Deriving Genetic Networks Using Text Mining

(HS-IDA-MD-02-205)

Elin Ohlsson (a98elioh@ida.his.se)

Institutionen för datavetenskap

Högskolan i Skövde, Box 408

S-54128 Skövde, SWEDEN

Masters Dissertation,

Study Program in Bioinformatics

Supervisor: Jenny Westerlund

Deriving Genetic Networks Using Text Mining

Submitted by Elin Ohlsson to Högskolan Skövde as a dissertation for the degree of M.Sc., in the Department of Computer Science.

02-06-07

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: _____

Deriving Genetic Networks Using Text Mining

Elin Ohlsson (a98elioh@ida.his.se)

Abstract

On the Internet an enormous amount of information is available that is represented in an unstructured form. The purpose with a text mining tool is to collect this information and present it in a more structured form. In this report text mining is used to create an algorithm that searches abstracts available from PubMed and finds specific relationships between genes that can be used to create a network. The algorithm can also be used to find information about a specific gene. The network created by Mendoza et al. (1999) was verified in all the connections but one using the algorithm. This connection contained implicit information. The results suggest that the algorithm is better at extracting information about specific genes than finding connections between genes. One advantage with the algorithm is that it can also find connections between genes and proteins and genes and other chemical substances.

Keywords: Text mining, Genetic network, Mendoza, Protein network.

1 Introduction

Bioinformatics have earlier based their research on gene databases like SwissProt, GenBank etc, but now a new possibility arise to use biomedical literature as a knowledge source when using bioinformatic algorithms. As new biological technologies arise, a new information-rich quantitative science emerges. On the Internet and in the literature there is a huge amount of data available that is represented in an unstructured form and the amount is still growing (Chang et al., 2001, Tanable et al., 1999). PubMed has abstracts that are available for most of the articles and the electronic publishers give access to full text articles. Different methods have therefore been developed that extracts the important knowledge into a more structured form (Chang et al., 2001). It would take too much time to read it all (Fukuda et al., 1998). If the data is gathered in a more structured form it can be highlighted in a well-organized and coherent manner (Tanable et al., 1999, Chang et al., 2001). If one manages to structure the information global views of structural and dynamic information of whole genome sequences and their corresponding gene activity patterns at the RNA and protein level are gained. By using genetic feedback networks the whole system of regulation can be viewed. In such a network every gene is described with respect to how it affects (activates, inhibits) the regulation of another gene (Kurhekar et al., 2002). Gene activity depends on the presence of one or several transcription factors, these are themselves regulated by other transcription factors and therefore a functional interdependence among a large group of genes is created. These groups of genes regulating the activity of each other are known as genetic regulatory networks (Mendoza et al., 1997). There are several different idealizations to genetic networks, a network can for example have a single input and a single output or it can have multiple inputs and multiple outputs. In the future developmental control genes will probably be studied less and less individually, but rather as components of complex gene regulatory networks (Theissen et al., 1999). Theissen et al. (1999) implies that learning to understand the origin and evolution of these gene networks will also help to clarify the origin and diversification of flowers. Mendoza et al. (1999) have done some of this by constructing a genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis by finding information in scientific jounals. They chose to represent the network by a qualitative or "logical" formalization of sets of interacting genes. This abstraction was done because very

little is known about how the regulatory interactions look and the values of the various parameters (Mendoza et al, 1999). At the end of year 2000 new information has been found that could change the molecular network of Mendoza et al (1999) but at the most fundamental level the genes seem to be accurate. In this project a new method is derived which from a text source e.g. Internet collect the data necessary to construct a genetic network. The method is then used to verify and enhance the network of Mendoza et al. (1999).

The data used for constructing the network were collected from the PubMed database (PubMed, 2002). The abstracts were collected using a script executed in Unix. A Perl program was then created that uses the information collected by the script and extracts relations between defined genes and the sentences they were extracted from. Using the abstracts as information source a network over *Arabidopsis thaliana* flower morphogenesis is created. Finally networks over some of the genes involved in the cell cycle in yeast are constructed. The problem definition of the work is that a genetic network can be derived using a text mining algorithm. The networks constructed are not sufficient to construct the whole network but one can use the algorithm to develop parts of networks. Another use of the algorithm is to extract valuable information from the abstracts about the genes one is interested in.

The thesis is ordered in the following way: In chapter 2 Genetic networks and genetic regulatory networks are described and the subject of text mining is discussed. In chapter 3 related and previous works are presented and discussed and the advantages and disadvantages with each of the methods are described. In chapter 4 the aim and objectives are described together with demarcates for this work. The differences from other work are explained and the motivation for the project is discussed. In chapter 5 a description of how the work was performed is given. In chapter 6 the results from the work are presented and in chapter 7, discussed. Finally in chapter 8 the conclusions are given and future work described.

2 Background

This chapter aims to give a fundamental understanding of the concepts of text mining and gene regulation. In this project, a text mining tool is developed, that are able to extract useful information for building a network from public data sources on the Internet like PubMed.

2.1 Text mining / Data mining

Text mining can be defined as:

“the process of analyzing text to extract information from it for particular purposes”

(Yeates, 2002).

The concept of text mining is practically the same as for data mining; they both deal with large amounts of data and are both part of a process called Knowledge Discovery (KDD). Text mining sort out large amounts of unstructured text-based data (Dci, 1999) while the primary goal for data mining is to develop usable knowledge regarding future events. The number of headlines in the KDD process varies depending on how detailed the description of the process are. In this description the process consists of seven different steps (Freeman, 1998). These steps are presented in figure 1 and described in sections 2.1.1-2.1.5. The steps cover the process of KDD from defining the problem to monitoring of the final model (Fayyad et al., 1996; Marakas, 1999). A model can be a way of performing the pattern matching needed to extract information from the text. It can be a program in perl or java or it can be an algorithm with statistical parameters.

