Mining Medical Literature

Based on the following papers:

1 Motivation

This lecture introduced the background of what it means to “mine” medical literature, why one would do it, and different paradigms and methods of carrying out the data mining of medical literature. Additionally, the related topic of finding synonyms in medical literature is explored, shedding more light on how literature can be mined for underlying meaning.

2 Basics of Mining Medical Literature

To understand the principles of mining the literature, we must first understand the goals of such an endeavor. The general categories of ‘medical, genomics, and proteomics research’, ‘Finding causal links between symptoms or diseases and drugs or chemicals’, and ‘Gene Comparison’ were given. These and other related data mining goals can be split up into two basic categories: Search vs. Discover.

2.1 Search vs. Discover

A fairly clear distinction can be drawn between mining goals that fall under the ‘search’ category, and those that fall under ‘discover.’ Search can be described as a goal-oriented approach: you have an idea of what you are looking for -- perhaps all genes related to a particular function -- and wish to mine the available data sources for relevant information (the actual methods of mining will be explored later). Discover is a more open-ended approach, where, as the name implies, you embark on a journey where you examine a large amount of data or text to look for patterns, implied relationships that have not yet been explored, or other correlations that were not previously guessed at.

The data being mined can be broadly categorized as being either structured or unstructured, though there are certainly shades of gray in this area. Structured data is in a form where it has been broken down into component parts in a differentiable way, meaning that the different parts are separately searchable and recognizable. Databases are the standard form of structured data. Unstructured data is generally in the form of text, found in papers, articles, and other scientific publications. With the advent of making most publications available online, it has become much easier to have access to the text of such articles in a way in which computers can search and manipulate them.

The following chart shows the actions performed by both Search and Discover methods, on structured data as well as on unstructured data:
<table>
<thead>
<tr>
<th>Search (goal oriented)</th>
<th>Discover (opportunistic)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structured data (database)</strong></td>
<td>Data retrieval</td>
</tr>
<tr>
<td><strong>Unstructured Data (text)</strong></td>
<td><strong>Information Retrieval</strong></td>
</tr>
</tbody>
</table>

### 2.2 Examples of the 4 method/data form pairs

To illustrate the differences between Search and Discover, and how each is applied to both structured and unstructured data sources, the following examples were shown:

**Data Retrieval:** Searching a company database of customer and product inventory records. This is done in order to retrieve a piece or pieces of stored information, such as “What is the address of Client A” or “How many widgets are currently in stock.” Examples of databases and systems supporting this type of search are SQL queries on DB2 or Oracle database systems.

**Information Retrieval:** A contemporary, visible example of this are the web search engines such as Google. They are goal-driven by the user’s query, but search unstructured data (text in web pages and documents in various formats such as HTML, PDF, etc.) for appropriate matches. How this is done will be explored further in a bit.

**Data Mining:** Like Data Retrieval, this operates on a structured data set. However in this case, a large amount of (usually historical) data is retrieved and examined in the hopes of finding previously unrecognized patterns or relationships. This is an *opportunistic* approach, and the classic example of what this can find is the “beer and diaper” case. It was found from mining the sales data from a large market chain that there were a high number of sales that involved both beer and diapers. This led to placing those two items closer together in the market.

**Information Mining:** This is the attempt to use patterns, trends, and/or domain knowledge in order to find previously unknown relationships and patterns – potential “gems” of information. This must all be obtained from unstructured data sources, such as local documents, scientific publications or abstracts, or web pages in different forms.

Approximately 90% of the world’s data is held in unstructured or semi-structured formats from which it is inherently harder to obtain relevant information. Thus, much effort has been expended to try to improve the information content extracted from these types of sources.
3 Medical Literature Mining and Finding Functionally Coherent Genes

To define a term used often in the exploration of this topic, “Functionally Coherent Genes” are groups of genes that exhibit similar experimental features. Some of these might be genes involved in amino acid metabolism, or in various body stress responses, or the low level electron transport.

Why is so much effort expanded in the text mining of medical literature? It has become plain in the scientific community that there are many multi-functional genes (genes which are involved in more than one biological process), as well as many instances where families of genes exist that are functionally coherent or similar. There has been a shift from studying individual genes to studying whole families of genes together. The complexity of the relationships that can be explored is increasing, and manually searching through text sources is becoming less and less feasible. However the large and increasing availability of online document sources begs for automated methods of information mining.

It is not easy to find functionally coherent genes: while most genes do exhibit multi-functional expression, many of these have just recently been discovered. Additionally, clustering genes by common experimental features often leads to grouping genes into many too many groups to easily identify those that are truly functionally relevant.

3.1 Medical Literature Mining: Methods

Great use has been made of current advances in statistical natural language processing, as well as taking advantage of the fact that while articles may discuss multiple functions, some address specific functionality in detail.

One basic tool used in text mining is the concept of Semantic Neighbors. Two articles are semantic neighbors if they use the same or similar sets of words. Semantic neighbors are established in the following way, as illustrated in the following picture:
• A small gene group of three autophagy genes are shown (boxes with dotted boundaries).
• Articles are scored relative to gene groups (such as this group of 3 genes).
• The boxes with solid boundaries show genes connected to articles that reference them.
• Articles about autophagy function are dark boxes with white lettering.
• Notice that only a few of the referenced articles for each gene are about autophagy, the function in which we are interested. That unites these genes in the group!
• The dark arrows indicate the semantic neighbors of article B.2, an autophagy article. The significance of this article to the group’s unifying function becomes apparent when we notice that many of its neighbors reference other genes in the same group.

