In silico discoveries for biomedical sciences

Haagen, H. van

Chapter 6

General discussion

This thesis presents *in silico* text- and data-mining techniques for the prediction of biologically related concepts. The methods were evaluated on protein-protein interaction data and genes associated with certain diseases. The main part of the research was the evaluation of the text-mining method called concept profiles. Later on we extended concept profiles with other non-textual information. The many hurdles and findings are discussed below. We conclude with the future directions where text-mining and data-mining can be improved.

1. Evaluating set creation

During this research a large part of the effort was needed to collect training and test data.

Collecting good data for the evaluation of a data-mining system is hard. Here we describe the problems we encountered.

1.1 Nature of biological data

The data used in this study has several characteristics that make the application of existing data and text-mining methods difficult. The world of biology is far more complex than a computer system can model. It is no simple 'black and white' or the use of TRUE and FALSE labels.

First, biological data is sometimes not reliable, and highly dependent on the context it appears in. For instance protein-protein interactions (PPIs) are recorded in protein databases and each database has a level of curation. Some protein interactions are very well described in databases like DIP. These PPIs are confirmed with several independent wetlab experiments or have a lot of literature evidence. Other protein interactions come from high throughput experiments and are recorded in a database like IntAct. High throughput experiments normally contain more false positives. The same holds for instance for the annotation of gene functions in the Gene ontology (GO). In an old release of the GO a gene is assigned a GO term describing a molecular function. In later releases sometimes the GO term becomes obsolete because it was wrongly annotated or the GO term is merged with another term.

Second, the current knowledge is limited and incomplete. Only a small fraction of the total interaction space (e.g. all protein-protein interactions in the human body) is described. This results in overestimation of the prediction performance because the performance is biased towards well studied proteins, i.e. biased towards only this small subset of protein-protein interactions.

Third, biological data change over time. For instance when two proteins are not known to interact, a system would label this protein pair as TRUE NEGATIVE. However in a wetlab experiment the two proteins were confirmed to interact. After this discovery the protein pair would be labeled as TRUE POSITIVE.

In an evaluation process, biological data should be used keeping these characteristics in mind.

1.2 Biological nomenclature

The nomenclature of biological names is not standardized. For genes or proteins there exist multiple accession numbers (e.g. Uniprot, Entrez Gene, or HUGO Gene Nomenclature Committee), synonyms, and abbreviations that all need to be mapped to single unique identifiers. To disambiguate genes in text is difficult because many genes share the same synonym, resulting in homonym problems.
For gene-disease relationships it is even harder. Many of the genes are assigned the name of the disease they are associated with. These samples cannot be used as a test sample. In addition the disease name as it is recorded in databases is hard to recognize in text. For instance Alzheimer disease had over 15 variants recorded in OMIM (e.g. Alzheimer type 2). In text normally this will be described as that they found a new type of Alzheimer disease. Hence not the concept Alzheimer type 2 is recognized but the generic concept Alzheimer disease.
Furthermore, concepts are related to each other in a hierarchical ontology. For instance the concept Duchenne muscular dystrophy (DMD) in the ontology is part of the concept muscular dystrophies. Once DMD is recognized in text as a concept, one could argue if it is informative that in the same text the higher level concept muscular dystrophy is recognized.

1.3 Minimum information requirements for text-mining

In chapter 5 we introduced the roll back analysis. This is a way to simulate a prediction over time. We imposed the constraint for gene-disease relationships in our test set that the two concepts should not be co-mentioned together before the relationship was discovered, to prove that they could have been predicted using the implicit information. However, before that first co-occurrence there should be enough information available which is sometimes also not the case. In order to build a concept profile for a concept we maintained a threshold of at least 5 abstracts where that concept is mentioned.
This limitation resulted in a set of only 18 gene disease pairs described in chapter 4 where the original list started out with roughly 5,000 gene disease pairs in HPRD. The same problem probably occurred in the article by Aerts et al. [1] where they obtained a small set of 10 monogenic and 6 polygenic diseases.

1.4 Curation and confirmation of biological data

One aspect of bioinformatics is that it is important to validate (or verify) every *in silico* prediction with a wet lab experiment. The results of the experiments described in this thesis required interpretation by expert biologists. This introduced a dependency on experts, who had to make time in their busy schedules. Luckily,

the biologists at our department were very helpful, but still the amount of work that could be asked of them was limited. Every bioinformatician would love to have his own private biologist.

