Clustering in Relational Biological Data

Aynur Dayanik AYNUR@CS.RUTGERS.EDU
Computer Science Department, 110 Frelinghuysen Road, Piscataway, NJ 08854 USA

Craig G. Nevill-Manning CRAIGNM@GOOGLE.COM
Google, Inc., 1440 Broadway, New York, NY 10018 USA

Abstract

The scientific endeavor of biology is becoming increasingly reliant on data in electronic form, and it is therefore necessary for biologists to manage and understand large quantities of data. Publicly available data including biological sequences, biological structures, and literature in the life sciences have grown to such an extent that computing is essential simply to store and access it. Here we describe a clustering approach by exploiting the relational structure of biological data to help with the next step: to enhance understanding of the data by combining techniques from information retrieval with those from bioinformatics. By computing over a network of sequence-structure-literature relationships it is possible to infer clusters of related articles, sequences and structures. This paper describes the general framework and its application to several biological domains.

1. Introduction

The growth of bioinformatics has coincided with the growth of the worldwide web. This happy coincidence, in conjunction with savvy policy on the part of publishers and the National Library of Medicine, has resulted in a body of data that is singularly well connected. For example, whenever a paper is published in the biological literature, any biological sequences or structures that were determined or analyzed in the course of the research must be submitted to the appropriate databases. The corresponding abstract in MEDLINE is then annotated with the ID of the sequence or structure. This linking allows researchers to find experimental data very easily once they have identified a paper of interest, or conversely to find an analysis of a particular sequence or structure. The National Center for Biotechnology Information (NCBI), part of the National Library of Medicine, provides online access to MEDLINE abstracts, GenBank sequences, and many other data types, through their Entrez system. The ability to browse data and literature seamlessly is important, but the underlying data has much greater potential.

Clustering is the task of grouping a set of objects into different subsets such that objects belonging to the same cluster are highly similar to each other. Convential clustering algorithms employ distance (or similarity) measure to form the clusters (Kirsten et al., 2000). On the other hand, graph partitioning algorithms exploit the structure of a graph to find highly connected objects. Rich relational structure of biological data can be represented as a graph for clustering biological data. Clustering biological data would be useful not only for exploring the data but also for discovering implicit links between the objects.

Here we describe a technique for clustering of biological objects: sequences, structures and literature. We use METIS, a multilevel graph partitioning system, to form the clusters. This process identifies subsets of nodes that are highly connected to each other, but are less strongly connected to the rest of the graph. These clusters are formed based on the pairwise relationships among biological data, so we can evaluate their topical cohesiveness by examining independent metadata such as Gene Ontology (GO) annotations and terms in MEDLINE abstracts. We also evaluate the clusters by hand for relevance, and find that the clusters are highly topical.

The organization of the paper is as follows. The next section describes the databases we used. In section 3, we describe the construction of a graph from the databases, and then present our graph partitioning approach in section 4. Section 5 describes the BioIR system we built. In section 6, we present and discuss the empirical results to assess the quality of clusters. Finally, we end with a summary.

2. Data Sources

In this section, we will briefly describe the data sources we used to construct our graph.

MEDLINE**:** MEDLINE is a digital collection of life science literature consisting of over twelve million abstracts. MEDLINE articles contain links to the sequences and structures that the article discuss. The MEDLINE collection we used contained about 100,000 abstracts.

SWISS-PROT**:** The Swiss Protein Database (SWISS-PROT) is a curated protein sequence database (Bairoch & Apweiler, 2000). The database contains high-quality annotation including descriptions of each protein's function. SWISS-PROT entries are cross-referenced to several other databases, including MEDLINE, PROSITE and the PDB. SWISS-PROT has about 120,000 protein sequences.

PDB: The Protein Data Bank (PDB) contains 3–D structural data of biological macromolecules (proteins and nucleic acids) (Berman et al., 2000). The PDB entries are also cross-referenced to the primary citations in MEDLINE and other databases including ENZYME and SWISS-PROT. PDB has about 20,000 structures.

3. Constructing the Graph

Using the relationships between biological data objects, we construct a weighted undirected graph where nodes correspond to entries from the databases listed in Section 2, including MEDLINE abstracts, protein sequences from SWISS-PROT, structures from PDB. Table 1 shows excerpts from a MEDLINE record that contains references to three structures in PDB, along with the title and abstract of the paper.