After finding the semantic neighbors in a document set, if all the articles about common functionality contain at least one gene in the group, then the group is functionally coherent. This is illustrated below:

Here, the rectangles are articles (about autophagy), the grey dots are genes, and the oval groups the genes into a functionally coherent group. Since each article references at least one gene in the group, the group is coherent.

A scoring method must be used to assess an article’s relevance to the gene group. The scoring method used is called Neighbor Divergence. The number of an article’s neighbors that reference the group is counted. The more of its neighbors that reference the group, the more that article is likely to be relevant itself. But how do we determine what a significant score is? It is posited that a random collection of articles (a non-coherent group) would have relevance scores that form a Poisson distribution based on the number of articles in the group. If the score is significantly above what we would expect randomly, it likely represents a biological function. Graphically, this can be shown:
If score distribution is different from Poisson then gene group represents biological function

3.1 Medical Literature Mining: Improving and Evaluating Results

There are some problems with this method that require the text to be filtered in various ways. One issue is that well-studied genes tend to have semantic neighbors that refer to the exact same gene. The neighbor might not actually be relevant to group function, but it will receive a false positive boost to its score. One way of countering this is to only consider articles that refer to different genes when generating the score.

A way must then be devised of evaluating the quality of the documents found. Are they all relevant? Or perhaps there are some false matches that have slipped into our group. Use is made of two concepts: Precision and Recall. As with many methods, there is an ROC (receiver-operator characteristic) curve that indicates the sensitivity and specificity of the evaluation. Precision and Recall capture these quantities in the following way: Recall indicates the percentage of total relevant documents that were actually found, and precision indicates what percentage of the retrieved documents were actually relevant as opposed to being false positive matches. This is graphically shown as:

\[
\text{RECALL: } \frac{A}{A+B} \times 100\%
\]

\[
\text{PRECISION: } \frac{A}{A+C} \times 100\%
\]
One way to determine if a functionally coherent group has been correctly assessed is to remove supposedly coherent genes with random, or non-coherent ones. The score of the group should monotonically decrease toward what we would expect for a random collection of articles. This was done for two different groups that were found: autophagy and ion homeostasis, and indeed that result was found in both cases with their group scores converging as their genes were replaced.

There are limits to what neighbor divergence can accomplish, for several reasons. First, while it helps us group genes, it doesn’t actually tell us their function. This can be overcome by taking the article that scores the highest for that group and examining it manually, or automatically if possible. Additionally, only the abstracts of articles are used for mining. While it may be possible to break entire articles up into parts, it remains a formidable task to attempt this on an entire-literature search.

Another mining approach that bears some exploration is Synonym Extraction.

4 Synonym Extraction

Synonym extraction refers to the process of mining literature to find equivalent terms (gene and protein terms, in this case). Why would be need to do this? Genes and proteins are often associated across the literature with more than one name. Indeed, more names are continually being added as new information about the structure or function of the gene or protein is discovered. It would help the search evaluation process if all the synonyms were known.

While current data sources such as GenBank and SWISSPROT do include synonyms, there are many problems with these sources:
- They are not up-to-date
- There is disagreement about whether some terms are synonymous or not
- It requires manually intensive curation and review to generate synonyms
- There is a need for an automated method of synonym extraction

4.1 How to find Synonyms?

It is currently a two-step process to find synonyms in the literature. The first step is a pass that semantically identifies the names of genes and proteins. This is done with the use of new “taggers,” which, as the name implies, tags the terms it finds. The second, more difficult step is to decide which, if any, of the terms found are synonyms.

Exploring the second step, what are the methods of determining of two terms are synonyms? Currently, if two terms represent the same biological substance they are synonyms. Meaning, if they represent identical function or sequence, they can be considered synonyms. Other approaches to this are to apply string matching, and to find abbreviations and match them to the full forms of the terms.

Some heuristics used: Synonyms often occur in the same sentence. Additionally, they are usually mentioned in the first few pages of an article. Synonyms are less likely to be introduced for the first time in an article deep into the article. This can be used to cut down the search space (at the risk of missing some few synonyms).
4.2 Detection approaches

The different approaches for finding synonyms are classified by how automated they are.

- Unsupervised (fully automated), which are based on contextual similarity – ‘Similarity’. This means that terms were used in close proximity to each other in order to be found.
- Semi-supervised, which extracts structured relations using patterns - ‘Snowball’
- Supervised - Text Classification/SVM
- Hand-crafted extraction – GPE

Ultimately, a combined system is used which combines the output of SnowBall, SVM and GPE. Each synonym pair is given a confidence score by each system, and the higher the product of their scores, the more likely a pair is to be accepted as a synonym.

Why use a combined system? To try to overcome the disadvantages of each individual system. Unsupervised systems run in cubic time in terms of the size of the dictionary of words being searched, but they find many false positives using their contextual based matching. ‘Snowball’ and SVM (machine-learning based systems) may capture pairs missed by GPE, but may include many more false positives. GPE, the manual knowledge-based system, is more parsimonious in finding synonyms, but the ones it finds are more likely to be correct, and it can be used in conjunction with the previous two to take advantage of the properties of each.

4.3 Synonym Conclusions

Ultimately, these systems do not perform incredibly well, though the combined system does outperform each of the others at every level of recall:

![Graph showing precision-recall trade-off for different methods]
5 Conclusions

There is still significant work to be done in the area of text mining. The current methods of finding functionally coherent genes, as well as synonym pairs have proven to be quite useful, but far from perfect. Improving these methods as well as developing entirely new methods may lead to faster paths toward deeper scientific understanding and discovery.