Use of Figures in Literature Mining for Biomedical Digital Libraries

Nawei Chen, Hagit Shatkay, Dorothea Blostein
School of Computing, Queen's University, Kingston, Ontario, Canada
chenn, shatkay, blostein@cs.queensu.ca

Abstract

The maintenance of biomedical digital libraries (including organism databases and protein databases) involves analysis of a large number of documents. Much work is done manually: curators study large numbers of biomedical documents while updating and annotating organism databases such as MGI (Mouse Genome Informatics) and Flybase (a database of the fruit-fly genome). We summarize the annotation process in organism databases, and describe some of the roles played by the Gene Ontology and by document databases such as PubMed. Efforts are ongoing to automate parts of the annotation process. Biomedical text mining contests, such as the TREC Genomics Track [6, 7], define annotation subtasks, and provide training and test data. So far, these efforts have focused on the analysis of the text content of documents. We are investigating the analysis of figures in biomedical documents; the information derived from figure analysis may later be combined with the information derived from text analysis. We present an algorithm for using figures in document triage; triage involves determining which documents are relevant to a given annotation task. In our triage algorithm, we segment figures into subfigures and classify the subfigures as Graphical, Gel, Fluorescence Microscopy, and Other Microscopy. A secondary classification into subcategories is performed by clustering, using clusters created from the subfigures in the labeled training data. The classifications of all subfigures in a document are combined to form a document descriptor. The document descriptor is then classified using a Naïve Bayes Classifier, as either relevant or irrelevant to the given annotation task.

1. Introduction

Much current work in biomedical literature mining aims to extract information and discover knowledge for populating biomedical digital libraries. As demonstrated in several recent surveys of biomedical text mining [1, 2, 3, 4], most biomedical literature mining methods focus on the analysis of text, typically from abstracts or keywords. Abstracts are available from PubMed[1]; keywords are often taken from MeSH terms[2]. The abstracts and keywords are used to perform various tasks, including document classification, named entity tagging (e.g. identifying protein/gene names), and information extraction (e.g. extracting interactions between proteins). However, much of the biological information contained in articles is not present in abstracts or keywords but rather in the body of the article. With the increasing availability of full-text documents, there is a trend toward using full-text documents for biomedical literature mining. We believe this raises new opportunities for the document image analysis research community. Figures and document layout structure in full-text documents may be used to improve literature mining.

In this paper, we focus on efforts to automate the annotation process in biomedical organism databases. Biomedical organism databases are important research tools for biologists. Annotation, in this context, means assigning attributes to biological entities in the databases, based on evidence that is found in biomedical publications and in other resources. Currently, annotation is performed by human curators; a daunting task given the tremendous increase in the number of biomedical publications over the past few years. Efforts to automate parts of the annotation process are ongoing. So far, research efforts have focused only on the analysis of the text itself [6, 8]. We are investigating a new direction, namely the analysis of figures in biomedical documents; we believe that the information derived

[1] PubMed, see appendix A.2. Appendix figures A.1 and A.2 illustrate the keywords and abstract information available for a sample abstract from PubMed.

[2] MeSH, see appendix A.3.1.

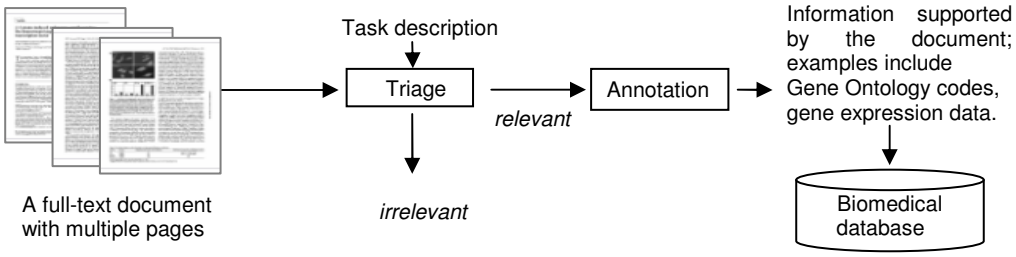

Figure 1. Document triage and annotation in a biomedical database. Document triage decides whether a document is relevant for annotation. Triage is task specific. Annotation extracts different types of information in different tasks. The supported information plus the link to the original document are stored into a biomedical database.

from figure analysis can later be combined with the information derived from text analysis to produce more complete and accurate results.

2. Overview of Organism Databases and Their Annotations

An organism database maintains information about biological entities (such as genes and proteins) associated with a specific model organism that is widely studied and used in biology and medicine. One such organism is the mouse; mouse genome information is stored in the MGI (the Mouse Genome Informatics) databases[3]. The fruit-fly is another such organism; FlyBase[4] stores the genomic data of the fruit-fly. Other more general and central databases have also been created; an example is the protein database Swiss-Prot[5]. There is an unprecedented growth in the amount of information stored in such databases, due to high-throughput biological techniques, such as whole-genome sequencing, and large-scale gene expression analysis.

2.1. Annotation of organism databases

The information in organism databases is typically focused on genomic and proteomic facts and hypotheses, concerning the genome and the genes it contains, as well as the proteins derived from these genes. Every piece of information, about a gene/protein becomes part of its *annotation*. Biomedical publications are often used as evidence for the annotation. For instance, there is often a reference to the publication in which the *gene sequence* was first shown, (where the *gene sequence* is the string of nucleotides {A, C, G, T} constituting the gene). Similarly there are references to the publication(s) in which the protein product(s) for the gene was first shown. Other aspects, such as the gene function and the subcellular localization of its protein, are also annotated and supported by references to the relevant biomedical literature.

A major activity of many model organism database projects is the annotation of genes and proteins using standardized codes for process and function from the Gene Ontology (GO)[6]. The Gene Ontology provides a controlled vocabulary that allows genes and gene products to be described in terms of their *molecular function*, *biological process*, and *cellular component*. Using this agreed-upon vocabulary within and across different biomedical databases enables more effective database searches by both computers and researchers.

As illustrated in Figure 1, annotation is task-specific. For example, one annotation task involves the assignment of GO codes to genes, while another involves the association of genes with their expression information. Annotation is preceded by *document triage*, which identifies the subset of documents that may contain evidence to support the annotation: these documents typically discuss experimental findings related to a particular gene or gene product.

[3] MGI, see appendix A.1.1.
[4] FlyBase, see appendix A.1.2.
[5] Swiss-Prot, http://www.ebi.ac.uk/swissprot

[6] See appendix A.3.2. for more discussion of the Gene Ontology and GO evidence codes.

2.2. Examples of GO annotation in MGI

In this section, we demonstrate how an article is used as a reference for GO annotations in MGI. This GO annotation task was simulated in the TREC Genomics Track 2004 [6]. Figure 2 shows the abstract of a reference article [13] with PubMed identifier 12235125 in the MGI database. Four genes are annotated based on this article: *Ctnnb1, Mitf, Myc* and *Tcf7*. The details of one of the genes, *Tcf7* are shown in Figure 3. *Tcf7* is annotated with GO codes with respect to the three GO hierarchies: *biological process*, *cellular component* and *molecular function*. This gene has a total of 9 GO annotations, shown in Figure 4. For each annotation, an evidence code and a reference code are assigned to justify the annotation. A GO evidence code describes the nature of the evidence that supports this attribution. For example, the evidence code IDA stands for *"inferred from direct assay"*.

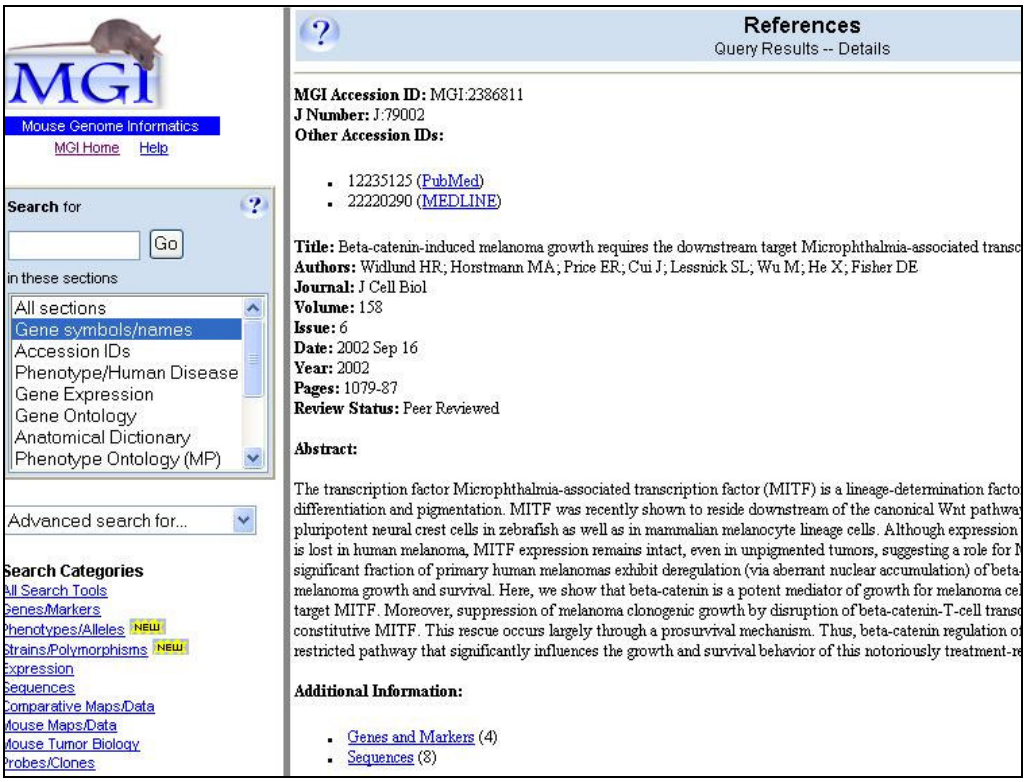

Figure 2. A sample document [13] viewed through the MGI database. The top of the screen shot shows four different identifiers for this document, including PubMed ID 122351215 and J number J:79002. Next, title and abstract are shown. The Additional Information at the bottom shows that this document provides evidence for annotation of 4 genes and 8 sequences. Details for one of these genes are shown in Figure 3. The screen shot was taken directly from the MGI's web site (http://www.informatics.jax.org/) in December 2005.

