.e. the probability of a term to belong to any other cluster. The score is the product of the in-cluster and the out-cluster probability. Formally,

$$TF = \frac{n_m^c}{n^c} \tag{1}$$

and

$$IDF = -\log\left(\frac{n_m}{N}\right) \tag{2}$$

where N is the total number of documents, n_m the number of documents in the whole set (corpus) containing term m, n^c the number of documents in cluster c and n_m^c the number of documents in cluster c containing term m. Then the score $s_m$of each term m is

$$s_m = TF \times IDF \tag{3}$$

Consequently, a ranked list of terms is generated, according to these scores.

In order to annotate the documents of a cluster, and subsequently the cluster itself, in a clutter-free manner, we manage the document coverage of each term in the ranked list using a threshold. The threshold controls how many new documents the candidate term describes, compared to the already accepted (and higher scoring) ones in the ranked list. In other words, we want to avoid complete overlap between the documents described by the candidate term and all the documents described by the already accepted terms. In our experiments, we set this threshold to one document.

Step (6):

The biological coherence of the MCL clusters is then validated by randomizing the assignment of the documents to the clusters (random clustering). Coherence is assessed by the χ^2-statistic (Weiss, 2002) and the counts of each term in each cluster, symbolized as $n_{m_v}^{C_k}$, where C_k is cluster k with k signifying the number of clusters. m_v is the v-th term. The terms used in this step are all terms contained in all documents of our set and not only the ones selected in the previous step of our methodology [Step (5)], thus avoiding any bias. Using the bootstrap method, we create 10 000 random clusterings (samples), where the number of clusters and their size remain the same as in the solution produced by MCL. The documents are randomly assigned to each cluster. Then for each cluster we count the number of documents that contain each term (Table 1—contingency table) and we calculate the χ^2-statistic for every sample (random clustering). Using the χ^2-statistic from the 10 000 samples, we build a histogram and a distribution graph based on kernel density estimation. If the χ^2-statistic calculated for the MCL clustering is at the right edge of our graphs, we can conclude that the terms are not independent of the clusters. Furthermore, we calculate the P-value for the χ^2-statistic for the MCL clustering.

Step (7):

The final step of the methodology involves the visualization of the clustering, documents and their titles and selected terms. This is achieved with BioLayout Express 3D (Goldovsky *et al*., 2005) that also allows the user to interact with results, for instance by searching for keywords, highlighting relevant documents, analyzing graph connectivity, linking nodes with external databases and so forth.

Table 1. A contingency table of clusters and MeSH terms

	Terms			
Clusters	m_1	m_2	…	m_v
C_1	$n_{m_1}^{C_1}$	$n_{m_1}^{C_1}$	…	$n_{m_v}^{C_1}$
C_2	$n_{m_1}^{C_2}$	$n_{m_2}^{C_2}$	…	$n_{m_v}^{C_2}$
…	…	…	…	…
C_k	$n_{m_1}^{C_k}$	$n_{m_2}^{C_k}$	…	$n_{m_v}^{C_k}$

It must be made clear that 2D or 3D space is only used for the presentation of the documents and the clusters to the end-user, and those coordinates of a citation have no further use. BioLayout uses a modified version of the Fuchterman and Rheingold graph layout algorithm in order to produce an 'aesthetically pleasing' layout of complex graphs (Goldovsky *et al*., 2005). Based on this algorithm, two documents that are similar, meaning they have a mutual connection with a high similarity score obtained from the 'related articles' tool of PubMed, will end up closer in the final graph than two documents which are weakly similar. Consequently, highly connected groups of similar documents will form tight clusters in the final graph.

In our methodology, apart from MCL and BioLayout Express 3D, we use perl scripts for processing the documents, as in Step (5), and the R statistical software package (R Development Core Team, 2007) for the statistical procedures as in Step (6).