Edges in the graph correspond to explicit links between entries encoded in the databases, such as the sequence annotations in MEDLINE abstracts, and pairwise similarity relationships between same type of objects. We use BLAST (Altschul et al., 1997), a sequence alignment technique, to compute similarities between protein sequences. We employ MG[1] (Witten et al., 1999), a full-text retrieval engine, to compute similarities between MEDLINE abstracts. We use the SCOP (Murzin et al., 1995), a database of hierarchical classification of PDB entries based on structural similarities, to relate PDB entries to each other. We assume a relationship between two PDB entries if they are in the same leaf of the SCOP hierarchy.

Figure 1 shows an example graph of biological entities, including edges between abstracts and sequences, abstracts and structures, and between sequences and structures as well as between same type of objects by similarity relationships.

We assign weights to edges as follows. We assign a weight of 100 to explicit edges encoded by the databases (as for 100% relatedness). We normalize the similarity scores between same type of objects computed by MG and BLAST to the range [1,100]. We assign a weight of 100 to PDB-PDB relationships obtained using SCOP since SCOP classifications are done by biologists.

PMID- 11807546
TI - Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin.
AB - Oedema factor, a calmodulin-activated adenylyl cyclase, is important in the pathogenesis of anthrax. Here we report the X-ray structures of oedema ...
SI - **PDB**/1K8T
SI - **PDB**/1K90
SI - **PDB**/1K93
...

Table 1. A sample MEDLINE file linking to PDB entries

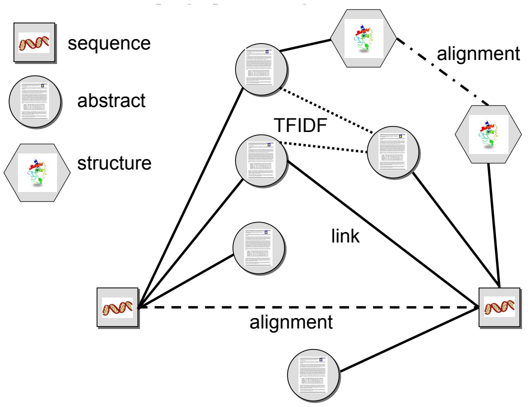

Figure 1. An example graph of sequences, abstracts, and structures related by explicit references and similarity relationships.

4. Graph Partitioning

The objective of graph partitioning is to partition the graph into k roughly equal parts such that the sum of the weights connecting different parts is minimized, thereby each part is highly similar. The graph partitioning problem is NP-complete. However, many heuristics have been developed that find a reasonably good partition.

Traditional graph partitioning algorithms compute a partition of a graph by operating directly on the graph, and they are usually slow. On the other hand, multilevel graph partitioning algorithms reduce the size of the graph by collapsing vertices and edges, partition the smaller graph, and then coarsen it to construct a partition for the original graph. These algorithms are generally fast and produce high-quality partitions. We chose the *pmetis* program pro-

[1] Available at http://www.cs.mu.oz.au/mg/.

vided by the METIS[2] software, a publicly available graph partitioning software package. The partitioning algorithm used by *pmetis* is based on multilevel recursive bisection described in (Karypis & Kumar, 1998).

5. The BioIR System

We built a system, called BioIR, to test our approach. We stored all the entities in the databases described in Section 2, and the relationships between them in a MySQL database. Then we created a graph using these tables as explained in Section 3, and stored the nodes and the edges between them in the graph back in MySQL.

The graph is not completely connected: there are many disconnected subgraphs. There is one large connected component as well as 864 connected components of size at most 100. We partitioned the largest connected component to obtain about 1000 clusters of 200 nodes each. We chose 1000 as the number of clusters since the PROSITE, database of sequence motifs/patterns, has about 1000 entries. Therefore, we can use PROSITE for quantitative evaluation of the clustering. Also, browsing clusters of size 200 would be manageable by biologists. We kept all other small size connected components as clusters themselves and stored all the resulting clusters in a MySQL database.