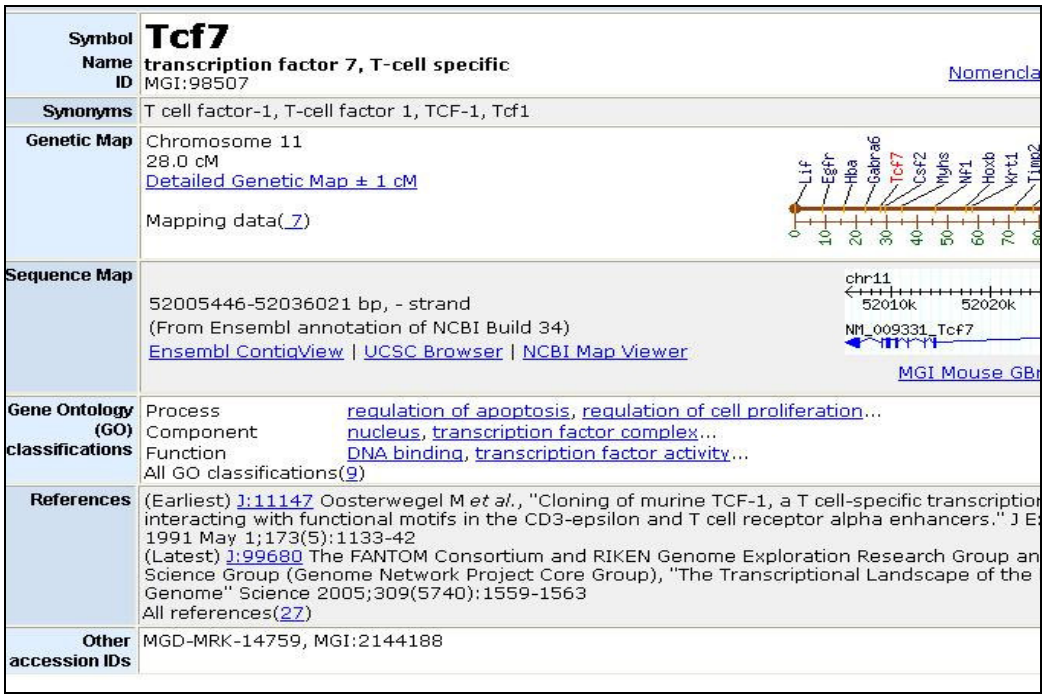

Figure 3. Details for gene Tcf7, one of the four genes mentioned in Figure 2. The screen shot shows some of the attributes of gene Tcf7, including synonyms, genetic map, sequence map, Gene Ontology (GO) classifications, references and other accession IDs. There are 9 GO annotations in total. Details are shown in Figure 4. The screen shot was taken directly from the MGI's web site (http://www.informatics.jax.org/) in December 2005.

Category	Classification Term	Evidence	Inferred From	Ref(s)
Biological Process	regulation of apoptosis	IDA		J:79002
Biological Process	regulation of cell proliferation	IDA		J:79002
Biological Process	regulation of transcription, DNA-dependent	IDA		J:79002
Biological Process	transcription	IEA		J:60000
Biological Process	Wnt receptor signaling pathway	IDA		J:79002
Cellular Component	nucleus	IEA		J:60000
Cellular Component	transcription factor complex	IEA		J:56000
Molecular Function	DNA binding	IDA		J:79002
Molecular Function	transcription factor activity	IDA		J:79002

Figure 4. GO annotations for gene Tcf7. The screen shot taken directly from the MGI's web site (http://www.informatics.jax.org/) shows that each GO annotation has a GO classification term, an evidence code, and reference IDs. Six of the nine GO annotations shown in this figure use the reference "J:79002". This refers to the sample document shown in Figure 2. Evidence code "IDA" stands for "inferred from direct assay", and "IEA" stands for "inferred from electronic annotation".

2.3. Ongoing efforts to automate annotation in organism databases

Annotation related to functional descriptions of genes and gene products is typically extracted manually: *curators* are employed to examine biomedical documents and find evidence to assign attributes to genes and gene products. This is a slow and labor-intensive process. Moreover, it is becoming more and more difficult to keep up with the increasing number of entities in the database (due to high-throughput experimental results) and the large number of biomedical publications that are to be reviewed for annotation. Therefore, efforts are underway to automate parts of the triage and the annotation procedures. While much work has been done on biomedical text mining during the past few

years, the utility of such new systems is unclear. The objective evaluation was very hard for early systems as there were no objective benchmarks. This issue was recently addressed by several evaluation efforts. The most notable evaluation contests to date include KDD Cup [5], TREC Genomics 2003-2005 [6, 7] and the BioCreAtIvE challenge [8]. In these contests, annotation subtasks were formally defined based on real tasks carried out by human curators. Training and test data sets labeled by human experts were provided, along with objective evaluation metrics. Text mining researchers participating in these contests applied a wide variety of techniques to a common problem. We believe that these tasks also provide an excellent research ground for document image analysis. In the following we briefly introduce the TREC Genomics Track. Our experiments, described in the next section, of using figures for biomedical literature mining are based on the subtasks that were defined in this track.

The TREC Genomics Track is an on-going contest sponsored by NIST, focused on information retrieval from biomedical text [6]. The 2004 track was its second year, and included an ad-hoc retrieval task and a categorization task. The categorization task had three sub-tasks, one was a document triage task and the other two were GO annotation tasks, defined to simulate the task performed by MGI curators [6]. The triage task is to classify a document as *relevant* or *irrelevant* for GO annotation, given a set of labeled documents as training data. The 2005 Genomics Track refined this task, adding three finer categories of information collected and catalogued by MGI [7]. The documents for the categorization task consisted of full-text articles from three journals over two years. The journals were *Journal of Biological Chemistry* (JBC), *Journal of Cell Biology* (JCB), and *Proceedings of the National Academy of Science* (PNAS). Articles from 2002 were designated as training data and those from 2003 as test data. The training and test sets consist of 5,837 and 6,043 articles, respectively. The true triage decisions were provided by MGI. While during the TREC contest the data was available to participants only, the data is now available to all by contacting NIST.

3. Using Figures for Document Triage

In this section, we present our work on using figures for the document triage task discussed above.

3.1. Importance of figures

Figures are often content rich and concisely summarize the most important results or methods used in a paper. Recently the importance of figure captions was noted for triage and annotation. Specifically, Regev et al. used figure captions for task 1 in the KDD Cup 2002 [3, 9] with notably good results. The task was to automate the document triage for FlyBase by identifying papers that contain information about gene expression in the fruit-fly. Regev et al. note that curators who manually review papers look primarily at the figures in the paper to ascertain the presence of experimental evidence. FlyBase curators have indeed mentioned that many of the experimental results are presented in figures and their captions [5]. Following this line of reasoning, Darwish and Madkour [10] have also used text extracted from figure captions for the triage task in TREC 2004.

Figures, specifically fluorescence images from biomedical articles, have been recently used to predict protein sub-cellular localization. Murphy et al [11] extracted figures from on-line biomedical journals and classified the segmented figure images into two classes: *fluorescence microscopy* images and *other* images. They used fluorescence microscopy images for protein sub-cellular prediction based on image features. This is also an image classification task, classifying an image into one of the sub-cellular localizations. Samples of fluorescence microscopy images from different sub-cellular locations are shown in Figure 5. By utilizing the associated captions, Murphy et al [11] aim to extract assertions such as “Figure N depicts a localization of type L for protein P in cell type C”.

As far as we know, as of yet, the figure images themselves have not been considered for triage and annotation. In this paper, we explore the possibility of using figures for the triage task. We plan to combine figure-based methods with text-based methods for triage, as we view these as complementary, rather than as competing approaches.