3 RESULTS

In order to evaluate our methodology, we performed control experiments based on two different datasets that have already been used for evaluating the performance of TextQuest (Iliopoulos *et al*., 2001).

The first dataset consisted of 1660 documents obtained from two different queries. The first query was 'escherichia AND pili', returned 830 documents and is relevant to the organism *Escherichia coli* and pili, which are surface organelles. The second query was 'cerevisiae AND cdc*', returned the same number of documents (830) and is relevant to the cell division control genes of yeast.

The second dataset was derived by using the following terms in the queries: 'anterior-posterior AND drosophila' plus 'dorsal-ventral AND drosophila'. Both queries are relevant to the developmental axes of *Drosophila*. Since the queries are closely related some documents returned by the two queries were the same. Thus, the number of unique documents in the second dataset was 465.

The two datasets reflect two control tests of increasing difficulty: in the first dataset, the desired solution would clearly be two large, disconnected clusters since the concepts are drastically different; in the second dataset, the desired solution is less clear although it should reflect some biologically relevant concept groups (Iliopoulos *et al*., 2001).

3.1 First dataset

Thirty documents were not related to the other 1630 documents according to PubMed. Using an inflation value of 1.2 for the MCL algorithm, the 1630 documents were divided in two clusters. The first cluster included 818 documents referring to the *E.coli* query,

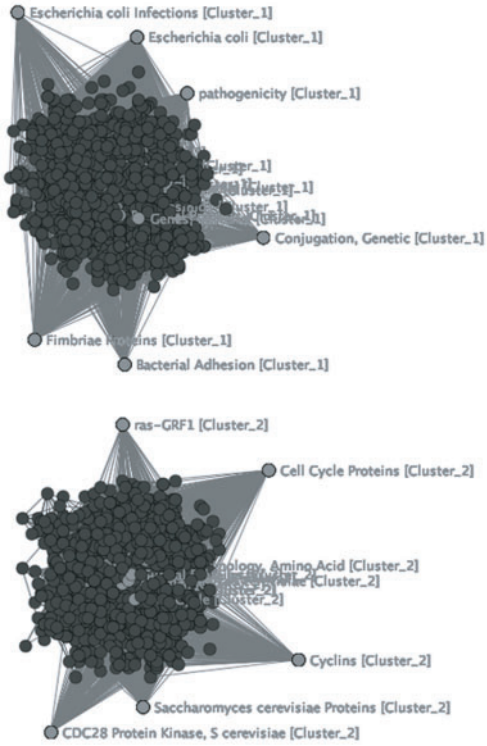

Fig. 2. BioLayout 2D graph of the documents (dark nodes) and the selected terms (light nodes) describing each cluster; changing the view to interactive 3D in the actual software would allow overlapping nodes to be seen clearly (we present only the 2D graph since it is more suitable for a 2D surface).

while the second cluster contained 812 documents referring to yeast and the cdc genes (Fig. 2).

Table 2 contains the terms selected using Step (5) of our methodology. Originally Cluster 1 contained 1349 unique terms and Cluster 2 contained 1551, but the selected ones were only 16 and 14, respectively. The terms described general concepts relevant to each query, like *bacterial adhesion*, *fimbriae proteins*, *cell cycle proteins*, etc.

Significantly, for different inflation values we observed clusterings with different granularities. For example, when we used an inflation value of 1.3, MCL produced four clusters. The first of the two additional clusters contained four documents relevant to gonococcal pili antigens and described by the MeSH term *Neisseria gonorrhoeae*. All the documents were from the query related to *E.coli* and pili. The second additional cluster contained three documents from the yeast query and was described by the MeSH terms *Ribonucleoprotein*, *U1 Small Nuclear* and *Schizosaccharomyces*. It is thus encouraging that PuReD-MCL is able, through the tweaking of a single parameter, to produce clusterings of variable detail. In this case two small subclusters were extracted from the original two clusters describing more specialized concepts.

3.2 Second dataset

Out of the 465 documents of the second dataset, 18 did not relate to any of the remaining 447 documents.