1.5 Circular reasoning

Another problem is that many databases have a certain level of redundancy. In machine learning a key step to evaluate a prediction system is to divide the data into a training set and a test set. The training data is used to train all the parameters that are used in the model of the prediction system. The test data is used to evaluate how well the system is able to correctly predict the labels of the test data. Training and test data should be independent. That is, no data that is used for training should be used for evaluation, else this could lead to an over estimation of the performance.

However for biological data it is sometimes not possible to divide the data into an independent training and test set. For instance when a wet lab experiment is done for investigating a PPI, the results will be described in an article and published, and the same result are stored in a database like DIP. To separate the database and article information is difficult. Therefore we introduced in chapter 3 and 5 the retrospective study (or roll back analysis) and do a prediction simulation over time to eliminate the bias. To do this, it is necessary to get access to old releases of databases. Most databases do not store previous releases for download. For bioinformatics purposes this would be extremely helpful to keep track of old releases.

2. Findings

2.1 Implicit information extraction and content

Chapter three and five showed that implicit information extraction works. The information, or the indirect links, that connects two concepts can be derived from the concept profile overlap. It seems that for PPI prediction the dominating concept is normally another protein already associated with one of the two proteins. For instance in chapter three CAPN3 was linked with PARVB via the intermediate protein DYSF. For gene disease relationships it is normally an associated phenotype, or another gene also known to cause another disease. For instance in chapter five RECQ4L was also associated with Rothmund-Thomson syndrome and therefore seems to be a good candidate for Baller-Gerald syndrome because these two syndromes show clinical phenotypic overlap. These two examples are indicative that the implicit information is meaningful and well explains the association between two concepts. We further observe that when an implicit link is found between concepts the link is normally one dominating concept. To verify this, more samples should to be evaluated.

2.2 Added value of other data sources.

In chapter four we investigated if concept profiles can be improved by adding other non-textual data sources. We found that some of the problems encountered with text-mining could be solved with the other data sources. We conclude that this may work but the performance is dependent on the sample you are looking at. It was shown for DYSF that the amount of information in additional databases besides the literature was poor. In the PKD1 case study the disambiguation problem was solved by microarray expression data. The nature and the amount of information from every source has its pros and cons dependent on the sample. Figure 1 shows an example of the AuC output for each database for DMD and HTT to illustrate that for each protein another data source is dominating. In many pattern recognition approaches it is usual to do feature selection for dimensionality reduction resulting in the most informative features. For instance in a microarray experiment the goal may be to look for differentially expressed genes. The number of genes checked start with 30,000 and after filtering (feature selection) the number of genes will vary from 10 till 100. However for combining data sources the number of available data sources, suitable for processing, is already limited. Making databases inter-operable is very important. As stated earlier the data source that is most informative changes with each sample (figure 1). A generic feature selection approach therefore seems not appropriate for biological data. A scientist should be able to select the data source he is interested in. Also on the basis of known knowledge and ROC curve analysis a scientist can get a feeling if the data source is informative for his samples (e.g. a protein). Added value of data sources and feature selection should be considered for each question separately.

2.3 Types of relationships

We did the collection of data for the relationship types 'protein interacts with other protein' and 'a gene when mutated causes a certain disease'. As discussed preciously the prediction performance is dependent on each sample. The same holds for types of relationships. This can be very well explained. For protein interactions 70% of the known PPIs recorded in databases cannot be traced back in PubMed abstracts because normally the interactions are stored in a table in full text. When a gene is found for a disease, the landmark paper will always co-mentioned the gene and the disease in the abstract, if not even in the title. After the landmark paper multiple occurrences happen in articles published after the landmark paper. We did an evaluation of the gene disease relationships in OMIM and found that ~83% of the known pairs have a co-occurrence in MedLine abstracts. The distribution is given in figure 3. We checked another relationship type, that of 'gene has function X' taken from the Gene Ontology. For this relationship type the distributions are given in figure 2. These figures clearly show