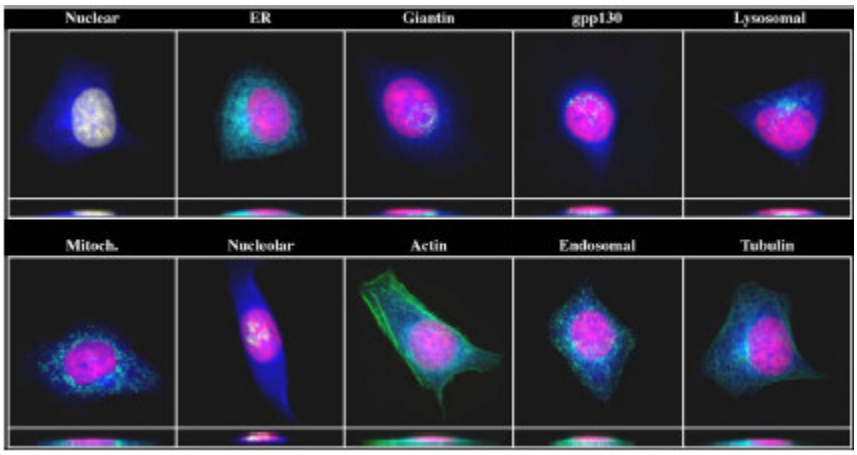

Figure 5. Typical images from ten defined classes based on sub-cellular localization of 3-D HeLa cell image collection (reproduced from [22] by permission of the author). We can see the images from different sub-cellular locations are visually distinct. Fluorescence microscopy images convey important information about the location and behavior of proteins, as well as many details of protein-protein interactions. They are important tools for cellular research.

3.2 Our method of using figures in document triage

Document triage can by viewed as a binary classification task. The input is a set of full-text documents. A document is classified as either *positive* (to be used in annotation) or *negative* (not to be used in annotation). To automate the task, we train a classifier using a set of labeled training documents (Section 3.2.1), and then apply the classifier to other documents (Section 3.2.2). Our basic idea is to create an image-based description for each document, and then apply a naïve Bayes classifier to these descriptions. This approach is adapted from Duygulu et al.'s work on image annotation [12]. Duygulu et al. described an image using a small vocabulary of "blobs", which are labels assigned to the clusters of all the segmented image regions in a collection of images.

3.2.1. Train the classifier. Our experiments use the triage data from TREC Genomics 2004, as described in Section 3.3. The training data consists of documents that have been labeled as either *positive* or *negative* by human experts. We further label the subfigures in the training data, as illustrated in Figure 6, and described in Step (3) below.

In our system, training documents are processed using six steps as follows:

(1) Figure extraction. The full-text documents are in XML format, obtained from TREC Genomics 2004. We extract captions and links to the figures from the XML documents, and then download the figure images themselves from the publisher's web site. A sample document [13] and an extracted figure are shown in Figure 6. We use 4,000 figures in training and testing as described in Section 3.3.

(2) Figure segmentation. As evident from Figure 6, each figure may consist of several subfigures. We segment each figure into its subfigures. A bottom-up segmentation approach, based on Connected Components (CCs) analysis [14], is used for this purpose. Sample figure segmentation results are shown in Figure 6. Unavoidably, errors do occur during the segmentation process.

The complexity found in figures is illustrated in Figure 7. Usually a figure has mixed types of subfigures and has no standard layout. Murphy et al. discussed lack of standards in figures of scientific publication and the difficulty of associating a sub-figure with corresponding captions [11].

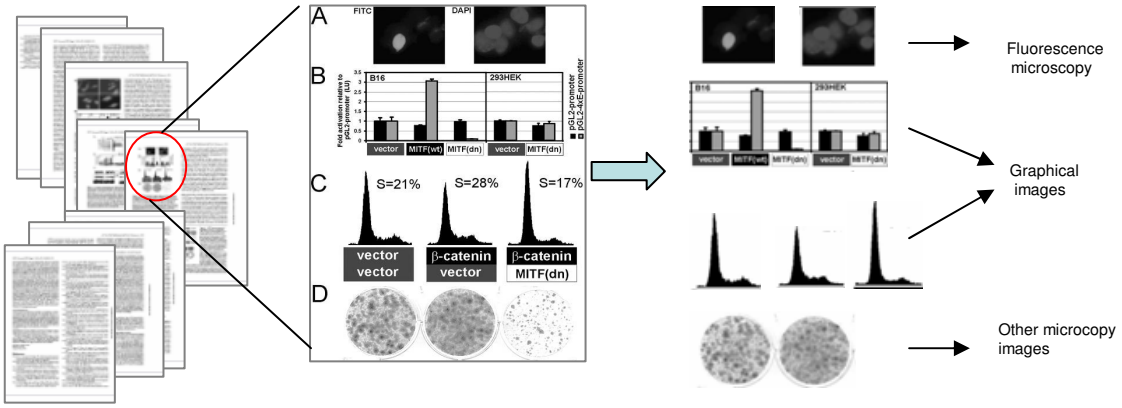

Figure 6. Preprocessing of figures in a sample input document. The document (PubMed Identifier 12235125 [13]) has nine pages and six figures. We extract all the figures from the document and save as image formats, such as JPEG, or GIF. One of the extracted figures is shown enlarged. Figure segmentation is based on Connected Components (CCs) analysis. Subfigures are extracted from each figure. The CCs whose bounding box areas are too small are discarded since they are most likely characters used to label figures. Subfigure classification uses a hierarchical classification scheme defined in Figure 8.

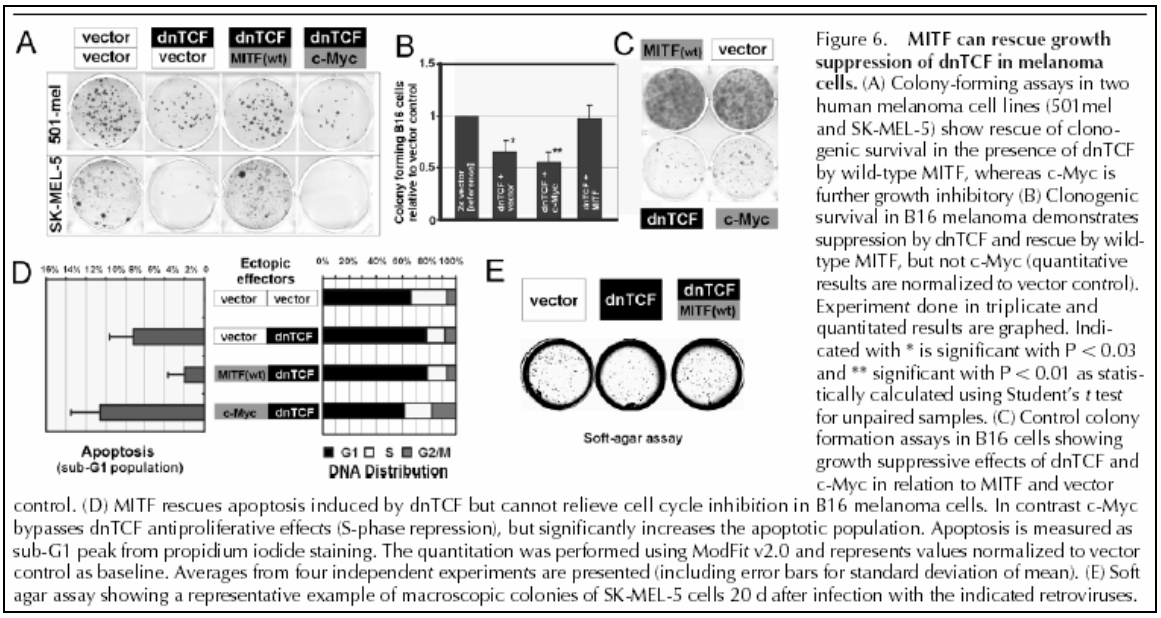

Figure 6. **MITF can rescue growth suppression of dnTCF in melanoma cells.** (A) Colony-forming assays in two human melanoma cell lines (501mel and SK-MEL-5) show rescue of clonogenic survival in the presence of dnTCF by wild-type MITF, whereas c-Myc is further growth inhibitory (B) Clonogenic survival in B16 melanoma demonstrates suppression by dnTCF and rescue by wild-type MITF, but not c-Myc (quantitative results are normalized to vector control). Experiment done in triplicate and quantitated results are graphed. Indicated with * is significant with $P < 0.03$ and ** significant with $P < 0.01$ as statistically calculated using Student's *t* test for unpaired samples. (C) Control colony formation assays in B16 cells showing growth suppressive effects of dnTCF and c-Myc in relation to MITF and vector control. (D) MITF rescues apoptosis induced by dnTCF but cannot relieve cell cycle inhibition in B16 melanoma cells. In contrast c-Myc bypasses dnTCF antiproliferative effects (S-phase repression), but significantly increases the apoptotic population. Apoptosis is measured as sub-G1 peak from propidium iodide staining. The quantitation was performed using ModFit v2.0 and represents values normalized to vector control as baseline. Averages from four independent experiments are presented (including error bars for standard deviation of mean). (E) Soft agar assay showing a representative example of macroscopic colonies of SK-MEL-5 cells 20 d after infection with the indicated retroviruses.

Figure 7. A sample figure and corresponding captions reproduced from the article [13] shown in Figure 6. The figure has mixed types of subfigures. There are challenges in extracting subfigures, and associating each subfigure with corresponding captions.