Table 2. The selected terms describing each cluster for dataset 1

Terms	
Cluster 1	Cluster 2
1. Bacterial adhesion	1. CDC28 protein kinase, *S.cerevisiae*
2. Bacterial proteins	2. Cell cycle proteins
3. Conjugation, genetic	3. Cell cycle
4. *Escherichia coli* infections	4. Cyclins
5. *Escherichia coli*	5. Fungal proteins
6. Fimbriae proteins	6. Genes, fungal
7. Fimbriae, bacterial	7. Mutation
8. Genes, bacterial	8. *Saccharomyces cerevisiae* proteins
9. Humans	9. *Saccharomyces cerevisiae*
10.Mannose	10.Sequence homology, amino acid
11.Plasmids	11.Cytology
12.Analysis	12.Enzymology
13.Immunology	13.Metabolism
14.Microbiology	14.Ras-GRF1
15.Pathogenicity	
16.Ultrastructure	

Since we wanted to compare our results with those of TextQuest, where the clusters were three, we experimented with MCL and inflation values that resulted in a similar number of clusters. An inflation value of 1.2 yielded two clusters, as described in the original control experiments (Iliopoulos *et al*., 2001). The first cluster contained 443 documents and corresponded mainly to the dorsal–ventral development of *Drosophila*, while the second cluster (four documents) covered the homeobox pbx1 protein, uncovering parts of development along the anterior–posterior axis. In more detail, three abstracts (PubMed ID: 7791786, 7565734, 10067897) in this cluster were fully interconnected, while the fourth document (PubMed ID: 7914870) connected also to documents outside the cluster, because it referred generally to homeotic proteins in *Drosophila*, as expected. Furthermore, using an inflation value of 1.3, we separated the documents to six, more granular, clusters. We chose to keep this last clustering solution, since it provided more details about the biological information in the documents and was directly comparable with a previous study (Iliopoulos *et al*., 2001).

Figure 3 presents a graph of the six clusters where the dark nodes are the documents and the light ones the selected terms. Also, Table 3 includes details about the number of documents and a short description of each cluster. We can see that the first cluster contained most of the documents and described the process of segmentation and embryonic patterning in general. Interestingly, the second cluster contained documents specifically related to the genetic machinery for the concept in cluster 1, the homeobox genes (Hox), which are the subject of intense research on merit: the Hox genes control the formation of segment specific structures in the anterior–posterior axis, are highly conserved, crucial for the development of *Drosophila*, and exhibit unique behavior, like colinearity (Lappin *et al*., 2006). Cluster 3 involved the sonic hedgehoc proteins (Shh) and their role in the zone of polarizing activity (ZPA), which are related to the differentiation in the anterior–posterior axis (Marigo *et al*., 1996). Cluster 4 contained documents about the Wnt/WG signaling pathway studied in *Caenorhabditis elegans*, *Xenopus* and *Drosophila*. The Wng/Wg pathway is evolutionary conserved

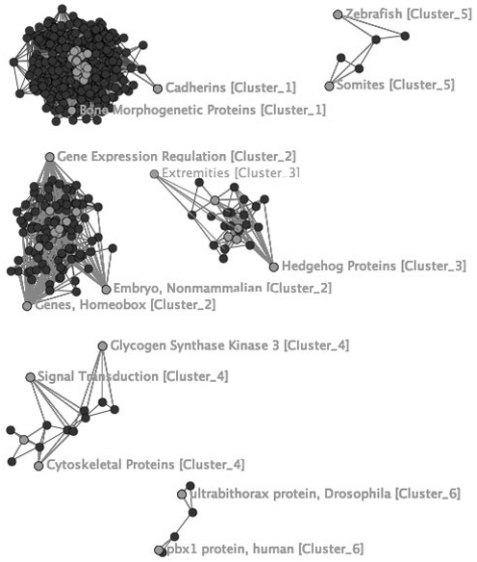

Fig. 3. The six clusters from the second dataset.

and plays an important role in normal development and cancer for many organisms, including *Drosophila* and human (Xiang, 2003). Wg is required to establish the anterior–posterior axis during embryogenesis, but it is also necessary for correct dorsal–ventral axis patterning in the wing imaginal disc, at different steps of development (Zhang *et al*., 1998).