that the known relationships are very different. Since gene/disease relationships almost always occur in PubMed abstracts, the association score is in general high.
The null distributions (or random distribution) tend to look the same. For knowledge discovery scientists are interested in new concept pairs (e.g. PPI, gene/disease) previously not recorded in any database but found by our text-mining system. Those are all the pairs from the null distribution. In chapter three we generated a null distribution of random protein pairs. In figure 2 and 3 the null distributions for gene/disease and protein/function are given respectively. The distributions look alike. To investigate if null distributions from different semantic types (e.g. protein pairs or gene-disease pair) as the same and can be treated universal we calculated match scores for 100,000 random protein pairs and 100,000 random gene-disease pairs. The results are plotted in figure 4. This plot clearly shows that the gene-disease pairs (blue) are different from the protein pairs (red) even though the two distributions both have a Gaussian characteristic. An explanation of this difference could lie in the fact that proteins or genes in general are more intensively described that diseases (all diseases besides OMIM are taken into account). Therefore concept profiles for protein/genes are more enriched which results in on average higher match scores. This result means that any pairs of two semantic types cannot be treated universally. For instance, when the match score for a protein pair is significant (e.g. $p<0.01$) calculated under the null hypothesis that any concept pair (regardless the semantic type) is not related, this same protein pair could not be significant (or at least is different) under the null hypothesis that protein pairs are not related.

3 Limitations of text-mining

Concept profiles show a better performance in predicting associations between concepts than the direct relationship approach (described in chapter two and three). However there are still limitations in prediction performance even for concept profiles. First, finding a new relationship between two concepts goes as far as there is information. This means that there must be sufficient information for both concepts. For concept profiles we formulated this that there should be at least 5 articles available for both concepts. Many diseases or proteins are rare that they have not been published about. In this case text-mining fails not because of technical shortcomings but just due to the lack of information.
Second, the lack of information can also be within the implicit information. If two concepts are related to each other is does not mean that they will always be linked with each other via intermediate concepts. If they do not, this is also not due to text-mining shortcomings but that there is no implicit links available.
Third, the biggest limitation is the accuracy of the disambiguation process. This is dependent on the style of writing of the author, i.e. which nomenclature he uses for words and if words are abbreviated. The problem of disambiguation lies in that

humans are more adapted to give names to entities that are easy to recognize and easy to remember on how they are named. Normally this is done using an acronym. In addition biologists do not make a standard convention about the word nomenclature for proteins. As Michael Ashburner [2] once said 'Biologists would rather share their toothbrush than share a gene name'.
It is shown in competitive conferences [3] where state of the art text-mining systems compete with each other that there is a maximum performance reachable (e.g. 0.88 and 0.50 recall and precision respectively). No computer system is ever able to retrieve a 100% score.

4 Future directions

We believe that text-mining and in particular concept profiles are indispensable in biological research. We foresee that text-mining will become a core technology in the so called semantic web. The semantic web is a name giving for a trend going on the Internet. The first trend in Internet development was called web 1.0. It was the collection of all static HTML pages with only plain text. The second generation is called web 2.0 where the web became interactive. Think of user input like credits cards, online bookstores or Wikipedia. The third generation is called web 3.0 or the semantic web. Here the plain text on web pages, blogs and published literature will be linked with each other in a web of concepts, where the links between concepts can be facts generated by information extraction (IE) or can be hypothesis being a novel relationship using text-mining techniques.
There is still a lot to gain in research and the development of text-mining and remarkable some of them are not computer oriented.

4.1 Community annotation

The first development is that of community annotation [4]. With the common technology the way that computers can read text have their limitations in terms of accuracy. Disambiguation remains a key aspect and hard to solve for many concepts. However with the future version of the semantic web and the millions of people on the internet every day this can be solved. A person on the web, a so called community annotator, can screen an article of interest that has been tagged by a text-mining system and correct the words that have been misclassified. With misclassification we mean that a word was too ambiguous to resolve or not recognized because the word does not appear as a concept in the ontology. For a human reader the disambiguation can be done manually even so the ontology can be updated with new concepts or synonyms for existing concepts. Or in the case of the Alzheimer example, a new type can be corrected for in an article years after publication. Since the internet contains millions of users every day, this annotation process increases the accuracy of tagged text over time.

4.2 Making standard nomenclature

The second improvement is for standardizing databases, identifiers and names for concepts. In the past many attempts have been made to come to a common ontology that is accepted world wide by all biology scientists. However thus far these attempts have failed for many reasons, some of which are unexplainable. An example is that in the past years companies developed their own databases for data-mining purposes. Once they published about their database (in online website form) it was used frequently over the coming month. After a period of time the database became old and not maintained. In the end the database 'dies' and is buried on the 'database graveyard'. For world wide co-operation we suggest that biologists get inspired by ICT companies and organizations like IEEE for whom standardizing is a well known principle (http://standards.ieee.org/). For instance, with the digital revolution many electronic devices came available for home users that need to be universal. A compact disk that can be played with any CD player that is bought in Germany or Japan. Or a personal computer where a soundcard works and fits in any motherboard. The universal exchangeability works in this field, hence, it may work in other fields like biology.