(3) Subfigure classification. We classify the subfigures using a hierarchical classification scheme defined below and shown in Figure 8. This classification scheme forms the basis for creating labels that capture image features in each figure. We plan to refine this classification scheme in the future. Currently, at the first level, images are classified into *Graphical* and *Experimental* classes. For the *Experimental* class, we define three subclasses: *Fluorescence Microscopy*, *Gel*, and *Other Microscopy*. The reason is that the three subclasses are visually distinct and obtained in different experimental settings. We manually labeled a few hundred subfigures in each class to train a classifier under this classification scheme[7]. We use two SVM (Support Vector Machine) classifiers: one at the root

[7] Appendix B provides more information about the subclasses we define.

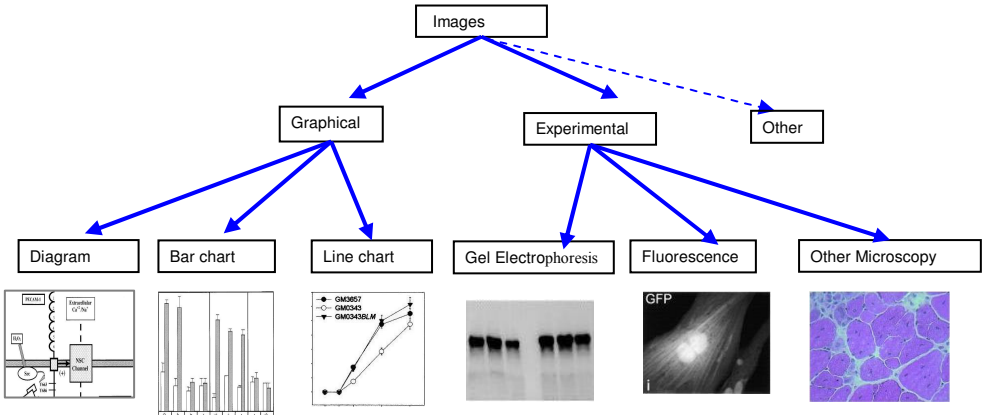

Figure 8. A hierarchical image classification scheme. A sample image for each class is given. More samples for each class are provided in Appendix B. At the first level, images are classified into Graphical and Experimental images. Other types of images found in publications include photographs of people, pictures of mice, etc. In our current work, we manually pre-filter the extracted subfigures to remove such Other images. At the second level, Experimental images are classified into Fluorescence Microscopy, Gel Electrophoresis, and Other Microscopy images. Graphical images are classified into Line Charts, Bar Charts, and Diagrams. In our experiments, Graphical images are not classified further; we focus on classification of Experimental images into Gel, Fluorescence, and Other Microscopy images.

level to classify the images into *Graphical* and *Experimental* images, and the other at the second level of the classification hierarchy to further classify *Experimental* images into one of the three subclasses. In our experiments, Graphical images are not classified further. Thus, every subfigure is assigned one of four class labels: *Graphical, Fluorescence Microscopy, Gel*, and *Other Microscopy*. Examples of subfigure classification results are shown in Figure 6.

The following image features are used by the SVM classifiers:

- Statistics based on gray-level histograms, including statistical moments up to three orders, and gray-scale entropy.
- Haralick co-occurrence textures, including contrast, energy, correlation, and Inverse Difference Moment [15].
- Edge direction histogram [16].
- Features based on run-length analysis of binary images.

The image feature vectors are normalized before classifying them with SVM classifiers and clustering them. The first level SVM classifier was trained using 1,600 subfigures (500 *Graphical* and 1,100 *Experimental*) and the accuracy is 95% using the ten-fold cross-validation testing method. The second level SVM classifier was trained using 1,100 *Experimental* subfigures (300 *Gel*, 500 *Fluorescence Microscopy* and 300 *Other Microscopy*) and the accuracy is 93%.

(4) Subfigure clustering. Next, we perform clustering to define fine-grained image classes automatically. In Step (3), all the training subfigures were classified into one of four classes. There are about 10,000 subfigures in the training data. We felt that for better accuracy we should partition the classes into subclasses, as four manually defined classes may not provide sufficient discrimination among thousands of subfigures. We thus use clustering for obtaining subclasses of *Experimental* subfigures. *Graphical* images are not clustered further. Since the number of subfigures belonging to the *Fluorescence Microscopy* class is significantly larger than the other two classes, the *Fluorescence Microscopy* subfigures are clustered into 20 clusters, while subfigures belonging to the other two classes are clustered into 10 clusters each. Currently, we choose the number of clusters heuristically. Different number of clusters could be used. In the future, we may test how this choice affects the classification performance.

Clustering should group together images with similar characteristics. The choice of image features is critical for the effectiveness of clustering. In the current clustering, we use all the gray-level image features used in the subfigure classification described in step (3). More discriminant features for each class may be used as well. Further discussion of document

image classification techniques is given in our earlier survey paper [17]. To summarize, subfigures within each subclass of *Experimental* are clustered in this step; the clustering results are used to assign a cluster label to each Experimental subfigure.

(5) Create an image-based feature vector of each document. Using the classification and clustering results from steps (3) and (4), we assign a label to each subfigure. For example, the top left subfigure in Figure 6 is assigned the label *F17*, where F stands for *Fluorescence* and *17* stands for cluster *17* in the clustering of *Fluorescence Microscopy* subfigures. The labels of all the subfigures in each document are combined to create an overall description of the document based on its image features. Then a feature vector is extracted from the description. The frequency of labels in each document is used as the feature vector. For example, the description of the document shown in Figure 6 is:

graphics graphics graphics F19 graphics graphics E2
F17 F9 F19 F16 graphics
graphics graphics graphics G6 G7
graphics G1 graphics G3 graphics
F17 G0 graphics graphics graphics graphics E7 F6 G6 E5 graphics
E1 graphics E5 G1 G4 graphics

In this description, G represents *Gel*, F represents *Fluorescence* and E represents *Other Microscopy*. The image description is created by combining the labels of 39 subfigures, drawn from the six figures in this specific document.

(6) Train a Naïve Bayes Classifier. Given the image-based description created in Step (5), we build feature vectors representing each document and train a Naïve Bayes Classifier using all the training documents. We use the MALLET toolkit for feature vector creation and document classification [18].

3.2.2. Executing the classifier. The results of the training phase (Section 3.2.1) are the clusters for each of the three *Experimental* subclasses (Step 4) and a classifier (Step 6). Given an input document, we classify it using the following procedure: First, the document goes through steps (1) - (3), the same figure-based preprocessing as in the training phase. Then each subfigure is assigned the cluster label of its nearest neighbor in the training set using the results of training Step (4). An image-based description is created containing a list of labels of all the subfigures in the document, similar to training Step (5). Then a feature vector is computed and fed into the Naïve Bayes classifier from training Step (6). The classifier labels the input document as either *positive* or *negative*.

3.3. Experimental results

To test our method, we used the data set from the categorization task of TREC Genomics Track 2004 (Section 2.3). In the experiments described here, we use only full-text documents from the *Journal of Cell Biology (JCB)* provided in TREC Genomics Track 2004. A summary of the training and test sets is shown in Table 1.

Table 1. The distribution of positive and negative documents in the training and test data sets.

	Total documents	Positive documents	Negative documents	Total fgures extracted
Training (JCB 2002)	256	26	230	1881
Testing (JCB 2003)	359	34	325	2549

The results are shown in Table 2 using the abbreviations: TP (True Positive), FP (False Positive), FN (False Negative) and Pos (number of positive samples in the test data). We use the same evaluation metrics as were used to evaluate the triage subtask in TREC Genomics 2004. As reported by Hersh et al. [6], the primary evaluation metric was the normalized Utility value, calculated based on equation (1). The constant 20 serves to weigh the evaluation toward positive answers: the cost of missing a relevant document is much greater than the cost of including an irrelevant one.

$$U_{norm} = \frac{(20 * TP) - FP}{20 * Pos} \qquad (1)$$

As we use only a subset of the training and test documents, our results are not directly comparable to those obtained in the TREC2004 Triage task. All 59 of the TREC 2004 Triage runs were based on full-text documents, including figure captions, but none used the analysis of figure images. In contrast, our results use only figure images, and make no use of text. As shown in Table 2, our results are roughly comparable to the average results in TREC 2004 runs. This is encouraging, suggesting that a combination of figure and text analysis may yield good results in the future.

Table 2. Classification results, using the evaluation metrics from [6]. Results from the TREC 2004 Triage runs are shown for an informal comparison. Due to the efforts involved in obtaining figure images, we only used a fraction of the test and training documents used in the TREC Triage task. Our testing used 34 positive and 325 negative documents, whereas the TREC 2004 Triage testing used 420 positive and 5623 negative documents.

	Utility $U_{norm} = \frac{(20*TP)-FP}{20*Pos}$	Precision $\frac{TP}{TP+FP}$	Recall $\frac{TP}{TP+FN}$	F-score $\frac{2*recall*precision}{recall+precision}$
Our system	0.3074	0.2791	0.3529	0.3117
Mean of 59 runs in TREC 2004 triage subtask (from [6], Table 6)	0.3303	0.1381	0.5194	0.1946

3.4. Future work

The current research is a preliminary exploration of the possibility of using figures for document triage. We believe that a refined classification scheme for subfigures is important for improving the result. Feature selection is also crucial to improve subfigure classification and clustering performance. In our future research, we will further investigate how human curators use figures in judging whether a document supports annotation, and how figures are used during the annotation process. Observing how humans handle the task is expected to suggest further ideas on how to automate (parts of) it.