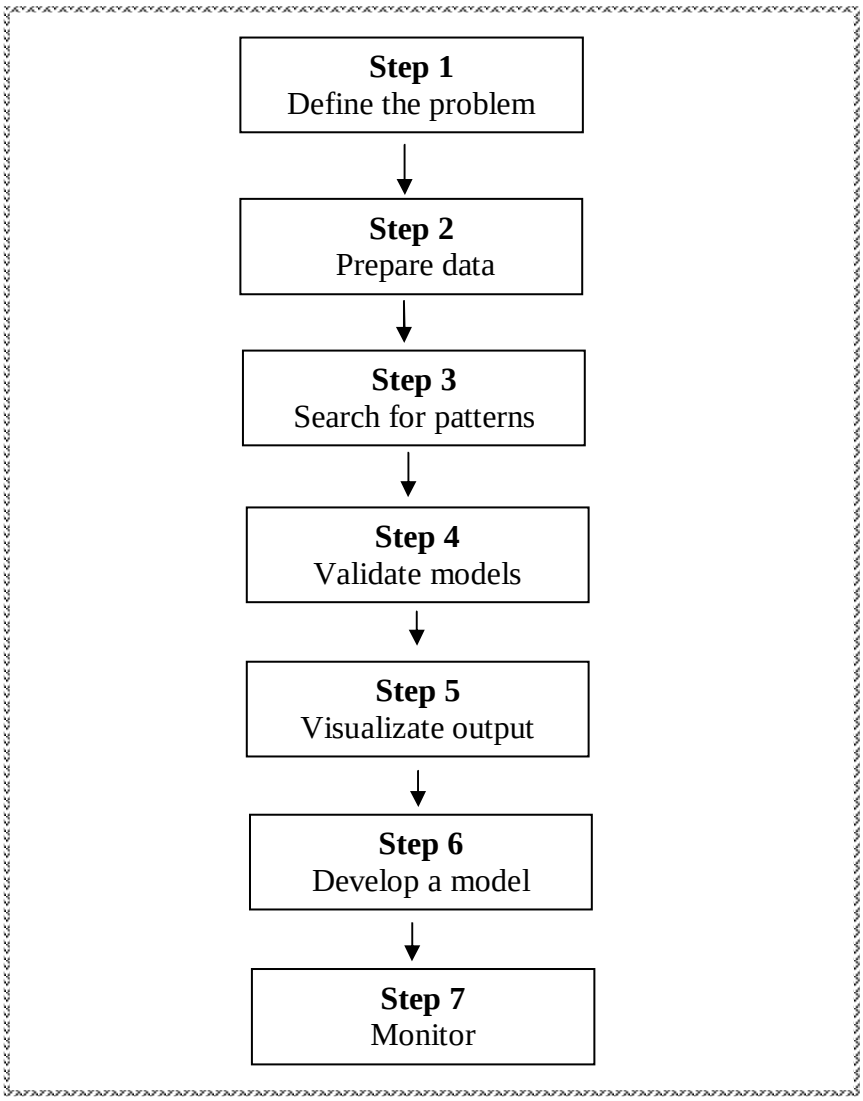

Figure 1. The different steps in a Knowledge Discovery process. Step 1: define the problem e.g. what are the goals and from where is the data collected. Step 2: Collect, transform and clean the data. Step 3: Search for patterns in the data e.g. Text mining/Data Mining. Step 4: Validate the models created in step 3. Step 5: Visualize the output to make it more interpretable. Step 6: Develop the model further by testing it on a new example. Step 7: Monitoring the model to maintain a correct prediction.

2.1.1 Problem definitions

The first step in the KDD process is to define the problem. An example of a problem can be what different kinds of proteins another protein is interacting with. In the KDD process one defines what the goals of the problem are and identify the data that is going to be used i.e. define the different data sources, from where one wants the data

to be collected. The data can be gathered from one or several sources on the Internet like databases, text-documents etc (Freeman, 1998).

2.1.2 Data Collection, Cleaning and Preparing

When the goals of the problem are defined and the identification of different data sources is finished, different data are collected from the data sources described in section 2.1.1. The data must be prepared, this is made by transforming the data collected from the different data sources. The data are transformed into the appropriate format that is going to be the input. If large quantities of text have been extracted it has to be filtered to be able to be processed. One way to filter a text is to translate the user query into relevance metrics, which then can be applied to large quantities of text automatically. Two other text filter methods are widely used: The first is to apply combinations of keywords to identify relevant documents, paragraphs or sentences. For an example, the filter method can specify that an abstract is relevant if it contains a sentence with both the name of the genes and a special word for example the word *inhibits* (Tanabe et al., 1999). The second text filter method that can be used is to use word frequencies to determine the relevance. This filter method identifies that a sentence is relevant if it contains words like gene, protein or inhibit scientifically more frequent than the average document does (Tanabe et al., 1999). There are also more complex methods to filter text. An example is surface clue evaluation that Tanabe et al. (1999) uses in their method MedMiner. Here one examines the contexts of the word to determine if it the word is relevant, in biology surface clue evaluation is used to find protein and gene names in text (Fukuda et al., 1998). Another example, shallow parsing, is when a document is parsed into a list of its markups and text items, using a single regular expression like a dot or a comma (Cameron, 1998). When extracting information on interactions directly from each article the first thing the system must do is to identify material names as gene and protein names. To identify technical terms from large quantities of unrestricted text is a challenging task for natural language processing (Fukuda et al., 1998). The data is also cleaned, by this means that data conflicts (not having the same value for the same data), outliers (unusual or exception values), missing data, and ambiguity are found and resolved. If different data sources are used they must be joined, here several

problems might occur like missing data fields, different time when the data was created and changes may have been done to the data files (Freeman, 1998).

2.1.3 Search for patterns e.g. Data Mining/Text Mining

In the knowledge discovery process, search for patterns is the data mining/text mining step where a model for pattern matching is built. The data collected in previous steps are first processed/transformed into a suitable format; see section 2.1.2 (Freeman, 1998). If the data is insufficient to make a correct prediction, one must enrich the data by adding words or other data sources. This can be done by adding additional synonym terms to the query, which can be found in the data extracted in the previous section. If, instead of text mining, data mining is used, it is sometimes necessary to generate three different samples from the original data. These three samples are then used for training, testing and validating the model (Freeman, 1998).

2.1.4 Validating the models

When the model is created it should be tested on a new example of data that was not used to build the model (Freeman, 1998). It is possible that an iteration of the whole process is necessary, this to be able to understand and incorporate the final result (Ekberg et al., 2000).

2.1.5 Visualization of output

When the extraction of information is finished and the model has been validated, a lot of text is probably gathered. Since the output from a text mining process results in a large amount of information, a good user interface is required. The user must be able to navigate the material easily and, if necessary, modify the original question if the result is unsatisfactory (Tanabe et al., 1999). By making a visualization of the output a better understanding is gained (Freeman, 1998).

2.1.6 Developing the model

To further develop the model it is important to predict additional cases and then examine the results. If the model is used in companies this may require building computerized systems that capture the appropriate data and generate a prediction in real time (Freeman, 1998).