4.3 Publish everything in blogs

A last improvement would be in the publication of negative results. In data-mining systems there is often the need to compare groups of data. For instance for a microarray this could be a treatment group of affected patients and the control group (reference group) of healthy people. In chapter one we compared the group of PPIs with the group of random protein pairs. There is no database available that explicitly describes that some proteins do not interact. Publication of any experiment ever done would be valuable for a computer scientist (and even for biologists so they do not reinvent the wheel). The publication can now be done via online blogs, which are generally publicly accessible.

4.4 Multidisciplinary environment

A complete non technical aspect of improvement is the communication between different disciplines. The background of today's bioinformatician in most cases is computer science with very little background in biology. In the same way today's biologist lacks the knowledge in the use of computers. The gap in communication between the computer scientist and biologists hampers the further development in bioinformatics research. For instance, biologist and most other disciplines, not engineering related, are sometimes not aware of what is already possible with today's technology. This results in reinventing the wheel or working with old school technology (e.g. massive storage in excel sheets that better could be stored in professional database systems like Oracle and MySQL). Engineers and computer scientists on the other hand have no idea that people are in need of their computer

and engineering skills. When the two worlds never meet they cannot benefit from each others knowledge. We would like to encourage organizations to strive to let biologists meet with bioinformaticians in order to learn from each other. For instance now there are conferences dedicated to bioinformatics research and mostly visited by bioinformaticians. Same holds for conferences mostly oriented for biology. It would be great if a conference was dedicated to present bioinformatics tools and ideas purely to biologists, and that biologists present their ongoing project to bioinformaticians and want feedback or a bioinformatic solution. Such a conference stimulates the increase of collaborations between biologist and bioinformatician.

1. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
2. Pearson, H., *Biology's name game.* Nature, 2001. **411**(6838): p. 631-2.
3. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., et al., *Overview of BioCreative II gene normalization.* Genome Biol, 2008. **9 Suppl 2**: p. S3.
4. Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., et al., *Calling on a million minds for community annotation in WikiProteins.* Genome Biol, 2008. **9**(5): p. R89.

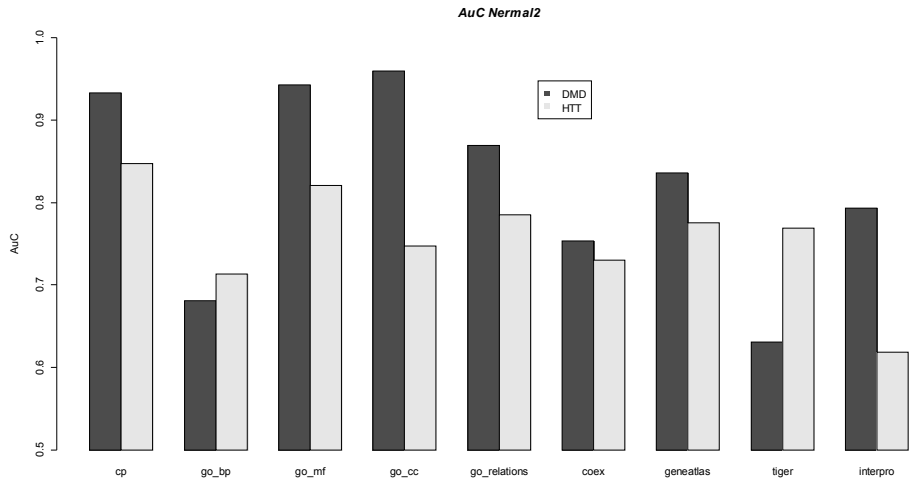

Figure 1. AuC result for DMD and HTT for different databases. The performance is dependent on the protein of interest. Tiger shows opposite behavior then InterPro.