We are also interested in combining the analysis of text, ontology, and figures for document triage and annotation tasks.

4. Conclusion

There are abundant research opportunities for document image analysis in support of the creation and maintenance of biomedical digital libraries. Figures and figure captions are information rich portions of documents. Due to their high availability, abstracts have been the target of most biomedical text-mining efforts to date. However, much useful information only exists in full-text papers. Figures and figure captions, which are only available within the full-text of biomedical articles, usually present important experimental findings. Figure images often play a key role in understanding the papers' results. Therefore, combining image analysis with text analysis is expected to help resolve ambiguity and improve the effectiveness of literature mining.

Automatic annotation of organism databases is expected to help organism databases keep up with the increasing number of biomedical publications. Conversely, annotated organism databases are important resources for literature mining in digital libraries, as the annotated data from biological databases can be used to train literature mining systems to perform useful tasks [5].

There are several challenges when applying document image analysis techniques to biomedical literature mining. In contrast to the millions of abstracts in PubMed, the number of full-text documents is still limited. Easy-to-use electronic versions (e.g. articles in XML format), with separately accessible figures and text are available only for some of the papers. For other cases (e.g. articles in PDF or image format), preprocessing has to be performed to separate text and figures and to associate figures with figure captions. This preprocessing is difficult and error prone. The variability of figures (as shown in Figure 7) provides challenging research opportunities for using figures to support literature mining.

Acknowledgments

We gratefully acknowledge the financial support provided by NSERC, Canada's Natural Sciences and Engineering Research Council and by the Xerox Foundation.

References

1. B. de Bruijn and J. Martin, "Getting to the (c)ore of knowledge: mining biomedical literature". *Int. Journal Med. Inf.* 2002; 67(1-3):7-18.
2. L. Hirschman, J. C. Park, J. Tsujii, L. Wong and C. H. Wu, "Accomplishments and challenges in literature data mining for biology". *Bioinformatics*, 18(12):1553-1561.
3. H. Shatkay and R. Feldman, "Mining the biomedical literature in the genomic era: an overview". *Journal of Computational Biology,* Vol. 10, No.6, 2003, pp. 821-855.
4. S. Dickman, "Tough mining, the challenges of searching the scientific literature". *Plos Biology* 2003, 1(2):144-147.

5. A. S. Yeh, L. Hirschman, and A. A. Morgan, "Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup". *Bioinformatics* 2003, 19:i331-i339.
6. W. R. Hersh, R.T. Bhuptiraju, L. Ross, P. Johnson, A.M. Cohen, and D.F. Kraemer, "TREC 2004 Genomics Track overview". *Proc of TREC 2004*, NIST Special Publication 2005. http://ir.ohsu.edu/genomics/
7. W. R. Hersh, A. Cohen, et al. "TREC 2005 Genomics Track overview". The Fourteenth Text Retrieval Conference - TREC 2005.
8. L. Hirschman, A. S. Yeh, C. Blaschke and A. Valencia., "Overview of BioCreAtIvE: critical assessment of information extraction for biology". *BMC Bioinformatics* 2005, 6(Suppl 1):S1, 1-10.
9. Y. Regev, M. Finkelstein-Landau, R. Feldman, et. al. "Rule-based extraction of experimental evidence in the biomedical domain - the KDD Cup (Task 1)." *SIGKDD Explorations* 4(2): 90-92 (2002).
10. K. Darwish and A. Madkour, "The GUC goes to TREC 2004: using whole or partial documents for retrieval and classification in the Genomics Track". *Proc of TREC 2004,* NIST Special Publication 2005.
11. R.F. Murphy, Z. Kou, J. Hua, M. Joffe, and W.W. Cohen, "Extracting and structuring subcellular location information from on-line journal articles: the Subcellular Location Image Finder". *Proc. IASTED Int. Conf. on Knowledge Sharing and Collaborative Engineering (KSCE-2004).* SLIF server web site: http://goblin.cbi.cmu.edu:8080
12. P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary". *Proc. Seventh European Conference on Computer Vision*, 2002. pp. 97–112.
13. H. R. Widlund, M. A. Horstmann, E. R. Price, et al. "Beta-catenin-induced melanoma growth requires the downstream target Microphthalmia-associated transcription factor". *Journal of Cell Biology* 2002 Sept. 16; 158(6):1079-87.
14. R. C. Gonzalez and R. E. Woods. Digital image processing. Prentice-Hall, 2002.
15. R.M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification". *IEEE Transactions Systems, Man, and Cybernetics, Vol. SMC-9, 1973, 610-621.*
16. A. K. Jain and A. Vailaya, "Shape-based retrieval: a case study with trademark image database". *Pattern Recognition*, 31(9), 1369-1390, 1998.
17. N. Chen and D. Blostein, "A Survey of Document Image Classification: Problem Statement, Classifier Architecture and Performance Evaluation". *Int. Journal of Document Analysis & Recognition*, to appear.
18. A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit". http://mallet.cs.umass.edu, 2002.
19. FlyBase Consortium, "The FlyBase database of the Drosophila genome projects and community literature". *Nucleic Acids Res.* 2003, 31:172-175.
20. J. McEntyre and D. Lipman, "PubMed: bridging the information gap". *CMAJ (Canadian Medical Association Journal),* 2001, May 1;164(9):1317-9.
21. The Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource". *Nucleic Acids Res. 2004*, 32:D258-D261.
22. Y. Hu and R. F. Murphy, "Automated interpretation of subcellular patterns from immunofluorescence microscopy". *J. Immunol. Methods 290:*93-105.K. 2004

Appendix A. Biomedical Digital Libraries

This appendix provides background information about organism databases (Section A.1), document databases (Section A.2), and standardized terms for annotation (Section A.3). Biomedical digital libraries, such as organism databases, protein databases or genomics sequence databases, store biomedical information and provide retrieval interfaces for researchers in biomedicine. All of these databases are constantly updated. Organism databases, such as FlyBase and MGI, annotate genes or gene products of specific organisms by finding evidence from the published literature. A set of standardized terms (in the form of *controlled vocabularies* or of *ontologies*) is important for sharing information across different databases and for supporting efficient information retrieval. Examples of such standardized vocabularies are MeSH (Medical Subject Headings) for indexing biomedical documents and GO (Gene Ontology) for annotation of genes and proteins.

A.1. Organism databases

We introduce here two examples of organism databases that are referred to in the paper.

A.1.1. MGI - Mouse Genome Informatics. The MGI (Mouse Genome Informatics) system is an initiative of the Jackson Labs (http:// www.informatics.jax.org/). It provides integrated access to data on the genetics, genomics and biology of the laboratory mouse. The mouse is the most common model organism for the study of mammalian biology and human disease. The major databases that MGI provides include the Mouse Genome Database (MGD), the Gene Expression Database (GXD), and the Mouse Tumor Biology database (MTB). MGI integrates experimental knowledge with information derived from both literature and online sources.

One of the goals of MGI is to provide structured, coded annotation of gene function from the biological literature. MGI participates actively in the

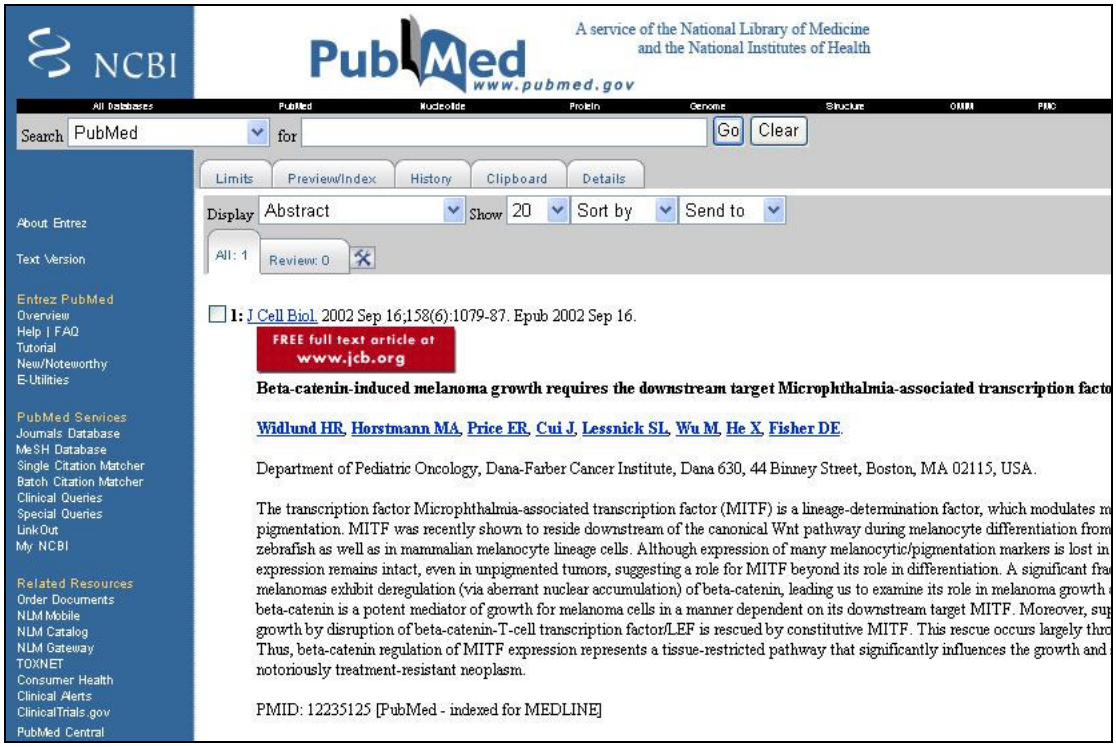

Figure A.1. The result of searching an article citation from PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi). The sample article with PMID 12235125 is retrieved, in response to the user typing 12235125. PubMed provides a link to the full-text article.

development and application of the Gene Ontology (GO).

