Partitioning a Graph of Sequences, Structures and Abstracts for Information Retrieval[1]

Aynur Dayanik
Computer Science
Rutgers University
aynur@cs.rutgers.edu

Craig G. Nevill-Manning
Google, Inc.
craignm@google.com

Rose Oughtred
Chemistry and Chemical Biology
Rutgers University
rose@rcsb.rutgers.edu

Abstract

In this paper, we consider the problem of finding the MEDLINE articles that describe functions of particular genes. We describe our experiments using the `mg` system and the partitioning of a graph of biological sequences, structures and abstracts. We participated in the primary task of the TREC 2003 Genomics Track.

1 Introduction

Computational biology deals with a wide range of entities, including DNA sequences, protein sequences, protein structures, gene functions and academic publications. Databases of these entities include explicit links between them. For example, MEDLINE abstracts often reference sequences and structures that are relevant to the publication. Sequences are related to the structures of the proteins that they encode. Sequences are also related to homologous sequences in other organisms. In this paper, we explore how these relationships can be used to enhance retrieval of relevant MEDLINE abstracts. Our intuition is this: two abstracts may not share a significant number of common terms, but if they are both connected to many common sequences and structures, then a researcher interested in one abstract should be alerted to the existence of the other.

An assumption of our work is that recall is more important than precision. In the context of a professional investigation, researchers are willing to spend more time evaluating possible relevant literature than, say, the average web searcher is willing to spend on evaluating pages returned from a casual search. In other words, the cost of a false negative is much higher than the cost of a false positive.

To test this idea, we create a graph of biological entities, where edges are defined by the explicit links between them. We then partition the graph to find clusters of topologically related entities, including abstracts. Finally, these clusters are used to adjust the ranking of abstracts returned by a simple text retrieval engine.

[1]This material is based upon work supported by the National Science Foundation under Grant No. 9986085.

The paper is organized as follows. The next section describes the primary task of the Genomics track. Section 3 describes the biological databases we used. In section 4, we describe the construction of the graph from the databases, and then present our graph partitioning approach in section 5. In section 6, we describe the details of the official runs. In section 7, we discuss our results. Finally, we summarize and discuss possible directions for our future work.

2 Primary Task

The primary task of the genomics track was an ad hoc information retrieval problem: For a given gene X, find all MEDLINE articles that focus on the basic biology of the gene X or its protein products. Basic biology includes isolation, structure, genetics and function of genes/proteins in normal and disease states [9]. The relevance judgements were obtained from the LocusLink database [11]. A portion of a sample LocusLink entry is shown in Table 1. A LocusLink entry contains references to MEDLINE database as well as brief descriptions of gene functions extracted from the MEDLINE articles. These references are called GeneRIF (Gene References Into Function). In the example shown in Table 1, the unique PubMed identifier 12482586 and the description attached to it define a GeneRIF in the LocusLink database for the gene *EIF4E* of the organism *Homo sapiens*. GeneRIFs were used as query relevants – *qrels* in the TREC terminology.

LOCUSID: 1977
ORGANISM: Homo sapiens
OFFICIAL_SYMBOL: EIF4E
OFFICIAL_GENE_NAME: eukaryotic translation initiation factor 4E
ALIAS_SYMBOL: EIF-4E
...
GRIF: 12482586—eIF4E is associated with 4E-BP3 in the cell nucleus and cytoplasm
GRIF: 11959093—Mutations in the S4-H2 loop of eIF4E which increase the affinity for m7GTP
...

Table 1: A sample LocusLink entry

The provided MEDLINE collection consisted of 525,938 MEDLINE abstracts, indexed between April 1, 2002 and April 1, 2003. The training and test queries were obtained from LocusLink entries, and consisted of 50 queries each. Each query was specific to a gene, and chosen from the LocusLink database with gene identifiers such as *official gene name, official symbol, alias symbol, organism, etc.* For an overview of the track, see [7].

3 Data Sources

We used several genomics resources in addition to the provided MEDLINE collection. In this section, we will briefly describe the data sources we used to construct a graph.

MEDLINE: MEDLINE is a digital collection of life science literature consisting of over twelve million abstracts together with some additional information associated with each abstract such as manually assigned MeSH terms and chemical names. Moreover, there are links from MEDLINE abstracts to the sequences and structures that the article discuss.