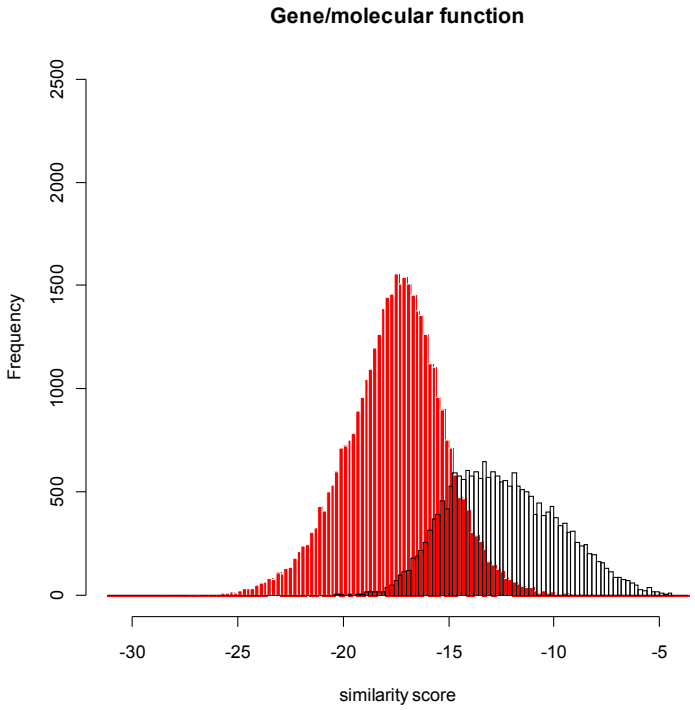

Figure 2. Gene/function distributions

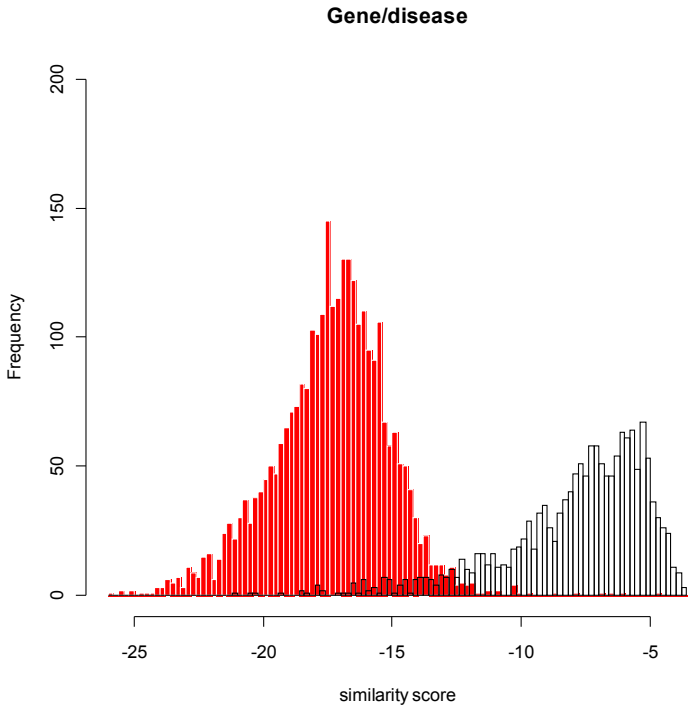

Figure 3. Gene disease distributions

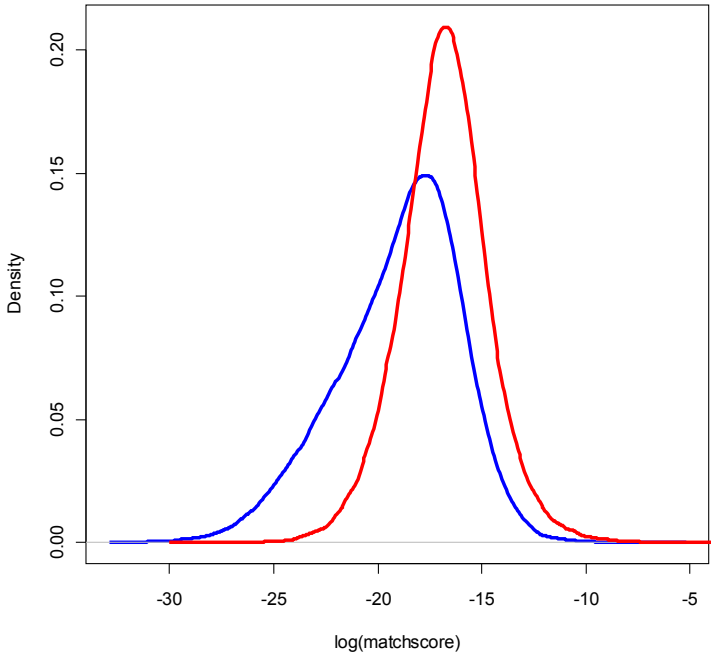

Figure 4. Density plot of random protein pairs (red) and random gene-disease pairs (blue)