2.1.7 Monitoring

After the model is finished it has to be monitored. New data is published constantly, which could change earlier results predicted by the model. Therefore a constant validation of the model is necessary to maintain a correct prediction. As Ekberg et al. (2000) says a model that is correct today may not be appropriate tomorrow.

2.2 Publicly available data sources

According to Elmasri et al. (2000):

"A database is a collection of related data",

By data Elmasri et al., (2000) here means:

"known facts that can be recorded and that have implicit meaning" (ch 1, pp4, Elmasri et al., 2000).

A database is a representation of some aspects of the real world. Elmasri et al., (2000) talks about a miniworld, and as new changes are made in this miniworld these changes are also seen in the database. The data in a database is coherent and has an inherent meaning, and the database also has a specific purpose within an intended group of users (Elmasri et al. 2000). The National Center for Biotechnology Information (NCBI, 2002) and the Weizmann Institute of Science (Weizmann, 2002) are maintaining several biological databases of which a few, related to this project, are described below.

2.2.1 PubMed

PubMed is a database that was designed to provide access to full-text articles at journal web sites and other related web resources. PubMed was developed by the National Center for Biotechnology Information (NCBI, 2002). A link feature is available that gives access to the journals in full text format. PubMed has also links to the other Entrez molecular biology databases and to the author's web page when this is available. The bibliographic information that PubMed gives access to is comprehensive: the medical database MedLine, out-of-scope citations, citations that precede the date that a journal was selected for MedLine indexing, and finally some additional life science journals which submit full text to PubMed Central and which receive a qualitative review by NLM (PubMed, 2002).

2.2.2 MedLine

MedLine is a database of more than 11 million bibliographic citations in over 4600 journals. MedLine covers the fields of medicine, nursing, dentistry, veterinary medicine, health care systems, and preclinical sciences, as well as additional life science journals since 1966 and new material are incorporated weekly. All the citations in MedLine are assigned subject headings from the National Library of Medicine MeSH (PubMed, 2002).

2.2.3 GeneCards

GeneCards were developed by the Weizmann Institute of Science and Crown Human Genome Center (Weizmann, 2002). GeneCards is a database that integrates some of the resources for human genes i.e. their gene names, their products and their involvement in diseases. The database is accessible through the Internet site of GeneCards (Rebhan et al., 2002).

2.2.4 Online Mendelian Inheritance in Man (OMIM)

The OMIM database is a catalog of human genes and genetic disorders. It was developed for NCBI and contains textural information, pictures and reference information (Brylawski, 2002).

2.2.5 The human genome organization (HUGO)

HUGO is maintaining a database over the human genome. The database is called The Genome Database (GDB). The Department of Energy and the National Institutes of Health is coordinating the human genome project. Some of the goals for the project is to identify all the approximately 30 000 genes in human DNA. In 2000 a sequence draft had been developed, one year ahead of plans, and now the process is to fill in the gaps. By 2003 they aim to provide a complete and high quality DNA reference (HUGO, 2002).

2.4 Gene regulation

Genetic networks can be considered as the interactions of biological macromolecules and the flow of regulatory information that controls development, behavior and homeostasis (Loomis et al., 1995). A set of pathways, produced by sequenced genes that interact with other pathways, constructs a genetic network (Altman, 2000). The nodes in the network represent the genes (or their products) and the interactions between the genes represent the regulatory and physical interactions among the RNA proteins and cis-regulatory DNA sequences (Loomis et al., 1995) The simplicity of a Genetic network is as Loomis et al. (1995) implies not as simple as for an ordinary network like a telephone network for example. If any of the interactions in such network breaks down, the whole structure will not fall, because the different wires are independent of each other. A difference is made to the genetic network that has been developed during years of evolution. Studies have been made on the effect a mutation has to a phenotype. The in vitro mutations can be an alternation or deletion of a specific gene. The results from the studies indicate that the loss of a specific gene can alter the phenotype. If a small mutation occurs a difference in phenotype is not significant or does not change the phenotype at all. An example of a very simple

network is illustrated in figure 2. If one of the genes were knocked out, the networks topology would fall apart (Loomis et al., 1995).

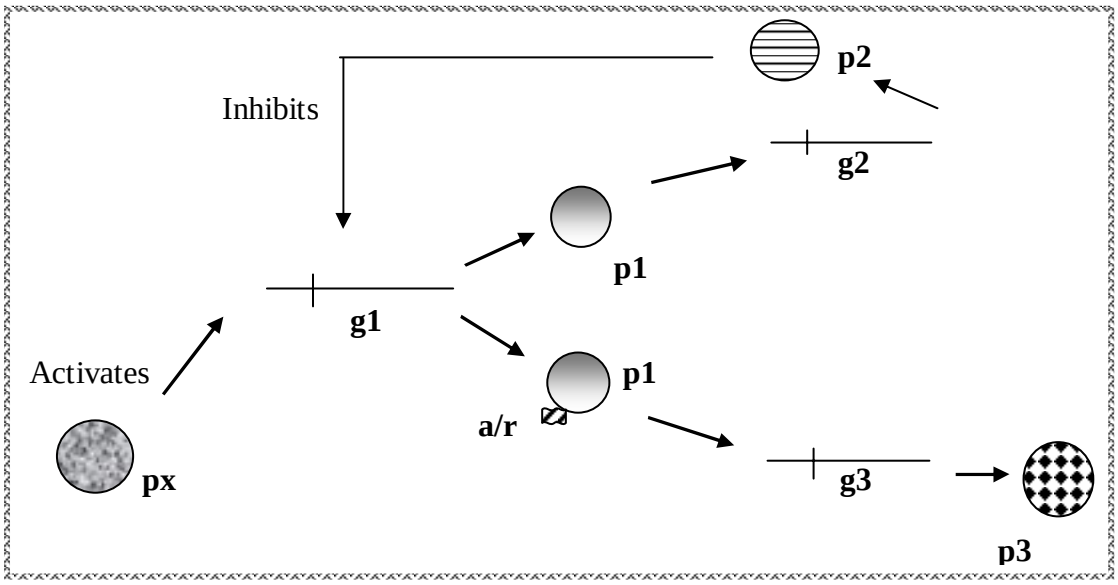

Figure 2. A schematic example of a very simple regulatory network. Genes are represented with g1, g2, g3, and proteins with p1 and p2 and activator/ repressor with a/r. A transcription factor in the form of a protein activates gene g1; the gene product from g1, i.e. p1, stimulates both g2 and g3. The protein product from g2, i.e. p2, inhibits the activity of g1 meanwhile g3 transcripts the final protein p3 that probably leaves the cell.. One of the protein p1 interacts with an activator or repressor to activate or repress the gene 3.