Cluster 5 described the morphogenesis of the zebrafish (tropical fish) brain and the genes that regulate it (Schier *et al*., 1996). It also included information about the genetic regulation of the somite formation in vertebrates (Rawls *et al*., 2000). Finally, Cluster 6 contained documents that described the binding of pbx1 protein (homeoprotein Extradenticle) to DNA and homeotic (Hox) protein Ultrabithorax; thus, the pbx1 protein affects development across the anterior–posterior axis of animals, as is already known (Passner *et al*., 1999).

It is intriguing to note that the documents related to the dorsal–ventral development axis were assigned to cluster 1, whereas development in both axis was described by cluster 4 (Wnt/Wg pathway) and cluster 5 (in vertebrates).

Table 4 lists the selected terms from Step (5) of the methodology for each cluster. Although the selected terms are rather generic, they can identify the different biological concepts of each cluster and help the user find specific information. Inside the parentheses there are the original number of terms contained in each cluster, before the selection of Step (5) of our methodology. In bold italics are the terms that are also found in TextQuest and underlined are the ones that are similar, but not exactly the same. Four out of the 12 are exactly the same, like 'mutation', 'phenotype', etc. The other eight terms are similar to the ones from TextQuest. For example, in TextQuest the term 'embryo' can be matched to 'Chick embryo' or 'Embryo, nonmammalian' in PuReD-MCL or 'kinase' in TextQuest matched to 'Glycogen Synthase Kinase 3' in PuReD-MCL. It must be noted that the terms of the third cluster of TextQuest referring to the early stages of the development are found in documents of the first and the second cluster in PuReD-MCL. This is expected since

Table 3. Cluster description of dataset 2

Cluster	Size	Short Description
Cluster 1	322	Segmentation and embryonic patterning
Cluster 2	88	Homeobox genes
Cluster 3	19	Sonic Hedgehoc proteins (Shh) and Zone of Polarizing Activity (ZPA)
Cluster 4	10	*C.Elegans*, *Xenopus* and *Drosophila* Wnt/WG signaling pathway
Cluster 5	4	Morphogenesis of zebrafish brain and somitogenesis
Cluster 6	4	Binding properties of pbx1 protein

PuReD-MCL did not produce a cluster referring solely to the early developmental stages.

These differences are partly due to the fact that TextQuest performs tokenization and stemming and thus some phrases are cut into several different words, or abbreviated (e.g. 'embryo' from 'embryology'). On the other hand, MeSH terms are usually more descriptive, they do not need any NLP processing since they are already built in computer readable format, and they refer to information in the whole document and not just the abstract or the title.

In order to statistically assess the biological coherence of the clustering, we developed Step (6) of our methodology where we use the χ^2-test. We applied χ^2 to our clustering and compared it to the χ^2-scores of random clusterings. We produced 10 000 random clusterings by keeping the size and number of clusters the same as in the original clustering (see Table 3 for dataset 2) and randomizing the assignment of documents to clusters.

The χ^2-score for the solution was 41 556.39, df = 17 366, P-value = 0. We also performed a Monte Carlo simulation (Hope, 1968) to calculate the P-value and the result was also significant (P-value = 0.00049). The comparison of the χ^2-scores between the solution and the random clusterings resulted in a significant difference. The random clusterings had a maximum score of approximately 30 000, which was much less than 41 556. Figure 4 depicts a histogram of the χ^2-scores from the random clusterings. Superimposed on the histogram is the graph (continuous line) of the kernel density estimation of the χ^2-scores distribution based on the random clusterings. Both graphs have a very small number of values around 30 000 while most of the scores are between 12 000 and 22 000.