GenBank: GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. Bibliographic references to MEDLINE articles are included for all published sequences.

Swiss-Prot: The Swiss Protein Database (Swiss-Prot) is a curated protein sequence database [4]. The Swiss-Prot entries are cross-referenced to several other databases, including MEDLINE, PROSITE and the PDB.

PDB and Ligands: The Protein Data Bank (PDB) contains 3–D structural data of biological macromolecules (proteins and nucleic acids) [5]. The PDB entries also contain references to small molecules, known as ligands. We used PDB structures as well as ligands to connect PDB structures based on their binding patterns. The PDB entries are also cross-referenced to the primary citations in MEDLINE and other databases including ENZYME and Swiss-Prot.

SCOP: The SCOP database provides a hierarchical classification of all proteins whose structure is known, including all entries in the PDB [10]. We represented the leaves of the SCOP hierarchy in our graph to be able to connect to the PDB structures.

PROSITE: PROSITE is a database of protein families and domains [6]. It consists of biologically significant sites, patterns and profiles. PROSITE provides cross-references to Swiss-Prot, PDB and ENZYME databases.

ENZYME: ENZYME is a repository of information relative to the nomenclature of enzymes [3]. ENZYME provides explicit links to Swiss-Prot and the PDB.

4 Constructing the Graph

We construct a weighted undirected graph where nodes correspond to entries from the databases listed in Section 3, including MEDLINE abstracts, DNA and protein sequences from GenBank and SwissProt, structures from PDB, patterns from PROSITE, classifications from SCOP and ENZYME, and chemical names from MEDLINE abstracts. Edges correspond to explicit links between entries encoded in the databses, e.g. the sequence annotations in MEDLINE abstracts.

Some nodes in the graph have high degree. Papers that describe genomic sequencing, for example, often reference all the sequences produced by the project. In these cases, the individual edges are not highly significant – the relationship between sequences from the same organism is not strong. So we assign these edges low weight. We assign weights to edges as follows. Let (u, v) be an edge in the graph. Then, the weight of the edge (u, v), $w(u, v)$, is computed as

$$w(u,v) = \min\left(\frac{1}{n(u,v)}, \frac{1}{n(v,u)}\right) \tag{1}$$

where $n(u, v)$ is defined as the number of edges between u and all other vertices of same type as vertex v, and $n(v, u)$ is defined as the number of edges between v and all other vertices of same type as vertex u.

As an example, consider a relation between the article A_1 and the sequence S_1. Let us say A_1 is related to 10 sequences in total, and S_1 is one of them. Also, assume that S_1 is related to 4 articles in total. Therefore, we will have an edge (u, v) corresponding to the relation between the article A_1 and the sequence S_1 in our graph with weight $1/10$ since $\min(1/10, 1/4) = 1/10$. Note that, for this particular example, when we compute the weight, we take into account only article-sequence relationships. In general, we consider the same type of relationships to compute edge-weights. We think that normalization is a fair method for assigning edge-weights because some objects are related to too many other objects of the same type and some only to a few.

5 Graph Partitioning

The objective of graph partitioning is to partition vertices of a graph in k equal subsets such that the total weight of edges connecting different subsets is minimized, thereby each subset is highly similar. The graph partitioning problem is NP-complete, but good heuristics exist. We employed a partitioning approach based on multilevel recursive bisection [8]. First, the size of the graph is reduced by collapsing nodes and edges to a few thousand nodes. Then the smaller graph is partitioned into two parts. Partitioning is repeated by uncoarsening each part one level up until all k subsets are obtained. We used publicly available graph partitioning software, METIS [1].

6 Run Descriptions

We employed the `mg` system as our retrieval engine [2, 12]. We submitted two official runs to the Genomics track. The first run was done using only the `mg` system while the second run was

obtained by reordering the retrieved abstracts by mg using the clusters of abstracts defined by the graph partitions.