2.4.1 Regulatory mechanisms

The regulation of transcription to produce mRNA and finally a working protein is very complex in eukaryotes. A class of proteins called transcription factors mediates the transcriptional regulation. These transcription factors bind to specific DNA regions, called cis-acting elements, and interact with RNA polymerase to enhance (activate) or reduce (repress) the expression of neighboring genes, see figure 2 (Weaver et al, 1999). Unlike prokaryotes, which utilize a single RNA polymerase to synthesize all the RNA in the cell, the eukaryotes have three different nuclear RNA polymerases: RNA polymerase I, which makes ribosomal RNA; polymerase II, that makes structural genes and genes for snRNA and finally RNA polymerase III transcribes tRNA genes and several other genes which in turn encode small cellular RNAs (Elseth et al., 1995, Weaver et al., 1995).

Often you can divide RNA polymerase II transcription factors into two sometimes-overlapping categories:

- General transcription factors

  General transcription factors are essential for initiation. They respond to specific DNA sequences that are common to most (all) promoters that RNA polymerase II recognizes.

- Regulatory transcription factors

  Regulatory transcription factors are not required for initiation. They react to specific sequences in certain promoters' and/ or are active in certain types of cells (Elseth et al., 1995).

2.4.2 Genetic regulatory network

It is commonly known that many of the acting genes in eukaryotes and prokaryotes encode transcription factors that are part of a network of regulatory interactions (Benfery et al., 2001). The transcription factors that regulate gene activity are themselves regulated by other gene products via posttranscriptional mechanisms (Weaver et al, 1999). The groups of proteins that regulate gene activity are functionally independent of each other (Noveen et al., 1998). These groups of genes regulating the activity of each other are known as genetic regulatory networks see illustration in figure 3. In such a network the proteins and their interactions with other material (other proteins, aminoacids) are not illustrated and thereby a more schematic picture is gained in difference from the regulatory network illustrated in figure 2. Every genetic regulatory network regulates a specific function as for example, the length of an organ (Noveen et al., 1998). One genetic network can also regulate several functions and are called multifunctional networks (Noveen et al., 1998). From their experiments Noveen et al., (1998) have drawn the conclusion that a particular gene may be involved in different networks and that these particular genes have different functions dependent on which network it is. In the same network though, the gene has a single and constant function throughout the time it is being expressed (Noveen et al., 1998). An important concept in genetic regulation networks is gene circuits. A feedback circuit (or feedback loop) is defined as a circular chain of

interactions; each element in the circuit has a direct or indirect effect on itself. The circuit is said to have a positive feedback when the effect on itself is positive, and the reverse the circuit is said to have a negative feedback when the effect on itself is negative.

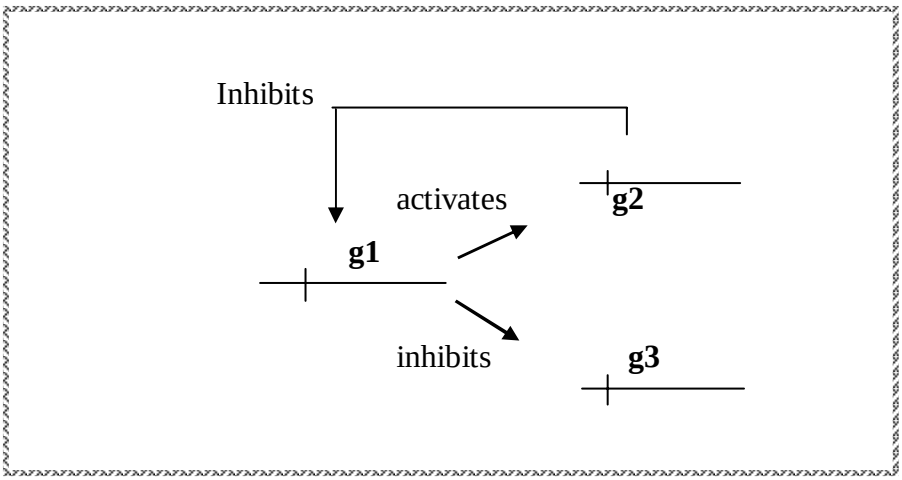

Figure 3. A schematic example of a very simple genetic network. Genes are represented with g1, g2, and g3. The gene g1 regulates (activates respective inhibits) the genes g2 and g3 and then g2 inhibits gene g1. Compare this network to the one in figure 3. In a genetic network as a difference from a genetic regulatory network the proteins are not illustrated, i.e. you can not see how the proteins interact with other proteins or aminoacids to control the genes.

In this project a Genetic network is going to be created using the information extracted by the text mining algorithm. The algorithm uses information from a database available on the Internet.

3 Related work

There are several text mining tools available for screening the literature for specific words or sentences. The works that are described in this project come from (in order of appearance): Tanabe et al. (1999), Dickerson et al. (2001), Blaschke et al. (1999) and Fukuda et al. (1998). The simplest of the methods are looking for co-occurring gene names within abstracts, while more sophisticated methods tries to identify those genes that often co-occur within documents. Then an examination of the sentences describing both of the genes is made to discover relationships between them (Chang et al., 2001).

Work has also been done on text mining manually to create a genetic network; the one described in this project is the work of Mendoza et al. (1998) and their regulatory network of *Arabidopsis thaliana.*

3.1 MedMiner

Tanabe et al. (1999) presents a system MedMiner that filters and organizes the literature by searching and querying multiple databases like PubMed and GeneCards. The system searches documents from PubMed and GeneCards for relevant facts that are specific to a predetermined domain. MedMiner consists of a list of dozen relationships words. These words are being used together with two terms that are specified by the user, these terms could be names of proteins etc. MedMiner has text filtering which make it possible to translate user queries into relevant metrics e.g. sentences that could be relevant to the problem. These metrics are then used to query the large amount of data. MedMiner uses two approaches for text filtering: The first is to apply combinations of keywords to identify relevant documents, paragraphs or sentences and the second is to use word frequencies to determine the relevance. MedMiner has also a carefully designed user interface, which allows the user to navigate the material easily, and the user is able to optimize the results by modifying their queries. The output from MedMiner is organized by relevance, which makes browsing more logical and efficient. The different steps are defined in figure 4.

The studies of Tanabe et al. (1999) have been applied to examine the correlations between the activities of different drugs and gene expression primarily gene-gene relationships observed in mRNA expression profiling experiments with cDNA microarrays and oligonucleotide chips. One must specify keywords in order to filter the results.