5.1. Identifying Descriptive Terms From Abstracts

We aim to identify words that best describe the set of documents in clusters by analyzing the MEDLINE articles of the clusters. These descriptive words can be used as index terms to identify the contents of the clusters. We identified the descriptive words as follows. We considered the words in the title and abstract of all articles in a cluster after eliminating stop words. We removed all punctuation, and converted all uppercase letters to lowercase. Then we ranked the resulting words by calculating p-values considering the entire set of MEDLINE articles in our collection. p-value calculation was described in subsection 6.4. We kept the top twenty most significant words, the ones having the smallest p-values, in our database for each cluster. We use the resulting set of twenty words to index the clusters, and build a search utility against this index using MySQL.

6. Experimental Results and Discussion

First, to quantify the quality of the produced clustering, we computed the entropy and purity of the clustering for SWISS-PROT and PDB entries by taking PROSITE and SCOP classifications as reference classifications. Note that we did not use PROSITE at all to obtain the clustering. However, we used SCOP to relate PDB entries. We are interested in seeing how well we recover the relational structure of PDB entries.

Second, we evaluated our system on several biological domains, described in subsection 6.2 by carrying out a user study to understand the quality of the clusters. Also, to quantify the quality of the sample clusters analyzed by a domain expert, we analyzed the SWISS-PROT to GO mappings and MEDLINE abstracts in clusters to extract common words to see their relevance to the topics of interests.

6.1. Evaluation of Overall Clustering quality

In general, two different metrics are used to measure the quality of a clustering. The first metric is the widely used *entropy* measure that considers how the various classes of objects are distributed within each cluster, and the second measure is the *purity* measure that considers the extend to which each cluster contained objects from primarily one class.

Let C_r denote a particular cluster of size n_r. The entropy of this cluster is defined as

$$E(C_r) = -\frac{1}{logq}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\;log\frac{n_r^i}{n_r}, \qquad (1)$$

where q is the number of classes in the dataset, and n_r^i is the number of objects of the ith class that were assigned to the rth class. The overall entropy of the clustering, where k is the number of clusters, is then defined as the sum of the individual cluster entropies weighted according to the cluster size:

$$Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(C_r). \qquad (2)$$

A perfect clustering will be the one consisting of clusters that contain objects from only a single class. In this case, the entropy will be zero. In general, the smaller the entropy values, the better the clustering solution is.

The purity of this cluster is defined as

$$P(C_r) = \frac{1}{n_r}max_i(n_r^i). \qquad (3)$$

The overall purity of the clustering is defined as a weighted sum of the individual cluster purities and is computed as

$$Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(C_r). \qquad (4)$$

[2] Available at http://www-users.cs.umn.edu/~karypis/metis/.

In general, the larger the values of purity, the better the clustering solution is.

Table 2 shows the entropy and purity values computed for SWISS-PROT and PDB entries using PROSITE and SCOP classifications as reference classifications, respectively. Recall that the closer the entropy value to 0, the better the clustering is. Also, the closer the purity value to 1, the better the clustering is.

As a baseline, we created clusters by randomyly assigning the objects in our graph to 1000 clusters and computed the entropy and purity measures for SWISS-PROT, PDB entity types. Table 2 also shows the average results of 10 random partitioning experiments. The average entropy value for 10 random partitionings is much higher than those of our graph partitioning, and the average purity value for 10 random partitionings is much lower than those of our graph partitioning. These suggest that we discover meaningful groupings by our graph partitioning method.

	SwissProt		**PDB**	
Method	Entropy	Purity	Entropy	Purity
METIS	0.1180	0.4129	0.1145	0.7334
Random	0.5596	0.0246	0.4358	0.0873

Table 2. Entropy and purity values for SWISS-PROT and PDB clusterings using PROSITE and SCOP classifications as references, respectively.

6.2. Biological Domains

The following biological domains were carefully examined by our domain expert, a Ph.D. candidate in Molecular Biology. We give a brief description of each domain below.

Calmodulin: Calmodulin is a ubiquitous intracellular receptor for calcium ions that functions by changing its shape upon binding to calcium so that it can bind to and activate/inactivate other proteins. Most proteins activitated by calmodulin are so-called CaM-kinases.