We compared our clustering with the results from TextQuest in order to explore the biological information revealed by each methodology. TextQuest defined three clusters (Iliopoulos *et al*., 2001), the first one containing documents about the process of segmentation and embryonic patterning, the second one about the embryonic dorsoventral axis specification in *Drosophila*, while the third cluster referred to genes involved in both the anterior–posterior and the dorsal–ventral axes, during oogenesis. In comparison, the PuReD-MCL algorithm detected smaller clusters with higher granularity and thus better specificity for certain terms, e.g. pbx1, Shh and ZPA, as shown (Table 3).

In summary, the original question posed by TextQuest was how are the two *Drosophila* embryonic axes established. Thus, the submitted queries in the PubMed had the keywords 'anterior–posterior' and 'dorsal–ventral'. Both, TextQuest and PuReD-MCL

Table 4. The selected terms describing each cluster for dataset 2.

Terms	
Cluster 1 (683)	Cluster 2 (307)
1. Bone morphogenetic proteins	1. Embryo, non-mamalian
2. Cadherins	2. Gene expression regulation
3. *Drosophila* proteins	3. Gene expression
4. ***Drosophila***	4. Genes, homeobox
5. *Drosophila melanogaster*	5. Genes, regulator
6. Female	6. Homeodomain proteins
7. Genes, insect	7. Polycomb protein, *Drosophila*
8. Insect hormones	8. Transcription factors
9. Morphogenesis	9. Transcription, genetic
10. ***Mutation***	10. Anatomy and Histology
11. ***Phenotype***	Cluster 4(102)
12. ***Proteins***	1. *Caenorhabditis elegans* proteins
13. Signal transduction	2. Cytoskeletal proteins
14. Transcription factors	3. Glycogen synthase kinase 3
15. Wing	4. Signal transduction
16. Growth and development	Cluster 5 (44)
17. Metabolism	1. Somites
18. Physiology	2. Zebrafish
Cluster 3 (146)	Cluster 6 (52)
1. Chick embryo	1. pbx1 protein, human
2. Extremities	2. Ultrabithorax protein, Drosophila
3. Gene expression Regulation, developmental	
4. Hedgehoc proteins	
5. Limb bud	
6. Mice	

Original number of terms obtained is indicated in parentheses.

answered the above question, but from a different perspective producing more than two clusters.

In future work, it will be crucial to devise more sophisticated learning sets to benchmark the performance of these algorithms/approaches, in the context of similar evaluations, for example, BioCreAtIve (Hirschman *et al*., 2005) and GENIA (Kim *et al*., 2003).

3.3 Performance evaluation

In order to provide a performance evaluation and further evidence of the scalability of our methodology we used the query 'microbial phenotype' in order to cluster 6524 documents from PubMed. The UNIX command 'time' was used in order to measure the time required to retrieve (download through an Internet connection) the 'related articles' xml files for each of the 6524 documents, process them, build the graph and cluster the graph using the MCL algorithm. The hardware used was an Intel 2.16 GHz Core Duo processor, 2 GB of RAM. The connection to the Internet was through ADSL 1 Mb download and 256 Kb upload. The downloaded 'related articles' information from PubMed was ∼112 MB. The inflation value for MCL was 1.2. The 'time' UNIX command produced the results in Table 5.

The output includes (i) the elapsed real time between the invocation and termination of the process ('real'), (ii) the user CPU time ('user') and (iii) the system CPU time ('sys'). The user time

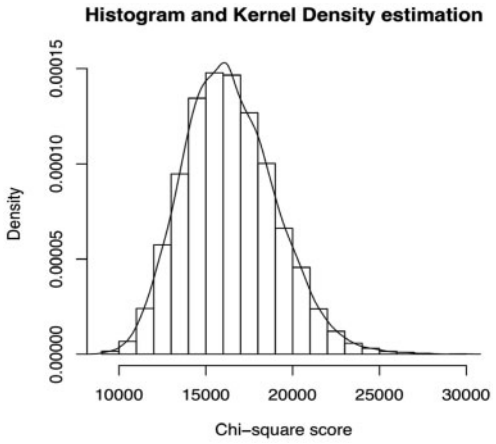

Fig. 4. Histogram and kernel density distribution graph of the χ^2-statistic of the random clustering for the second dataset.