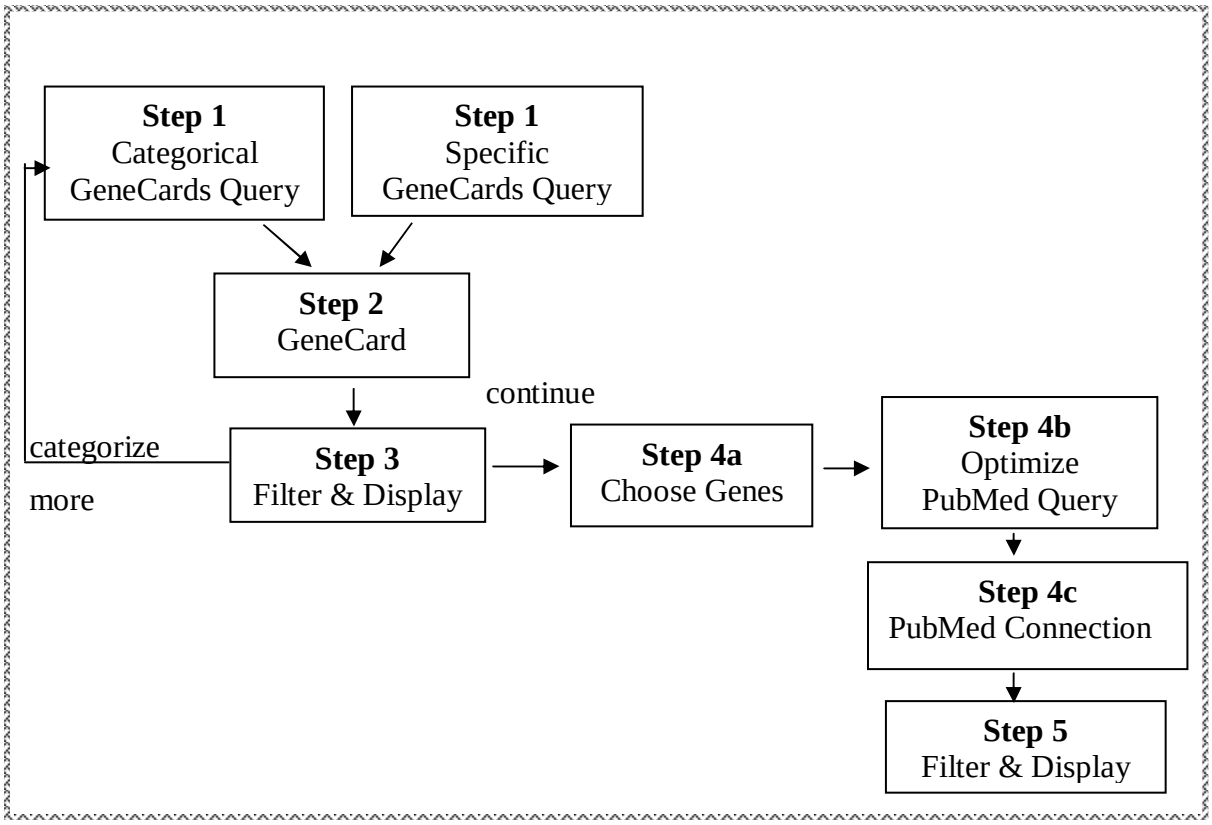

Figure 4. A schematic picture over how the MedMiner system works. First the genes or a concept of interest are specified by the user. This is then sent as a query to the GeneCards database. MedMiner then processes the result from the GeneCards entry. The process is then repeated for each of the two genes entered in step 1. The returned list is filtered automatically by matching the retrieval against the local gene database, which consist of the set of genes on an array, this to exclude any unsatisfactory results. The list can also be filtered manually. A PubMed query is then formulated and submitted. Finally, the user filters the results by applying a combination-of-keywords method to the titles and abstract, and the citations that pass the relevance filter are grouped according to the particular relevance rules triggered.

A schematic view of how MedMiner works is presented in the following steps:

Step 1. The user specifies the genes of interest. This can be done in two different ways, either by specifying specific gene names (e.g. gene names located on a cDNA microarray or oligonucleotide chip) or by a general concept that can be used to find genes (e.g. apoptosis). In step 2 the gene names specified in step 1 are sent to the GeneCards database as a query. In the resulting dataset from GeneCards, genes that exist on the user-specified chip are highlighted and the synonyms of these genes are extracted. The genes that are related to the query but are not located on the specific oligonucleotide chip are also highlighted; these genes could be interesting for future chip design. This step is repeated once for each of the genes the user specified in step 1. In step 3 the dataset is filtered and modified manually or automatically. Automatic filtering is performed by matching the retrieval against the local gene database that consists of a set of genes located on an array. After the automatic filtering the user can also filter the dataset manually. In step 4 a, b, and c the gene, drug and/or disorder names are formulated into a PubMed query. The results from the query are the citations from relevant abstracts, and the user can choose to specify a publication date or enter additional search terms to reduce the amount. In step 5 the user filters the dataset by applying the combinations-of-keywords method to the titles and abstracts. The abstracts that contains one of the keywords and a relationship is extracted and grouped according to the relationship. If a sentence contains several relations it will be represented in each of the corresponding groups (Tanabe et al., 1999).

3.2 PathBinder

Dickerson et al (2001) have made a system called PathBinder, which, from MedLine and PubMed, extracts relevant sentences (passages) about protein relationships like those containing terms that indicate relevance to signal transduction or metabolic relationships (Dickerson et al., 2001). In their study, Ding et al. (2002) has suggested that sentences, defined by the text between two punctuations, were significantly better in information retrieval than phrases, defined by the text between any non-word character (e.g. punctuation, comma, colon, semicolon), with respect to effectiveness. Therefore PathBinder rely on the sentence unit rather than abstracts, phrases or other units. In figure 5 the process of PathBinder is described. It consists of a half dozen

steps that ends up by integration with other software to continue the creation of a complete metabolic pathway. First a user edits an input of a biomolecule of interest. Then the synonyms are extracted by accessing the HUGO and OMIM databases and the sentences are extracted. Each URL is then downloaded and the sentences are processed in a more user-friendly form. A graphical presentation can also be created.