Run using mg: We indexed the following sections from the MEDLINE abstracts: title, abstract, MeSH terms and chemical names. We slightly modified mg to be able to tokenize biological terms properly. It currently forms words as a sequence of letters, digits and the following special characters: (,), [,], ', -, ',' and /. Note that these special characters cannot be the first or last character of the word. It is clear that these special characters appear in gene names and synonyms. For example, 1,25-dihydroxyvitamin, dead/h and cyclin-dependent are now treated as single indexing terms.

mg performed case-folding but not stemming. Query parsing was done identically to document parsing. We formed queries from the gene names and synonyms. We eliminated duplicate words and stopwords from the queries. The mg system includes support for ranked queries, where similarity is evaluated using the cosine measure. We issued ranked queries.

Run using mg but ordering results by clusters: The graph is disconnected, with one large graph of about 500,000 nodes, and 224,440 smaller graphs, the largest of which has 1,500 nodes. The smaller graphs are considered single clusters. The large graph is partitioned to produce 5000 clusters of about 100 nodes each.

In this run, we first obtained the search results using mg, and then grouped them by clusters. We assigned each group the highest mg score in that cluster. We ordered the groups first by group scores, and then in each group, by the mg scores. The intuitive idea of reranking mg search results is that if it can identify some qrels at the top ranked results, we can push more qrels to the top results by using the additional information about their relatedness, i.e., being in the same cluster.

7 Results

This section reports our results obtained using the mg system and the clusters. The orginal mg achieved a mean average precision (MAP) of 0.2759 whereas the modified mg achieved a MAP of 0.3054 on the training data. Since the modified version increased the performance in terms of MAP, all the results reported for the test data were obtained by the modified version of mg. While mg achieved a MAP of 0.3054 on the training data, ranking mg search results by clusters achieved a MAP of 0.3191. However, the mean average precisions for mg and using clusters are 0.1652 and 16.36, respectively, on the test data.

Figure 1 compares mg for the top 1000 retrieved articles to the best and median systems using the test data. In general, our performance is close to that of the median systems. We

tried to understand the possible reasons for this low performance, and noticed that retrieval is quite sensitive to query formulations. For example, for test topic 7, the mg system identified *only* one qrel as relevant out of four known qrels for this topic. In Figure 1, mg exhibits very low performance for topic 7 compared to the other participating systems. The reason is that the query for this topic includes "syndecan 4", however, three qrels contain only "syndecan-4", and only one qrel contains both "syndecan" and "syndecan-4". We indexed "syndecan-4" as a single word for those three qrels, and therefore mg cannot locate them when the query is "syndecan 4". In fact, when we change the query to include only "syndecan-4" (even omitting gene symbols), mg identified all four qrels as the $4th$, $5th$, $7th$ and $8th$ documents. Another example is that "1,25-dihydroxyvitamin" appears in MEDLINE abstracts, whereas test topic 12 expresses it with an additional space, i.e. as"1,25- dihydroxyvitamin", thereby breaking it up into two words. Therefore, different variations of gene names and symbols are an important issue to be considered when preparing queries.

Figure 1 also compares mg and ranking using clusters for the top 100 retrieved articles using the test data. As can be seen from the figure, ranking by clusters significantly improved the MAP for eight queries, but significantly decreased for the other eight queries. However, upon manual inspection, we find out that many qrels fall into same clusters. We believe that clusters can be useful for researchers by itself or in conjunction with a retrieval method to support browsing similar entities.