Table 5. UNIX 'time' command results

Type of time measured	Time
real	26 min 43.898 s
user	8 min 46.848 s
sys	0 min 18.464 s

is the time used by the program itself and any library subroutines it calls. The system time is the time used by system calls invoked by the program (directly or indirectly). The sum of user plus system is the total direct CPU cost of executing the program. To cluster a dataset 1000 times larger, the time requirement would be of the order of 8000 min, or 5 days or so—on a faster server, this could be reduced to a matter of hours. Also note that the difference between real and user plus system time, is the sum of all of the factors that may delay execution, for example delays due to network speed. Furthermore, there are parameter dependencies, e.g. as the inflation value in MCL increases, the time required for the clustering decreases.

4 DISCUSSION AND FUTURE WORK

PuReD-MCL can efficiently produce meaningful and interpretable clusters of PubMed documents by creatively using a combination of existing and well-established algorithms and tools. It is the first time that the MCL algorithm is used in the field of biomedical text mining, although it has been used before in the computational linguistics field for synonym dictionary improvement (Gfeller *et al*., 2005) and word sense disambiguation (Dorow *et al*., 2005) for the French language.

The fact that the method is unsupervised and relies mostly on existing systems is a major advantage, keeping its computational complexity to an absolute minimum. Although, the use of precomputed sets of 'related articles' from PubMed enables us to avoid applying computationally expensive NLP techniques to the documents, this has two shortcomings. First, it makes the method suitable only for documents contained in the PubMed database. Second, the relatedness of the documents is based on an Euclidean distance (cosine), which suffers from severe defects as to keyword weightings and correlations (Mochihashi *et al.*, 2006).

Other measures, proposed elsewhere (Mochihashi *et al.*, 2006), could reveal different relatedness between the same documents and thus affect the clustering.

Nevertheless, the ability of MCL to produce clusterings of different granularities with the adjustment of the inflation value allows a detailed analysis of the information contained in the documents through a form of controlled hierarchical clustering. In particular, starting from a graph where the documents are connected in the minimum number of clusters, the user can use MCL to start splitting the documents into subclusters. Gradually, an increase of the inflation value will produce an increasing number of clusters, until there are no more connections in the graph that can be eliminated, resulting to the maximum number of clusters.

In the future, we intend to cluster complete sets of PubMed-available documents for specific organisms, such as *Drosophila*, or human, to capture and explore different research topics and new discoveries through high-throughput text mining.

ACKNOWLEDGEMENTS

We would like to thank all the members of the CGU for insightful discussions and comments.

Funding: N.D. is supported by an ENTER grant from the General Secretariat for Research and Technology of the Hellenic Ministry of Development. The CGU at CERTH is supported by the Networks of Excellence BioSapiens (contract number LSHG-CT-2003-503265) and ENFIN (LSHG-CT-2005-518254), both funded by the European Commission.

Conflict of Interest: none declared.

REFERENCES

Ananiadou,S. *et al*. (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol*., **24**, 571–579.

Brohee,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488.

Dorow,B. *et al*. (2005) Using curvature and Markov clustering in graphs for lexical acquisition and word sense discrimination. In *MEANING-2005, 2nd Workshop organized by the MEANING Project, February 3–4, 2005*, Trento, Italy.

Enright,A.J. *et al*. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.

Gfeller,D. *et al*. (2005) Synonym dictionary improvement through Markov clustering and clustering stability. In *International Symposium on Applied Stochastic Models and Data Analysis 2005*, Brest, France. pp.106–113.