Chemotaxis: This is a bacterial signaling pathway involved in chemotaxis. Repellents activate receptors that, with the assistance of CheW, activate CheA. Attractants inhibit CheA. CheA activates CheY which causes the flagella to rotate such that the bacteria tumble. CheZ inactivates CheY.

Rhodopsin and Gt: Rhodopsin is a 7-pass transmembrane G-protein linked receptor containing a pigment, 11-cis-retinal. Light changes the structure of the pigment which causes Rhodopsin to bind with transducin (Gt), a trimeric G-protein. Upon binding, Gt looses its alpha subunit which diffuses and binds GMP phosphotase, activating it and eventually leading to signaling.

U1 U2 U5 U4 U6 spliceosome: The spliceosome is a protein, RNA complex reponsible for splicing introns out of nascient mRNA during its maturation. U1, U2, U4, U5 and U6 are among the different snRNPs present in Eukaryotic nuclei - they consist of both protein and small RNA molecules.

Ubiquitin: Ubiquitin-dependent protein degradation plays a role in many cellular processes including transcriptional regulation, cell cycle progression and DNA repair. Ubiquitin is a highly conserved 8kDa protein whose many cellular functions are mediated by its covalent ligation to other proteins.

Apoptosis: Apoptosis, or programmed cell death, plays a fundamental role during tissue development, injury and degeneration. The biochemical pathways of programmed cell death are also used to destroy cells with damaged DNA and cells that are infected with viruses.

p53 Signaling Pathway: p53 is a transcription factor whose main function is to prevent the cell from progressing through the cell cycle when DNA damage has occurred. p53 may either halt the cell cycle until the DNA can be repaired or else it may cause the cell to undergo apoptosis.

Insulin Signaling Pathway: Insulin, a small protein that acts as a hormone, is secreted by the pancreas in response to increased glucose levels in the blood. Most cells of the body have receptors which bind insulin. Upon binding of insulin, the cell activates other receptors designed to absorb glucose from the blood stream into the cell. Insulin is a necessary hormone and insulin deficiency or resistance results in diabetes.

6.3. Expert Analysis

The domain expert evaluated sixteen clusters – two clusters for each topic of interest, e.g, calmodlin, apoptosis, etc. Three different types of entities were considered (PDB, SWISS-PROT and GO terms) to determine how many of them were relevant to the topic of interest. Although GO was not used to obtain the clusters in any way – therefore, they are not in the clusters, GO terms were assigned to the clusters using SWISS-PROT to GO mappings as described in subsection 6.4. Also, overall cluster qualities are reported for each cluster manually examined.

Table 3 shows the evaluation results judged by the domain expert. For each entity type, a relevancy score between 1 and 10 was assigned where 10 means all entities of that particular type are highly topical and 1 means that none of them are relevant. Almost all entity types for all clusters have high scores. Therefore, we can conclude that all the sections evaluated by the expert are highly relevant to the topics considered.

Topic	Cluster	PDB	SW	GO term	Overall
calmodulin	1794	10	10	8	10
calmodulin	1815	5	7	5	5
rhodopsin	1402	10	9	10	10
rhodopsin	1400	3	5	10	7
spliceosome	1634	10	10	10	10
spliceosome	1648	N/A	8	6	7
chemotaxis	1072	9	9	9	9
chemotaxis	1071	5	6	4	5
apoptosis	1670	10	10	10	10
apoptosis	1669	10	9	10	10
ubiquitin	1665	7	2	8	8
ubiquitin	1666	7	10	9	8
insulin	1473	10	10	10	9
insulin	1472	10	7	10	9
p53	1674	10	8	10	9
p53	1722	7	7	7	8

Table 3. Evaluation of sample clusters by our domain expert. Scores range from 1 to 10, where 10 means all of the objects are relevant, and 1 means none of them are relevant.

Biologists note that the SWISS-PROT to GO mapping is incomplete because not all SWISS-PROT sequences are fully annotated. For example, in one cluster for the topic “apoptosis”, the SWISS-PROT gene for E1B is not annotated with apoptosis even though it is involved in apoptosis. Similarly, in another cluster for the topic “apoptosis”, the SWISS-PROT annotation for the CASP-1 genes do not refer to apoptosis, but some MEDLINE articles indicate that it is involved in apoptosis. So, since the SWISS-PROT GO annotation is incomplete, the relevance scores that we obtain based on the GO terms through automated means may underestimate the relevance of the cluster contents.