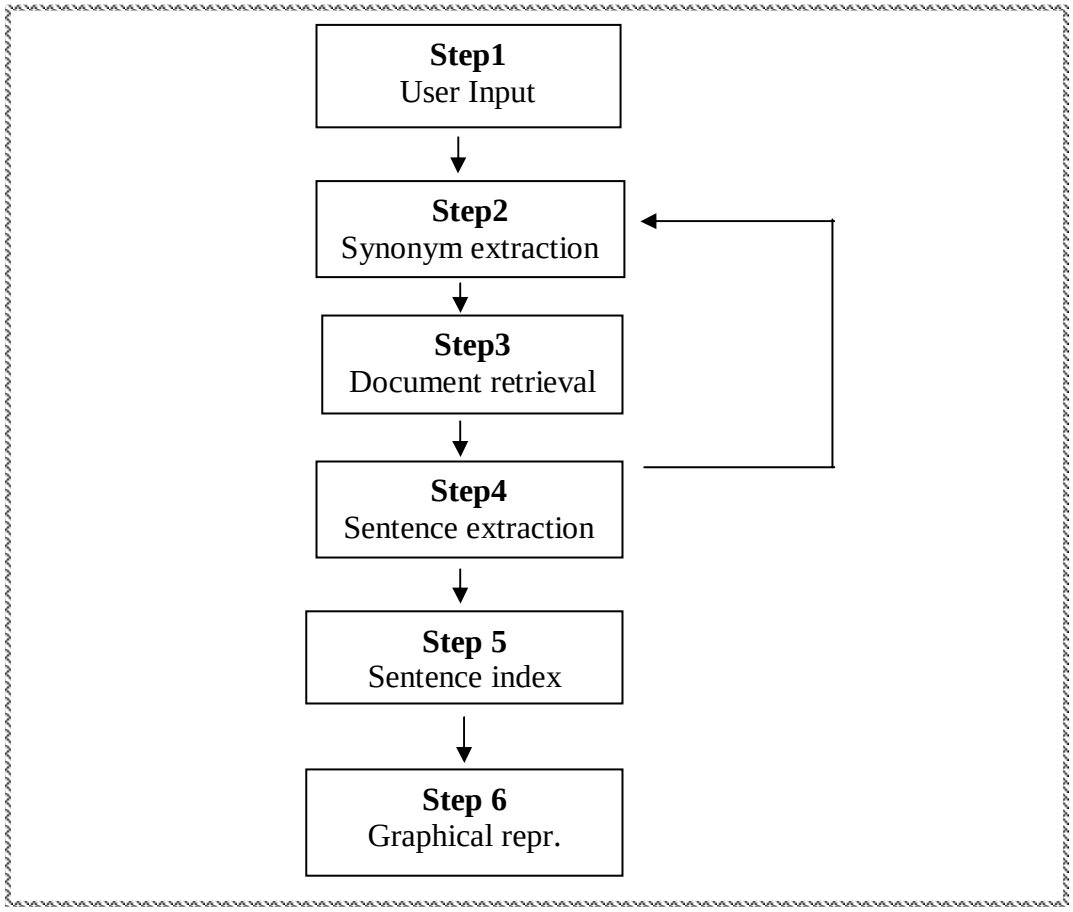

Figure 5. The steps in the PathBinder algorithm. Step 1; the user prints name of biomolecule of interest. Step 2; a user editable file combined with a more advanced module, access the HUGO and OMIM databases and extracts synonyms. Step 3; PubMed is accessed and queried by using the terms in Step 1. The output from this step is a list of URLs with high relevance probabilities. Step 4; each URL is downloaded and scanned for relevant sentences. Step 5; the collection of sentences are processed into a more user-friendly form. Step 6; A graphical representation can be created with the index (Dickerson et al., 2001).

In the first step the user prints biomolecule names that are part of a pathway as input. In step 2 the synonyms for the biomolecule names edited in step one are extracted. This is done by combining a file that the user can edit and a module that automatically access the HUGO and OMIN databases. In step 3 the program accesses PubMed and queries it with the input in step1. The output should, according to Dickerson et al. (2001), result in a list of URLs with high relevance probabilities. The URLs are downloaded and scanned for pathway–relevant sentences containing relevant information, in step 4. From the relevant sequences new biomolecule names are extracted. Steps 2 – 4 are repeated using the new names as input. In step 5, the sentences collected in the previous step are presented as a multi-level index and displayed by a web browser. If a sentence is clicked on, the original document is shown. Finally, in step 6 the index can be used to create a graphical representation where the biomolecule names are connected with the verbs by a line and forming a web–like relationship diagram of the information (Dickerson et al , 2001).

3.3 Blaschke

Blaschke et al. (1999) has also made a system for the extraction of information about protein-protein interactions from scientific journal abstracts available in the MedLine database. Their system is based on the counting of the number of sentences containing protein names separated by interaction verbs. Predefined protein names are used and a number of verbs are made that describes different events. Sentences derived from sets of abstracts will contain a significant number of protein names connected by verbs that indicate the type of interaction between them. The complexity of semantic analysis is avoided by pre-specifying a limited number of possible verbs. The design of the system relies on the peculiarities of this knowledge domain (Blaschke et al., 1999). Their system can be used in the construction of macromolecular networks. The method is briefly outlined in figure 6.

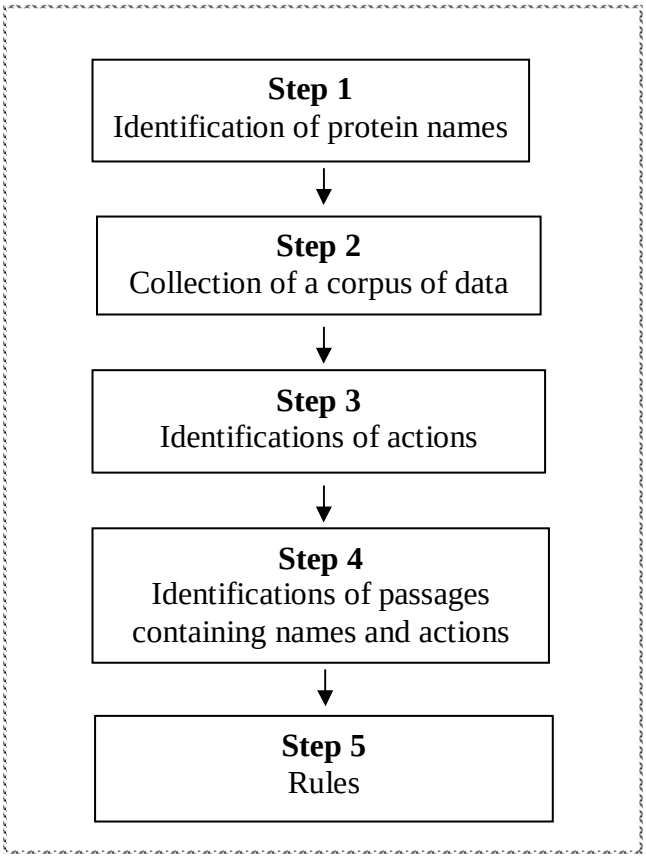

Figure 6 The method that Blaschke et al. (1999) used to extract biological information. Step 1; User prints a protein name. Step 2; The program collects all the abstracts that contains the name, related abstracts are also found using the neighbors utility. Step 3; Identification of different actions is done using a set of 14 predefined words. Step 5; The text is fragmented by using a set of rules. Further description is found in the text.