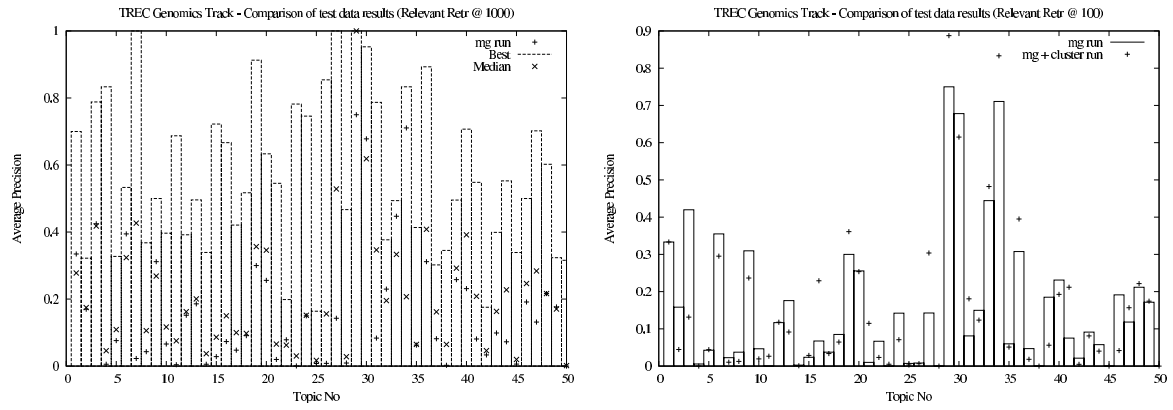


Figure 1: Comparison of mg for the top 1000 retrieved articles to the best and median systems using the test data (left), and Comparison of mg and clustering results for the top 100 retrieved articles using the test data (right)

We analyzed one cluster to assess its quality. Figure 2 shows the qrels and their corresponding clusters for test topic 3. Test topic 3 have thirteen qrels, and eleven of them fall into

same cluster (the other two that fall into two other separate clusters are PMID 12167712 and PMID 12186496). We picked the cluster containing these eleven qrels. Figure 3 and 4 present the Chemical names, PDB and GenBank entries in this cluster. These figures aslo show the descriptive words extracted automatically from the MEDLINE abstracts in the cluster based on the word frequencies in the cluster and in the entire corpus of MEDLINE articles.

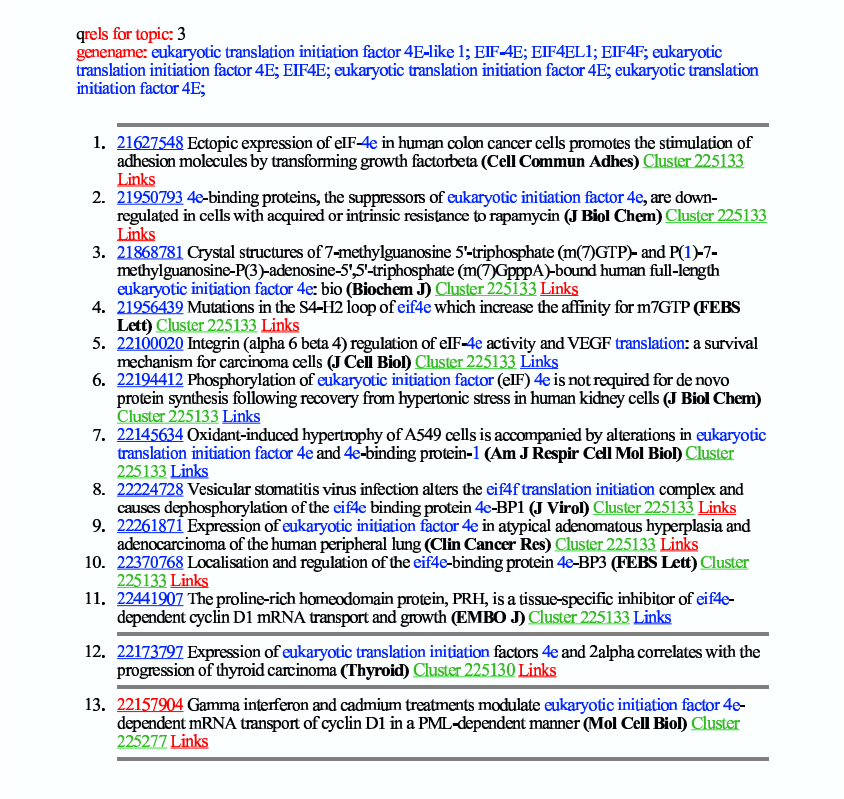
Our scientific domain expert carefully analyzed this cluster and found it to be highly relevant to the gene of interest, eukaryotic initiation factor 4e (eif4e). The descriptive words are meaningful with respect to the eif4e gene and to the biological pathway within which the eif4e protein is involved. The chemical names are also relevant to the mechanism of eif4e and all of the associated PDB macromolecular structures contain the eif4e protein. In addition, approximately 60% of the Swiss-Prot and GenBank sequences in this cluster are relevant to one another in that they all play a role in the biological process of protein synthesis.