Another interesting point is that our domain expert first thought that 30S ribosomal protein in cluster 1665 was unrelated to ubiquitin, therefore assigned a score of 2. However, the immediate links to MEDLINE articles as provided by our system suggested that it should be relevant. We asked to reconsider whether 30S ribosomal protein could be related to ubiquitin, and the SWISS-PROT sequences in this cluster were reevaluated. After examining some immediate neighbors (MEDLINE articles) of these entities in the graph, our domain expert found out that they were indeed relevant to ubiqutin, and now believes that the GO terms assigned to these clusters (nucleus and structural constituent of ribosome) are very relevant to ubiquitin. This ’discovery’ aspect of our system is important – it demonstrates that the clusters can bring to light relationships that are not obvious at first glance.

6.4. Correlation between clusters and GO categories – GO Term Assignment to Clusters

The Gene Ontology Consortium (2000) produces a controlled vocabulary for genes and gene products, called GO. GO[3] provides three structured networks of defined terms to describe gene product attributes. These three GO ontologies are referred to as Biological Process, Molecular Function and Cellular Component.

To show how much correlation we obtained between clusters and GO categories, we assigned GO terms to the clusters using the SWISS-PROT to GO mappings. Before explaining how we did this, it is important to note that we did not use GO to construct the graph; we just use it to provide a biological validation.

p-value Calculation:

Consider a two class population, and suppose we take a sample of size n from this population. Let A and B represent the number of objects for two classes, and N be the total number of objects. Let a, b and n be the corresponding numbers in our sample population. Thus, $A + B = N$, and $a + b = n$. Let us define the hypotheses

$$H_0 : \text{type-}A \text{ objects appear at random,}$$
$$H_a : \text{not at random.}$$

We reject the null hypothesis H_0 if

$$\text{p-value} = \textstyle\sum_{x \geq a} \binom{A}{x}\binom{B}{n-x}/\binom{N}{n}, \qquad (5)$$

i.e., the probability of observing at least a numbers of class A at random is close to 0, e.g., p-value ≤ 0.001. Since the calculation of (5) is computationally expensive, we use an approximation instead. If the objects were selected with replacement, then the number of type-A objects in a sample of size n will have approximately Binomial distribution with success probability $p = A/N$, and the p-value becomes

$$\text{p-value} = \textstyle\sum_{x \geq a} \binom{n}{x} p^x (1-p)^{n-x}. \qquad (6)$$

We consider the entire set of Gene Ontology (GO) annotations for the clusters. For each GO term for SWISS-PROT sequences within a specific cluster, we compute a p-value as in (6) where

N = total number of SWISS-PROT sequences,

A = actual number of A GO category SWISS-PROT sequences,

n = number of SWISS-PROT sequences in the cluster,

a = number of A GO category SWISS-PROT sequences in the cluster,

p = the proportion of the SWISS-PROT sequences containing A GO category.

[3] http://www.geneontology.org/

We give a general GO biological assessment to a cluster based on those GO annotations with p-values of less than 0.001.

Table 4 presents the GO term assignments to the selected sample clusters. As can be seen from these tables, GO terms assigned to the clusters are also highly topically related.

6.5. Analysis of Abstracts by descriptive keywords

Table 5 shows the twenty most significant words extracted from the articles in clusters as described in subsection 5.1. These words are highly topically relevant to the main topics considered for all but one cluster.

7. Summary

The relational structure of biological data has heretofore mainly served to facilitate browsing. Here we have used the implied graph as a computational object, partitioned it using standard techniques, and thus produced clusters of biological objects. These clusters exhibit strong topicality, as measured by both quantitative and qualitative manual evaluations, and by concentration of keywords and protein classifications. Because computation can be done on a large scale, these clusters reveal relationships that manual traversal of the graph do not. Furthermore, we believe that treating the graph as a computational object has applications in addition to producing topical clusters– for example, to information retrieval and data mining. In our future work, we plan to investigate statistical relational learning algorithms to predict links between biological objects for knowledge discovery.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 9986085. We thank David Synder for evaluating the sample clusters, and for his helpful comments. We thank Rose Oughtred for helpful discussions.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI–BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, *25*, 3389–3402.