A.1.2. FlyBase - a database of the Drosophila genome. FlyBase provides integrated access to the fundamental genomic and genetic data of *Drosophila* (fruit-fly) and related species (http://flybase.bio.indiana.edu). The fruit-fly is one of the most studied eukaryotic organisms and was a central model for the human genome project.

FlyBase provides value-added, annotated genetic and genomic data from the *Drosophila* literature. All the information in FlyBase is attributed, that is, associated with a specific bibliographic citation [19]. FlyBase, like MGI, is one of the founding members of the GO Consortium. FlyBase annotates gene entries with GO terms and GO evidence codes.

A.2. Document databases: PubMed

PubMed (http://www.pubmed.gov) is an on-line service provided by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) (http://www.ncbi.nlm.nih.gov). It provides access to over 15 million citations of biomedical articles, mostly from MEDLINE (http://www.nlm.nih.gov/pubs/factsheets/medline.html), going back to the 1950s. PubMed includes links to full text articles and other related resources. Title, authors, abstract are freely available. PubMed indexes documents by MeSH terms (appendix A.3.1), identifiers, authors, and organisms. PubMed ID provides a unique identifier for each article. When an article is used as a reference in an organism database, a PubMed ID is always provided to link to the article. Figure A.1 shows a screen shot from using the PubMed search engine to find the article whose PubMed ID is 12235125. Figure A.2 shows the complete indexing information by the National Library of Medicine associated with the sample article shown in Figure A.1.

PMID- 12235125 OWN - NLM STAT- MEDLINE DA - 20020917 DCOM- 20021017 LR - 20050315 PUBM- Print-Electronic IS - 0021-9525 VI - 158 IP - 6 DP - 2002 Sep 16 TI - Beta-catenin-induced melanoma growth requires the downstream target Microphthalmia-associated transcription factor. PG - 1079-87 AB - The transcription factor Microphthalmia-associated transcription factor (MITF) is a lineage-determination factor, which modulates melanocyte differentiation and pigmentation. MITF was recently shown to reside downstream of the canonical Wnt pathway during melanocyte differentiation from pluripotent neural crest cells in zebrafish as well as in mammalian melanocyte lineage cells. …(part of the abstract) AD - Department of Pediatric Oncology, Dana-Farber Cancer Institute, Dana 630, 44 Binney Street, Boston, MA 02115, USA. FAU - Widlund, Hans R AU - Widlund HR FAU - Horstmann, Martin A AU - Horstmann MA FAU - Price, E Roydon AU - Price ER FAU - Cui, Junqing AU - Cui J FAU - Lessnick, Stephen L AU - Lessnick SL FAU - Wu, Min AU - Wu M FAU - He, Xi AU - He X FAU - Fisher, David E AU - Fisher DE LA - eng GR - AR 43369/AR/NIAMS GR - HL 07574-20/HL/NHLBI PT - Journal Article DEP - 20020916	PL - United States TA - J Cell Biol JID - 0375356 RN - 0 (Cytoskeletal Proteins) RN - 0 (DNA-Binding Proteins) RN - 0 (Luminescent Proteins) RN - 0 (Trans-Activators) RN - 0 (Transcription Factors) RN - 0 (lymphoid enhancer-binding factor 1) RN - 0 (microphthalmia-associated transcription factor) RN - 146409-33-8 (beta catenin) RN - 147336-22-9 (Green Fluorescent Proteins) SB - IM **MH - Animals** **MH - Apoptosis** **MH - Cell Division** **MH - Cell Line** **MH - Comparative Study** **MH - Cytoskeletal Proteins/*physiology** **MH - DNA-Binding Proteins/genetics/metabolism/*physiology** **MH - *Gene Expression Regulation, Neoplastic** **MH - Green Fluorescent Proteins** **MH - Humans** **MH - Luminescent Proteins/metabolism** **MH - Melanoma/*genetics/metabolism/pathology** **MH - Melanoma, Experimental/genetics/metabolism/pathology** **MH - Mice** **MH - Promoter Regions (Genetics)** **MH - Research Support, Non-U.S. Gov't** **MH - Research Support, U.S. Gov't, P.H.S.** **MH - Trans-Activation (Genetics)** **MH - Trans-Activators/*physiology** **MH - Transcription Factors/metabolism/*physiology** **MH - Transfection** **MH - Tumor Cells, Cultured** EDAT- 2002/09/18 10:00 MHDA- 2002/10/18 04:00 PHST- 2002/09/16 [aheadofprint] AID - 10.1083/jcb.200202049 [doi] AID - jcb.200202049 [pii] PST - ppublish SO - J Cell Biol 2002 Sep 16;158(6):1079-87. Epub 2002 Sep 16.

Figure A.2. Keywords and abstract information available in MEDLINE for a sample PubMed article. The article with PMID 12235125, is also shown in Figure A.1. The MeSH terms are in bold font in the second column.

A.3. Standardized terms for annotation: MeSH and GO

A.3.1. MeSH: Medical Subject Headings. MeSH is the National Library of Medicine's controlled vocabulary (http://www.nlm.nih.gov/mesh). It consists of sets of terms organized in a hierarchical structure, which allows searching at various levels of specificity. Examples of MeSH hierarchical structures are shown in Figure A.3. At the most general level of the hierarchical structure, the headings are broad such as "Investigative Techniques". More specific headings are found at lower levels, such as "Electrophoresis, Gel, Two-Dimensional" and "Microscopy, Fluorescence." There are 22,997 descriptors in MeSH as of November 2005 and it is constantly updated. As discussed in appendix A.2, the MeSH thesaurus is used for indexing biomedical articles for PubMED.

Figure A.3. Samples of hierarchical structures of MeSH. The screen shot was taken from the National Library of Medicine's MeSH browser (http://www.nlm.nih.gov/mesh).

A.3.2. Gene Ontology. The Gene Ontology (GO) (http://www.geneontolgy.org) project began in 1998 as collaboration among the mouse, yeast, and fruit fly model organism groups, and has grown to include many other databases. As stated by the GO consortium, their goal is to produce a structured, precisely defined, and controlled vocabulary for describing the roles of genes and gene products in any organism [21]. By providing a common language for annotations, the information in organism-specific databases has the potential to be integrated. GO has over 19,000 terms as of December 2005; it is actively maintained and continually expanded.

The GO vocabularies are categorized into three directed acyclic graphs: *Cellular Component*, *Molecular Function* and *Biological Process*, providing hierarchical structures for describing genes and gene products. Organism databases use GO codes to annotate specific genes and gene products. An example of part of the GO *molecular function* hierarchy pertaining to the MGI databases is shown in Figure A.4.

Each GO annotation assigned to a gene or a protein must be accompanied by a document identifier, such as a PubMed identifier, and an evidence code. The evidence code indicates what kind of evidence is found in the cited source to support the association between the gene and the GO term.

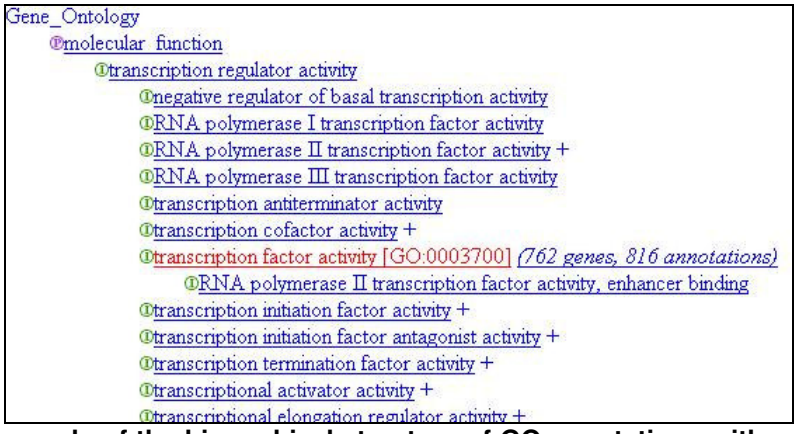

Figure A.4. An example of the hierarchical structure of GO annotations with regard to molecular function. Here, 762 genes in the MGI databases are annotated (once or several times) with GO code GO:0003700 (transcription factor activity); each annotation has an evidence code with reference to a publication or other source of evidence. The screen shot was taken from the MGI's Gene Ontology browser (http://www.informatics.jax.org /searches/GO.cgi?id=GO:0003700).