Let us summarize the expert's analysis of this cluster:

Quality of cluster: Very high
Descriptive words are meaningful
Relevancy: Very good
Chemicals: All relevant
PDB structures: All relevant
GenBank: 8 out of 13 relevant
Swiss-Prot: 3 out of 5 directly relevant

8 Conclusions

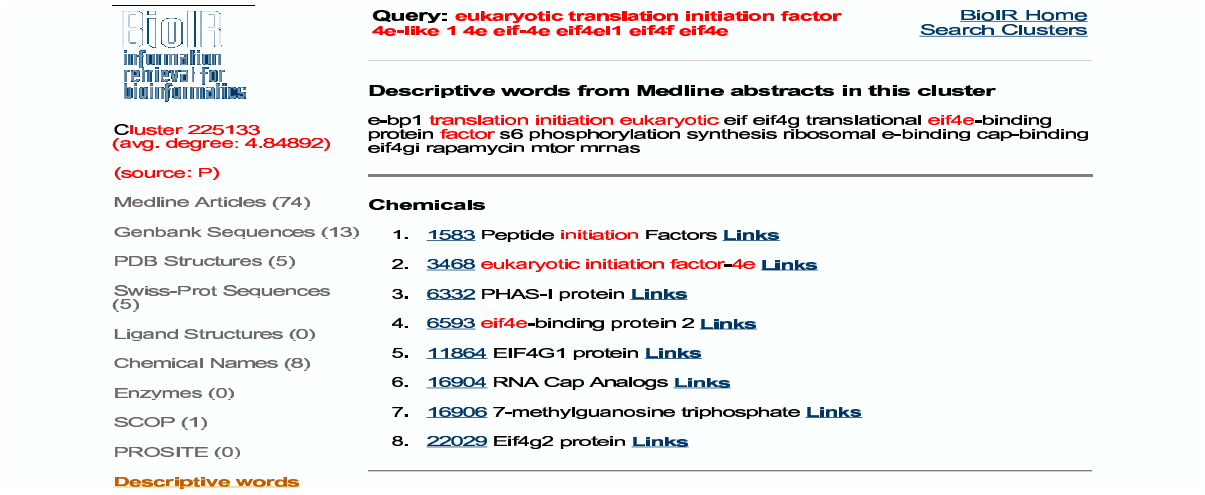
Our primary goal was to demonstrate to the Information Retrieval and Bioinformatics communities experiments that involved the open-source `mg` system and graph partitioning for an information retrieval problem in genomics. Our results are close to the median performance of the participating systems in the Genomics track. Even though we observed some high-quality clusters, we did not obtain a significant improvement over `mg` alone using our clusters for retrieval purposes. In the future, we plan to carry out more experiments in order to better understand the quality of the clusters. Moreover, we want to try other retrieval methods together with our clusters in order to determine how the initial retrieval method affects the performance of the retrieval using clusters. We also want to investigate the effects of creating clusters during the retrieval process from the neighborhood graphs of retrieved abstracts.



qrels for topic: 3
genename: eukaryotic translation initiation factor 4E-like 1; EIF-4E; EIF4EL1; EIF4F; eukaryotic translation initiation factor 4E; EIF4E; eukaryotic translation initiation factor 4E; eukaryotic translation initiation factor 4E;