In step 1 the user prints a protein name. All the abstracts that contain the protein name are collected in step 2. The abstracts can be collected from five different data sets: set 1 consists of MedLine abstracts which directly references from each of the Drosophila SwissProt entries. Set 2 contains abstracts from MedLine's fly Base; set 3 contains abstracts which were collected by adding the neighbor's utility to set number 2. Set 4 contain Medline abstracts, which in the Medical Subject Headings (MeSH) list contained any of the protein names and the word Drosophila. MeSH is a vocabulary consisting of terms or subject headings arranged in an alphabetic and a hierarchical order. Finally set 5 contains MedLine abstracts developed by adding neighbor's utility to the related abstracts collected in set 4. Blaschke et al. (1999) then used the neighbor's utility to search for related abstracts. In step 3 different actions are

identified by using a set of 14 predefined words indicating actions which are related to protein interactions. This list is presented below: (Blaschke et al., 1999).

- acetylat-e (-ed, -es, -ion)
- activate-e (-ed, -es, -ion)
- associated with
- bind (-ing, -s, -s to, /bound)
- destabiliz-e (-ed, -es, -ation)
- inhibit (-ed, -ing, -s, -ion)
- is conjugated to
- modulat-e (ed, -es, -ion)
- phosphorylat-e (-ed, -es, -ion)
- regulat-e (-ed, -es, -ion)
- stabiliz-e (-ed, -es, -ation)
- suppress (-ed, -es, -ion)
- target

In step 4 the passages that contained both the protein names and the action identified earlier are extracted from the original text. In step 5 the text is fragmented using grammatical separations that segment the text into phrases. A phrase is the text between special characters e.g. dots or comma. The different text segments are separately dealt with by applying a series of simple rules based on protein / verb arrangement and proximity. This was made by selecting those text fragments that contained at least two protein names and one action verb. The easiest construction to interpret is “protein A – action – protein B”, there are also other variants “action – protein A – protein B” which are more difficult to interpret because they need a future extension that remembers the action before the protein names. A grammatical separation is difficult to use when an interaction is described over a longer passage such as:

“…in wild oat aleurone, two genes, alpha-Amy2/A and alpha-Amy2/D, were isolated. Both were shown to be positively regulated by gibberellin (GA) during germination …” (Ding et al., 2002, pp. 2).

Here one must use a variable that remembers the text before the non-word character in order to capture the content i.e. that Amy2/A and alpha-Amy2/D both were regulated by gibberellin (GA).

3.4 Proper

Fukuda et al. (1998) presents a method to extract material names by using surface clue on character strings in medical and biological documents. Surface clue is when one examines the contexts of the word to determine if it the word is relevant. Their method does not require any specific term dictionary prepared in advance and has the same accuracy for words regardless for when it was defined. They deal in their report with the problem of the different functions of a protein i.e. one protein can have a lot of different functions and be spelled in different ways.

Their method extends to extract all words in the document that could be protein names in five different steps. The words are extended to adjacent words or other annotations to develop whole sentences. Then the dependencies between blocks are rebuilt by different rules. Improper rules and annotations are excluded. PROPER, as their system is called, uses the characteristics of proper noun descriptions and does not require any specific term dictionary prepared in advance. It extracts material names from the sentence with high accuracy regardless of whether it is already known or newly defined and whether it is a single word or compound word. As a result, their method can completely correspond to the variation in expression. The classification of proteins is used. Below in figure 7 is a schematic picture over their process.

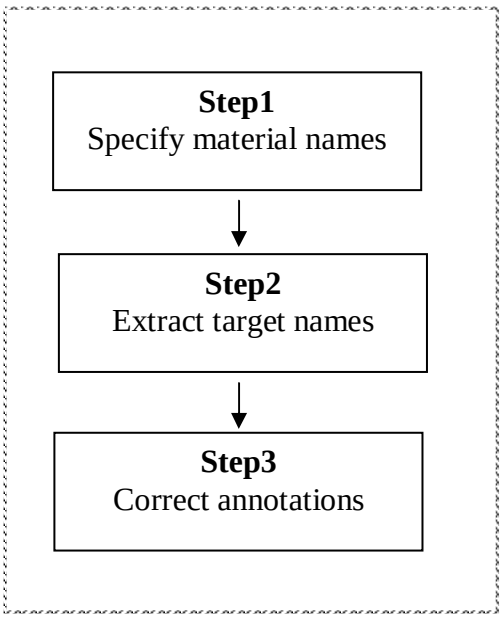

Figure 7 Step1: specify the material names, these could be protein names or domain names. Step 2: Extract the target names. Step3: Correct annotations.

In the first step, the user specifies what kinds of material names to be printed. These material names could be protein names in which kinas, receptor, ligand, enzyme and compound are also included. In step 2 the material names are extracted. This is made by first using a core term extraction from text that has been divided into tokens and then rebuilding the core blocks and the dependencies. The core term extraction is divided into 5 different steps where different words are included or excluded in the sentence dependent on whether the word has upper cases, numerical figures, are too long, have special symbols etc. The extracted core terms are annotated in the text. The rebuilding is then made by extending the core terms using adjacent words to the core term concatenated with other annotations. The result of this process is the noun-phrases without conjunctions and prepositions restored. The rebuilding of the core terms is then made by using two concatenation rules, the first rule is surface clue when the terms are connected if the terms are adjacent to each other and in the second a Part-of-speech (POS) tagger are used. A POS tagger connects a longer sentence if it lies between two core terms or if there is a determiner of preposition before the core term. It can also connect if it is a single upper case letter or a word representing Greek letter. In the third step different methods that correct wrong annotations are used to exclude improper annotation. The first of the two rules they applied extract those words that remained as a single word. This could happen if the word is very ordinary. Ordinary are for example the words that describe a relation between two objects. These words are Fukuda et al. (1998) refereeing to in their report as feature terms or f terms. The second rule is applied when the last word of a rebuild sentence is not a noun, this will happen in the case when the core term is not a noun as in the case of “Src-related”.