Appendix B. Examples of images in the six classes defined in Figure 8

We show here some more examples of images for the six classes of subfigures defined in Figure 8. Six classes of subfigures were defined manually by inspecting segmented subfigures from sample documents. We inspected about 250 documents with approximately 1,880 figures. The total number of extracted subfigures is about 11,000. About 100 subfigures that fell outside of the six classes were pre-filtered manually. These include pictures of mouse and photographs of people. We chose representative subfigures for each defined classes. Manually defining classes is a subjective process, which relies on knowledge of subfigures obtained from different biological experiments. Our current classification scheme can (and should) be refined. Figure B.1 shows the samples of Line Chart; Figure B.2 shows the samples of Bar Chart; Figure B.3 shows the samples of Diagram; Figure B.4 shows the samples of Gel; Figure B.5 shows the samples of Fluorescence Microscopy; and Figure B.6 shows the samples of Other Microscopy.

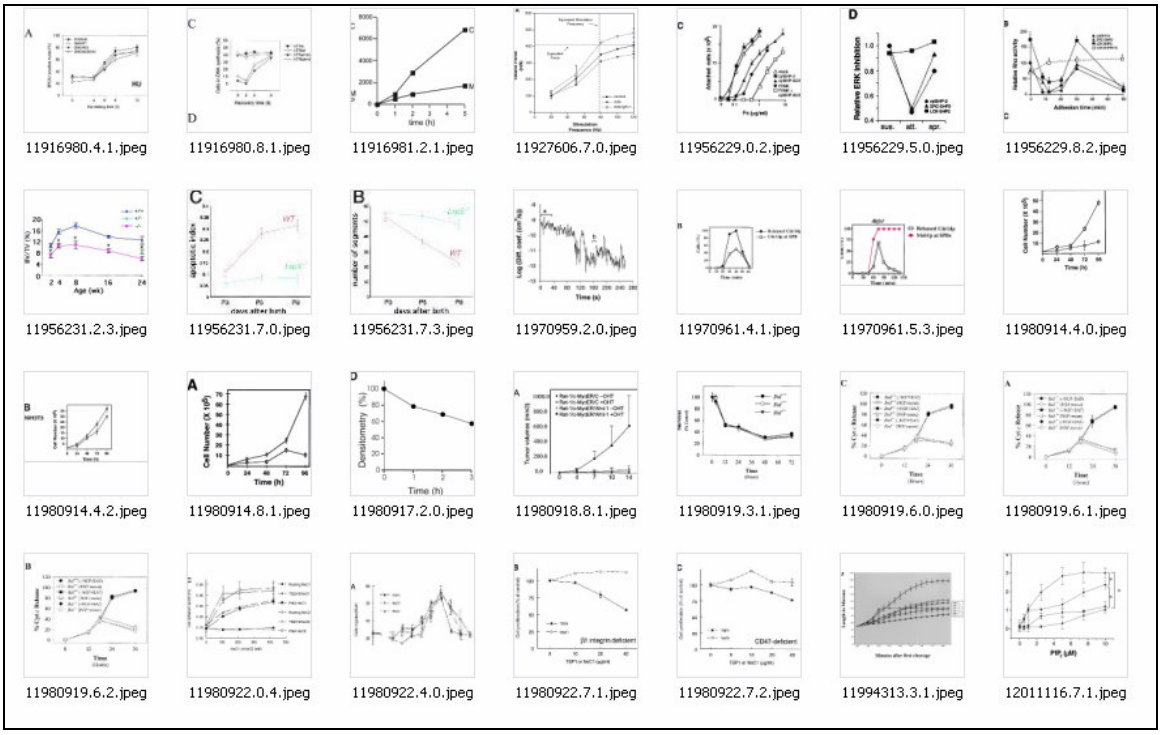

Figure B.1. Samples of Line Chart

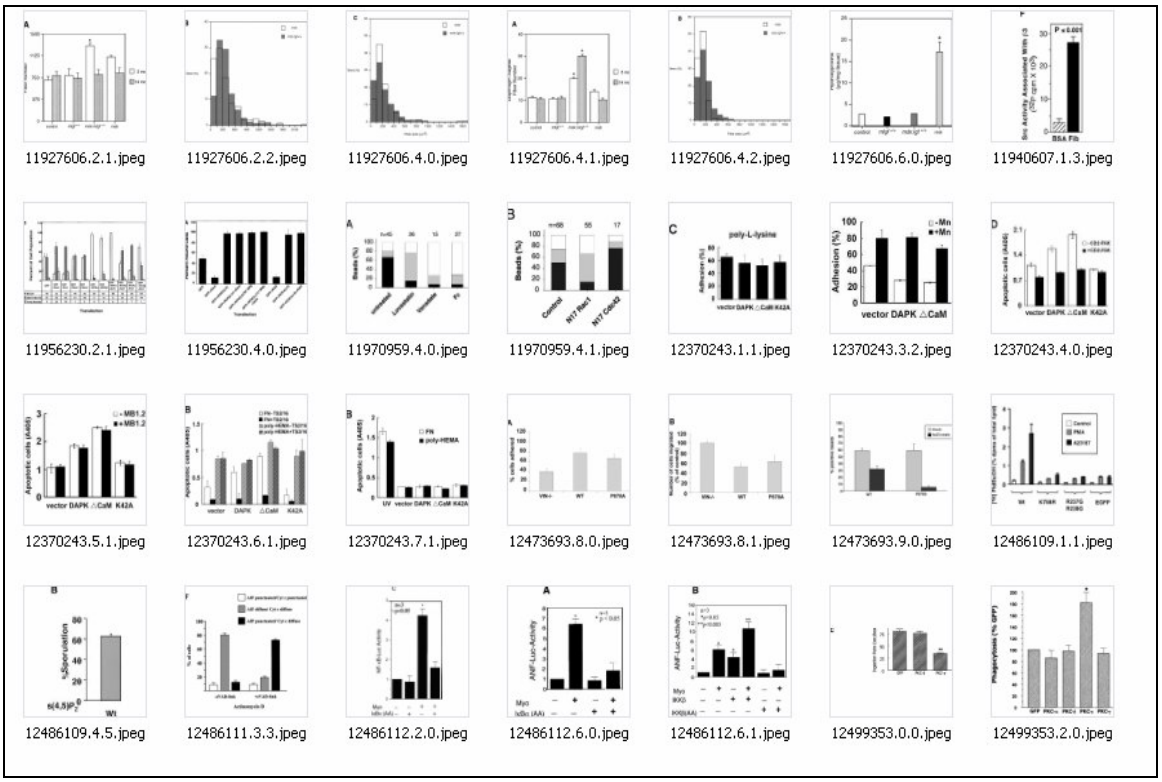

Figure B.2. Samples of Bar Chart

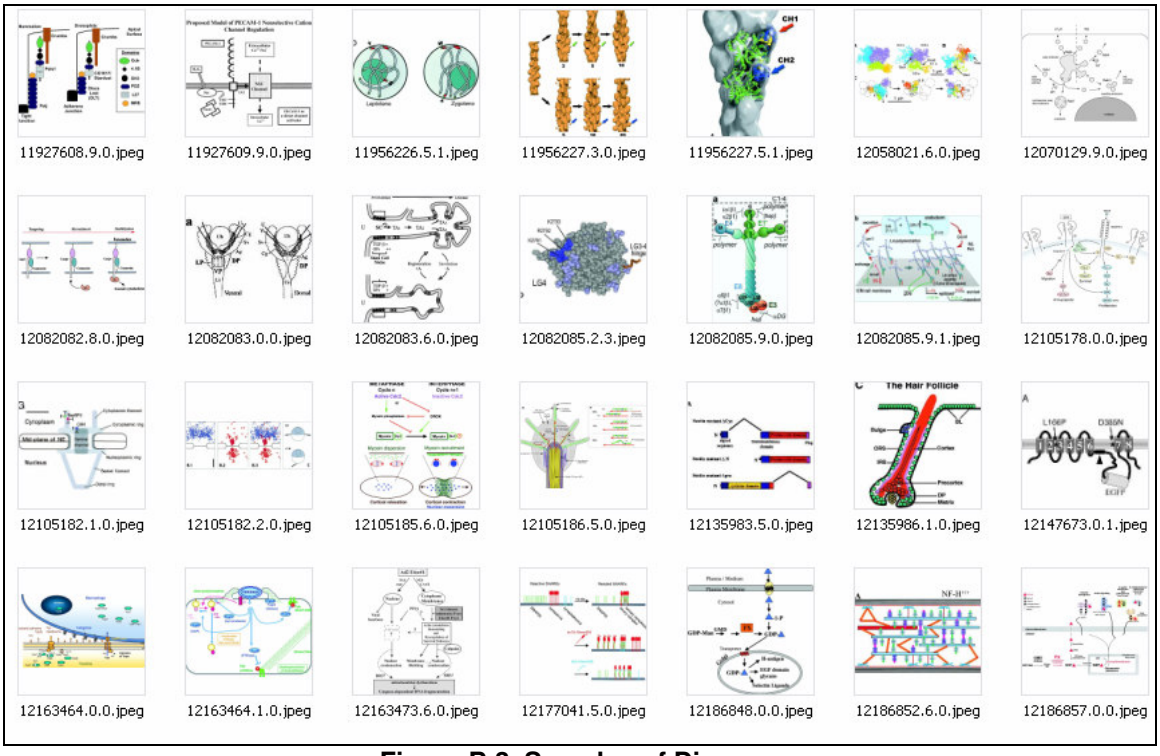

Figure B.3. Samples of Diagram

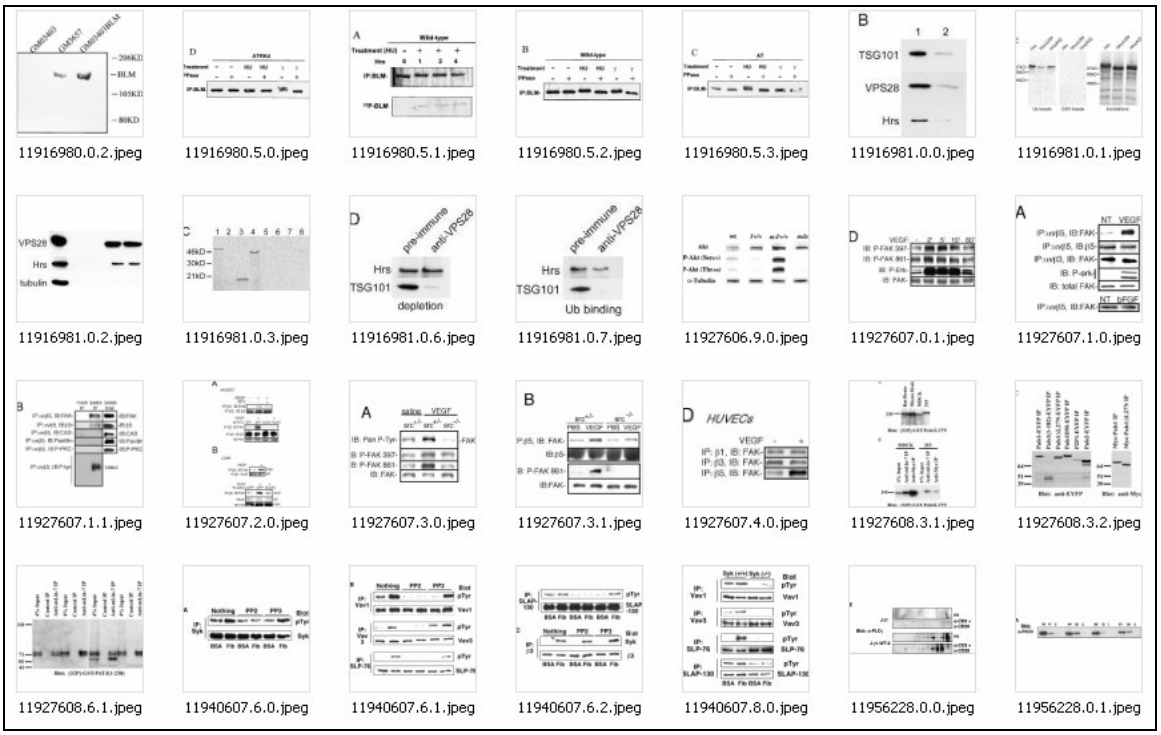

Figure B.4. Samples of Gel

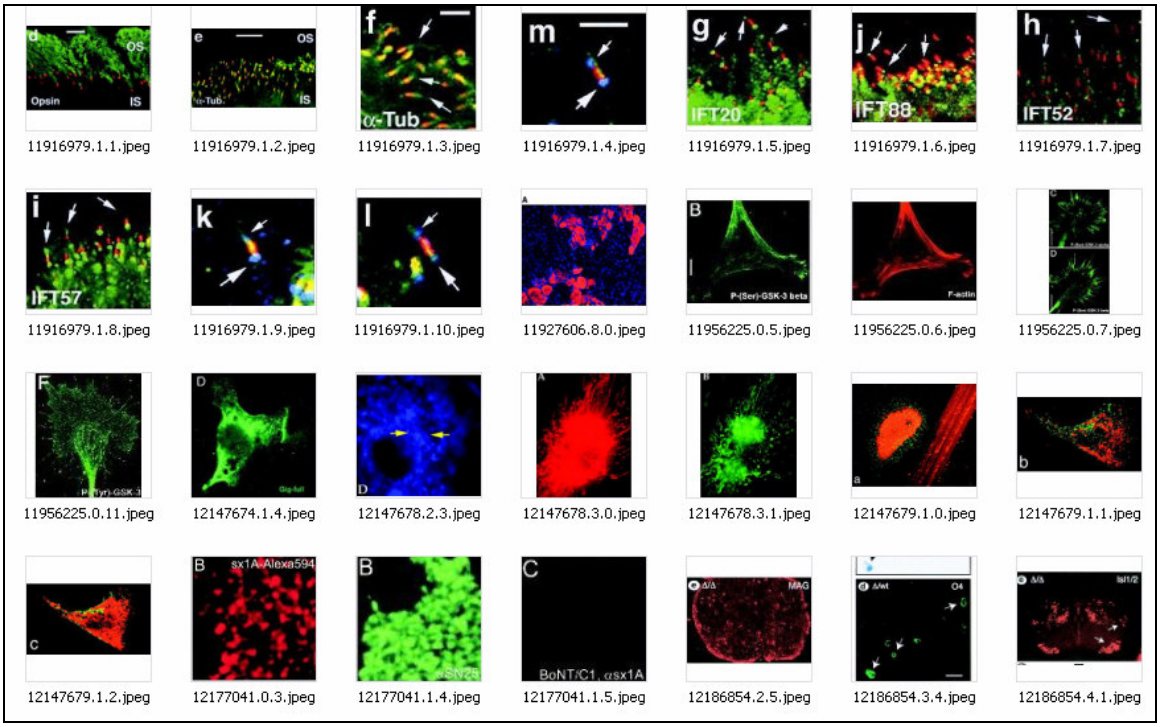

Figure B.5. Samples of Fluorescence Microscopy

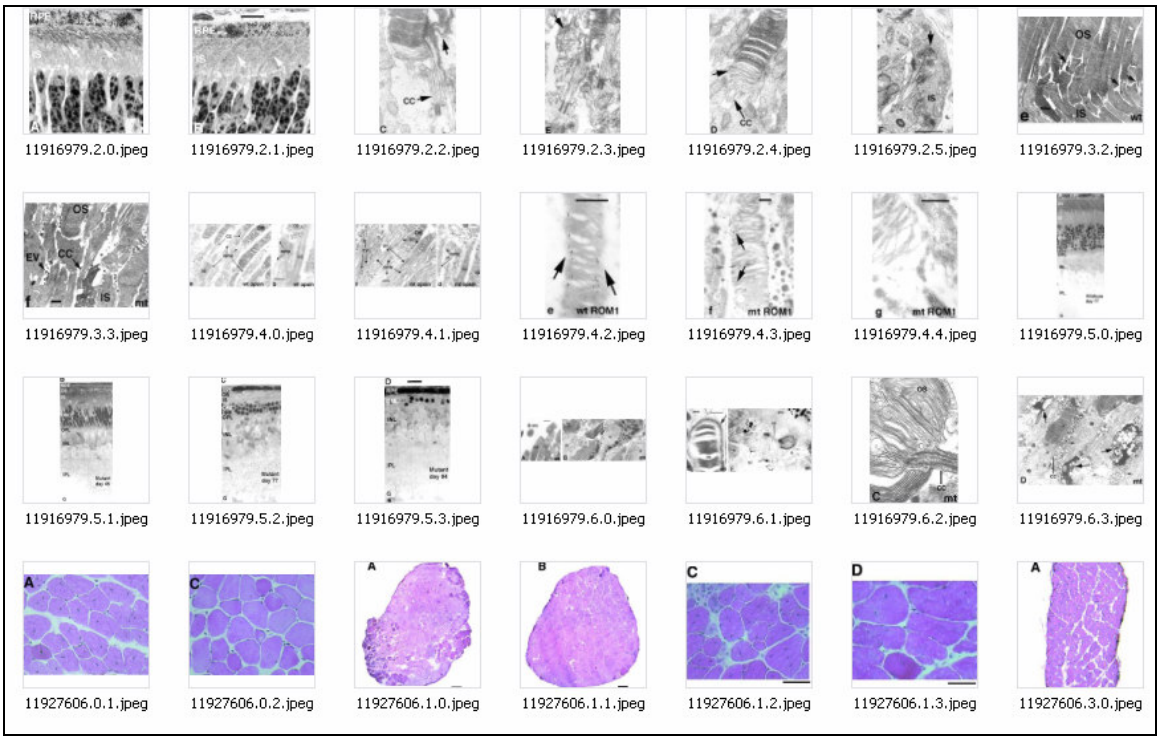

Figure B.6. Samples of Other Microscopy