1. 21627548 Ectopic expression of eIF-4e in human colon cancer cells promotes the stimulation of adhesion molecules by transforming growth factorbeta **(Cell Commun Adhes)** Cluster 225133 Links
2. 21950793 4e-binding proteins, the suppressors of eukaryotic initiation factor 4e, are down-regulated in cells with acquired or intrinsic resistance to rapamycin **(J Biol Chem)** Cluster 225133 Links
3. 21868781 Crystal structures of 7-methylguanosine 5'-triphosphate (m(7)GTP)- and P(1)-7-methylguanosine-P(3)-adenosine-5',5'-triphosphate (m(7)GpppA)-bound human full-length eukaryotic initiation factor 4e: bio **(Biochem J)** Cluster 225133 Links
4. 21956439 Mutations in the S4-H2 loop of eif4e which increase the affinity for m7GTP **(FEBS Lett)** Cluster 225133 Links
5. 22100020 Integrin (alpha 6 beta 4) regulation of eIF-4e activity and VEGF translation: a survival mechanism for carcinoma cells **(J Cell Biol)** Cluster 225133 Links
6. 22194412 Phosphorylation of eukaryotic initiation factor (eIF) 4e is not required for de novo protein synthesis following recovery from hypertonic stress in human kidney cells **(J Biol Chem)** Cluster 225133 Links
7. 22145634 Oxidant-induced hypertrophy of A549 cells is accompanied by alterations in eukaryotic translation initiation factor 4e and 4e-binding protein-1 **(Am J Respir Cell Mol Biol)** Cluster 225133 Links
8. 22224728 Vesicular stomatitis virus infection alters the eif4f translation initiation complex and causes dephosphorylation of the eif4e binding protein 4e-BP1 **(J Virol)** Cluster 225133 Links
9. 22261871 Expression of eukaryotic initiation factor 4e in atypical adenomatous hyperplasia and adenocarcinoma of the human peripheral lung **(Clin Cancer Res)** Cluster 225133 Links
10. 22370768 Localisation and regulation of the eif4e-binding protein 4e-BP3 **(FEBS Lett)** Cluster 225133 Links
11. 22441907 The proline-rich homeodomain protein, PRH, is a tissue-specific inhibitor of eif4e-dependent cyclin D1 mRNA transport and growth **(EMBO J)** Cluster 225133 Links
12. 22173797 Expression of eukaryotic translation initiation factors 4e and 2alpha correlates with the progression of thyroid carcinoma **(Thyroid)** Cluster 225130 Links
13. 22157904 Gamma interferon and cadmium treatments modulate eukaryotic initiation factor 4e-dependent mRNA transport of cyclin D1 in a PML-dependent manner **(Mol Cell Biol)** Cluster 225277 Links

Figure 2: Test topic 3 qrels and their clusters – A screen snapshot from our BioIR system



BioIR
information retrieval for bioinformatics

Query: eukaryotic translation initiation factor 4e-like 1 4e eif-4e eif4el1 eif4f eif4e

BioIR Home
Search Clusters

Cluster 225133
(avg. degree: 4.84892)
(source: P)

Medline Articles (74)
Genbank Sequences (13)
PDB Structures (5)
Swiss-Prot Sequences (5)
Ligand Structures (0)
Chemical Names (8)
Enzymes (0)
SCOP (1)
PROSITE (0)
Descriptive words

Descriptive words from Medline abstracts in this cluster

e-bp1 translation initiation eukaryotic eif eif4g translational eif4e-binding protein factor s6 phosphorylation synthesis ribosomal e-binding cap-binding eif4gi rapamycin mtor mrnas

Chemicals

1. 1583 Peptide initiation Factors Links
2. 3468 eukaryotic initiation factor-4e Links
3. 6332 PHAS-I protein Links
4. 6593 eif4e-binding protein 2 Links
5. 11864 EIF4G1 protein Links
6. 16904 RNA Cap Analogs Links
7. 16906 7-methylguanosine triphosphate Links
8. 22029 Eif4g2 protein Links

Figure 3: Chemical names assigned to the cluster having eleven qrels out of thirteen qrels for test topic 3 – A screen snapshot from our BioIR system