Bairoch, A., & Apweiler, R. (2000). The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, *28*, 45–48.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*, 235–242.

Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, *20*, 359–392.

Kirsten, M., Wrabel, S., & Horvath, T. (2000). Distance-based approaches to relational learning and clustering. In *Relational data mining*, 213–230. Springer-Verlag New York, Inc.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, *247*, 536–540.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, *25*, 25–29.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images*. San Francisco, CA: Morgan Kaufmann. 2 edition.

Chemotaxis cluster 1072					
go id	**go name**	**a**	**n**	**A**	**-log(p_value)**
GO:0007600	sensory perception (BP)	52	83	412	243.02
GO:0006935	chemotaxis (BP)	33	83	300	144.67
GO:0030435	sporulation (BP)	19	83	404	66.11

Apoptosis cluster 1670					
go id	**go name**	**a**	**n**	**A**	**-log(p_value)**
GO:0006915	apoptosis (BP)	65	88	318	337.88
GO:0008234	cysteine-type peptidase (MF)	40	88	462	164.63
GO:0016787	hydrolase (MF)	43	88	10857	49.19
GO:0005634	nucleus (CC)	13	88	5856	8.04

Insulin cluster 1473					
go id	**go name**	**a**	**n**	**A**	**-log(p_value)**
GO:0019838	growth factor binding (MF)	31	40	48	223.3
GO:0005179	hormone (MF)	8	40	1082	19.81
GO:0005180	peptide hormone (MF)	3	40	276	9.1

Table 4. GO assignments of the sample clusters. In the GO term column, the GO annotation types are shown in parentheses: BP, MF and CC stand for biological_process, molecular_function and cellular_component, respectively. a is the number of the particular category SWISS-PROT sequences in the cluster, n denotes the number of SWISS-PROT sequences in the cluster, and A is the actual number of the particular GO category SWISS-PROT sequences.

Topic	**Cluster**	**Descriptive words extracted from the MEDLINE articles in clusters**
calmodulin	1794	calmodulin n-cam calmodulin-dependent cam-dependent cams cam-binding adhesion ca ca2 cabp neural brain ng-cam kinase calcium domain calcium-binding chicken calcium-dependent molecule
rhodopsin	1402	bacteriorhodopsin retinal schiff rhodopsins chromophore light-driven halorhodopsin proton pump halobacterium photocycle pharaonis asp85 transmembrane light phototaxis opsin visual pumping retinal-binding
spliceosome	1634	splicing snrnp u1 u2 ribonucleoprotein spliceosome sr nuclear sf2 asf rna pre-mrnas rna-binding u2af factor rs alternative factors splice spliceosomal
chemotaxis	1072	chemotaxis chea cheb chew cher response regulator bacterial swimming chez flagellar phosphorylation phosphotransfer chemotactic salmonella typhimurium swarm transduction flim crystal
apoptosis	1670	apoptosis death apoptotic fadd cd95 programmed necrosis apo-1 fas-mediated fasl flice mort1 cell daxx fas-induced effector tumor cells death-inducing signaling
ubiquitin	1665	polyubiquitin ubiquitin-specific ubiquitin-like deubiquitinating ubp ubi ubiquitin-dependent extension ubiquitins conjugates fusion ubiquitin-beta-galactosidase ubiquitin-encoding degradation ribosomal nedd8 repeats ubiquitin-activating tetraubiquitin pr
insulin	1473	insulin-like igfbps igfbp igfbp-1 diabetes igf growth autophosphorylation igfs factor-binding receptor igfbp-2 insulin-stimulated igfbp-3 mellitus igfbp-5 igfbp-4 igf-i factor igf-binding
p53	1674	tumor suppressor cancer p53-dependent tumors p53-binding p53-mediated cancers lines li-fraumeni carcinomas tumor-suppressor mutations cell human tp53 tumour damage breast carcinoma

Table 5. The most significant words extracted from the MEDLINE articles in each cluster; sorted in ascending order of p-value. The extracted words within each sample cluster are highly topically related.