Goldovsky.L. *et al*. (2005) BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl. Bioinform.*, **4**, 71–74.

Hirschman,L. *et al*. (2005) Overview of BioCreAtIve: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6** (Suppl. 1), S1.

Hope,A.C.C. (1968) A simplified Monte Carlo significance test procedure. *J. R. Stat. Soc. B*, **30**, 582–598.

Hu,Z.Z. *et al*. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.

Iliopoulos,I. *et al*. (2001) TextQuest: document clustering of medline abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput*., **6**, 384–395.

Izumitani,T. *et al*. (2004) Assigning gene ontology categories (GO) to yeast genes using text-based supervised learning methods. In *Proceedings of the IEEE Computational. Systems Bioinformatics Conference (CSB 2004),* IEEE Computer Society, Stanford, USA, pp. 503–504.

Kim,D.J. *et al*. (2003) GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**, i180–i182.

Lappin,T.R. *et al*. (2006) HOX genes: seductive science, mysterious mechanisms. *Ulster Med. J.*, **75**, 23–31.

Manning,C.D. and Schutze,H. (1999) *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, USA.

Marigo,V. *et al*. (1996) Sonic hedgehog differentially regulates expression of GLI and GLI3 during limb development. *Dev. Biol*., **180**, 273–283.

Mochihashi,D. *et al.* (2006) Learning an optimal distance metric in a linguistic vector space. *Syst. Comput. Jpn*., **37**, 12–21.

Nenadic,G. *et al*. (2003) Terminology-driven mining of biomedical literature. In *Proceedings of the 2003 ACM Symposium on Applied Computing*, ACM, Florida, USA, pp. 83–87.

Passner,J.M. *et al*. (1999) Structure of a DNA-bound ultrabithorax-extradenticle homeodomain complex. *Nature*, **397**, 714–719.

R Development Core Team (2007) *R: a language and Enviroment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radhakrishna,C.R. and Bhaskara,M.R. (1998) *Matrix Algebra and its Applications to Statistics and Econometrics*. World Scientific Pub Co Inc., Singapore.

Rawls,A. *et al*. (2000) Genetic regulation of somite formation. *Curr. Top. Dev. Biol*., **47**:131–54.

Raychaudhuri,S. *et al*. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res*., **12**, 203–214.

Salton,G. (1970) Automatic text analysis. *Science*, **168**, 335–343.

Schier,A.F. *et al*. (1996) Mutations affecting the development of the embryonic zebrafish brain. *Development*. **123**, 165–78.

Struble,C.A. and Dharmanolla,C. (2004) Clustering MeSH representations of biomedical literature. In *Proceedings of BioLINK 2004*, Association for Computational Linguistics, Boston, May 6 2004, pp. 41–47.

Theodosiou,T. *et al*. (2007) Gene functional annotation by statistical analysis of biomedical articles. Int. *J. Med. Inform*., **76**, 601–613.

van Dongen,S. (2000) Graph clustering by flow simulation. PhD thesis, University of Utrecht. Available at http://micans.org/mcl/lit/svdthesis.pdf.gz (last accessed on July 17 2008).

Weiss,N.A. (2002) *Introductory Statistics*. 6th edn. Addison-Wesley, USA.

Wheeler,D.L. *et al*. (2007) Database resources of the National Centre for Biotechnology Information. *Nucleic Acids Res*., **35**(Database issue), D5–D12.

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.

Xiang,Y. (2003) A wingless flight. *PLoS Biol.*, **1**, e49.

Yamamoto,Y. and Takagi,T. (2007) Biomedical knowledge navigation by literature clustering. *J. Biomed. Inform*., **40**, 114–130.

Zhang,J. and Carthew,R.W. (1998) Interactions between Wingless and DFz2 during Drosophila development. *Development*, **125**, 3075–85.