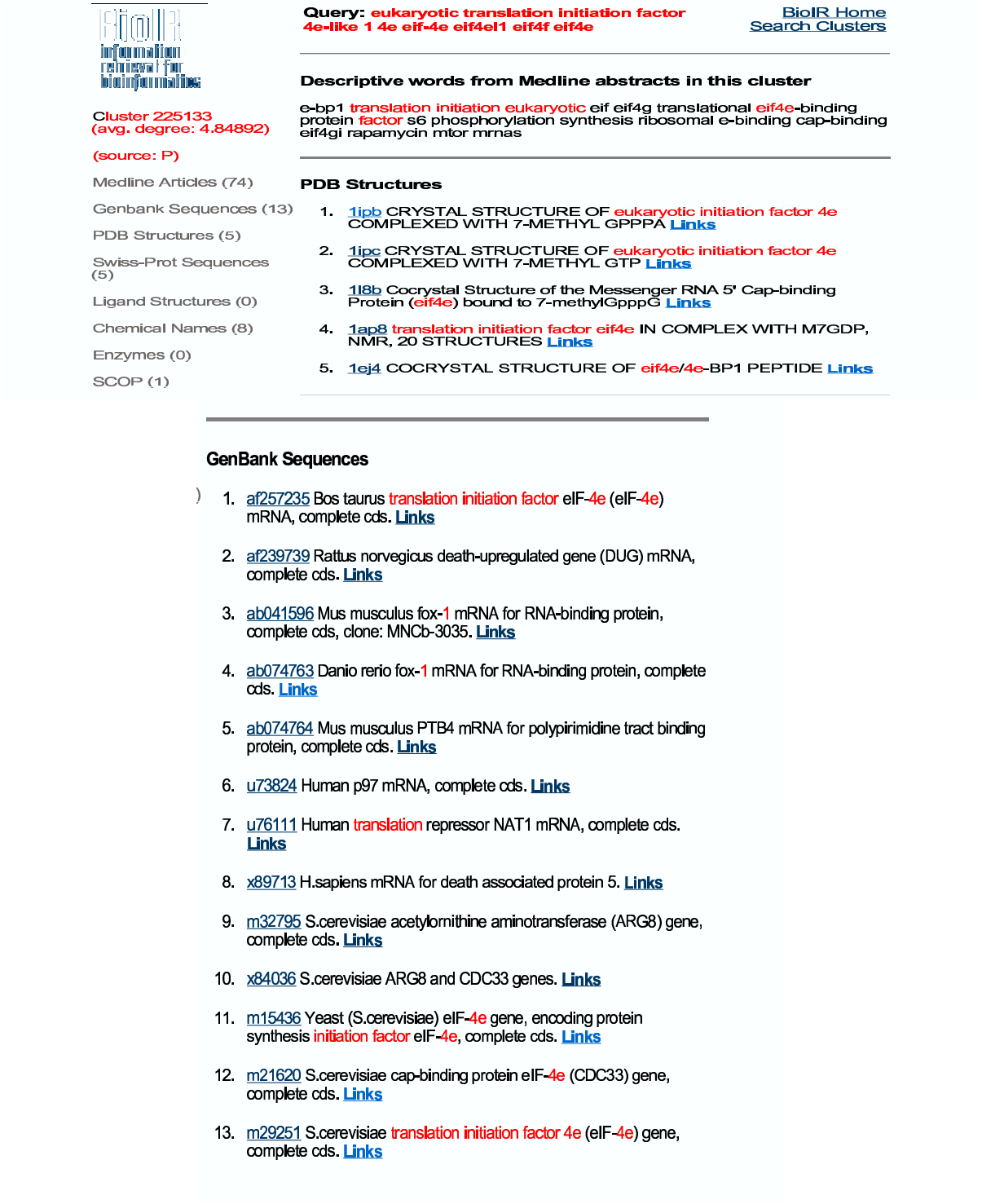


Figure 4: PDB and GenBank entries in the cluster having eleven qrels out of thirteen qrels for test topic 3 – A screen snapshot from our BioIR system

References

[1] METIS: Family of Multilevel Partitioning Algorithms. http://www-users.cs.umn.edu/~karypis/metis/.

[2] MG software. http://www.cs.mu.oz.au/mg/.

[3] A. Bairoch. The ENZYME database in 2000. *Nucleic Acids Res*, 28:304–305, 2000.

[4] A. Bairoch and R. Apweiler. The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, 28:45–48, 2000.

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

[6] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J Sigrist, K. Hofmann, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleic Acids Res*, 30:235–238, 2002.

[7] William Hersh and Ravi Teja Bhupatiraju. TREC Genomics Track Overview. In *Proceedings of the Twelfth Text Retrieval Conference, TREC-12*, Gaithersburg, MD, 2003.

[8] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

[9] J. A. Mitchell, A. R. Aronson, J. G. Mork, L. C. Folk, S.M. Humphrey, and J.M. Ward. Gene indexing: Characterization and analysis of NLM's GeneRIFs. In *Proceedings of the AMIA Annual Symposium*, pages 460–464, 2003.

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

[11] K.D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, 29(1):137–140, 2001.

[12] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, CA, 2 edition, 1999.