4 Thesis statement

Earlier work in text mining has a very narrow focus on either very specific target areas, for example human proteins, or a too broad focus with no abilities to focus on specific terms. Most of the methods available focus on gene-to-gene interaction without the possibility to view information from text sources like MedLine for only one gene. The viewing of only one gene could be useful in order to find new unexpected data, which might not be true for both of the two genes specified. The use of text mining methods for constructing gene-networks would be time saving and give better results. This as the new data will be easy to view when only the important parts are shown and more data can also be collected during the same time range, which would give a more reliable result. The network created by the algorithm can be used in science as a valuable source to view how genes are interacting with each other. One can in an easy and fast way get an overview how the genes are connected to each other and in what way the interaction is changed when new data is available.

4.1 Aim and objectives

The aim of this project is to develop a method to derive a genetic network using text mining methods.

A text mining method is going to be developed that verify a genetic network developed manually. Building a new network tests the method to see if it is manageable to construct a network over other genes than the network verified.

To obtain this aim the following objectives are stated:

Selecting data sources. Looking at the characteristics and the capacity of the database identifies the database/databases that are going to be used. To be able to create the network the database / databases must contain necessary information e.g. published articles or other reviewed scientific information that includes the studied area.

Identification of an interesting network. Searching the literature for known gene interactions will identify the network that is going to be used. The network should have been developed by manually searching for genes and their interactions in the literature e.g. available on the Internet. This is important in order to be able to verify the network.

Develop a text-mining algorithm. A new algorithm is created by guidelines of earlier work and new ideas. The algorithm is going to search databases for articles and abstracts for the specific gene of interest. The goal is to create a network of specified genes by using the literature collected.

Evaluation of algorithm. The text mining algorithm is evaluated by verifying the network identified in the second objective. The algorithm is presumed to have the correct function if the text mining algorithm correctly predicts the genetic connections identified in the second objective.

Creation of a new network using the algorithm. Deriving of a network by using the text mining algorithm should be done on a new set of genes, this to test the system if it can create another network. The genes on which the network is being made have to be identified to be able to use the algorithm.

Analysis of the results. The results are analyzed by examine the relationships between the genes in the model network. To see if the information is sufficient to create a network. An examination of the correctness of the relation is made by looking at the abstract for the specific relation to see if there is something in the text that is contrary to the result developed by the algorithm. For example negative sentences, implicit information or if the specified gene names can have other meanings.

4.2 Problem definition

The problem definition is that a genetic network can be derived using a text mining algorithm. The algorithm extracts all the information necessary to derive the network from public data sources. The network is equal to a defined network, by a manual reading shown able to be recreated from the text in a defined data source used when creating algorithm. All interactions that exist in the network developed by Mendoza et al (1998) but do not exist in the network developed by the algorithm, are then considered as false.

4.3 Demarcates

The prerequisites in this project are that all the genes for which a network is being created are given and that not any new genes are added on the basis of the result from the program. All the abstracts that are going to be searched by the program are accessible from a text file. If the network has the connections A is regulating B and B is regulating C se figure 8. And the algorithm finds the relations that A is a regulator of C, and that B regulates C. It is not presumed that A is regulating B on the basis of the results.

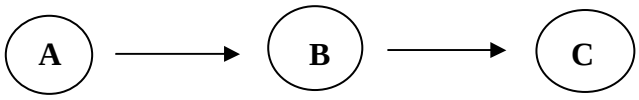

Figure 8. A figure of how dependencies can develop. In this network A regulates both B and C and B regulates C.

4.4 Difference from existing work

In this study it has been clear that a lot of work has been done in the area of text mining. The approach in this work is to create a network over known genes to be able to extract valuable information. A network over known gene interactions has not been developed earlier from the text mining methods available, but the algorithm to extract the gene interactions is similar to existing approaches. For example Tanabe et al.

(1999) and Dickerson et al (2001) have both developed methods that use a text mining method to extract interesting abstracts from PubMed. Blaschke et al., (1999) have developed a text mining method that identifies proteins and their relationships in abstracts from MedLine. In this project the algorithm uses an interactive list where the user can change the relationships depending on the resulting sentences extracted by the Perl program. By doing this, one can also create a network over proteins, which has other relationships involved than genes. In this project, the abstracts are also collected from PubMed which gives a comprehensive covering of different organisms since the database has publications from many different magazines. I am further creating a view system that makes it possible to view sentences containing the information about a gene; this system will make it easier to find new information about genes. The view system will also be applicable to proteins, aminoacids, chemical substances etc, this because a gene may be affected not only of other genes but also on these different compounds.

5 Method

This chapter gives a description of how the work is performed and the different choices that are made during the process.

5.1 Selecting data sources

Several databases are available directly from the Internet. In this project a database, PubMed, from the NCBI (National Center for Biotechnology Information) is going to be used. NCBI are maintaining and providing access to many of the databases of genes and proteins (NCBI, 2002). The database that is used in this project is specified by three criteria:

- The database should have a comprehensive covering of articles, from many of the different phyla. This is because it is making it possible to create network not only over one distinct category, e.g. the genome over *Arabidopsis thaliana*.
- It must be able to search the database in an easy way, i.e. the database must be accessible to the algorithm, and it must have a way to be searched in UNIX.
- The search results must come in a fashion that is able to process in an algorithm. An example is a page with all the abstracts in text format, i.e. to access the results there is no need to address an URL link and find the result on another page, all the results are collected at one place.

The databases were finally selected looking at their content and ability to show the result in an appropriate way, see section 6.1.

5.2 Identification of an interesting network

The identification of a network was made using published literature available on the Internet. The network has to be manually constructed from the information found in scientific texts. This in order to be able to reconstruct the network by using scientific abstracts from the database identified in section 5.1. All the genes in the network have to be known and all the interactions found between the genes must be described in the

article about the construction of the network. The information about the genes involved in the network has to be available on the Internet or in another format i.e. a text file accessible by the computer, so the Perl program described in 5.3 can access the information.

5.3 Develop a text mining algorithm

The following demands for the text-processing program were in this project stated as:

- The program must be able to process a text file containing abstracts from PubMed.
- There must be a function that makes it possible to print the gene names of interest.
- There must a function to view the relations between the defined genes, and a possibility to change these relations.
- The program must be able to find and extract the information available in the text for all of the gene names printed by the user.
- For each combination of two gene names the program searches for a relation, if a relation is found it is printed on the screen.
- If the program does not find any relation, the abstracts containing the genes are printed and a possibility is given to change the relations.

The algorithm performs a number of steps, presented in figure 9:

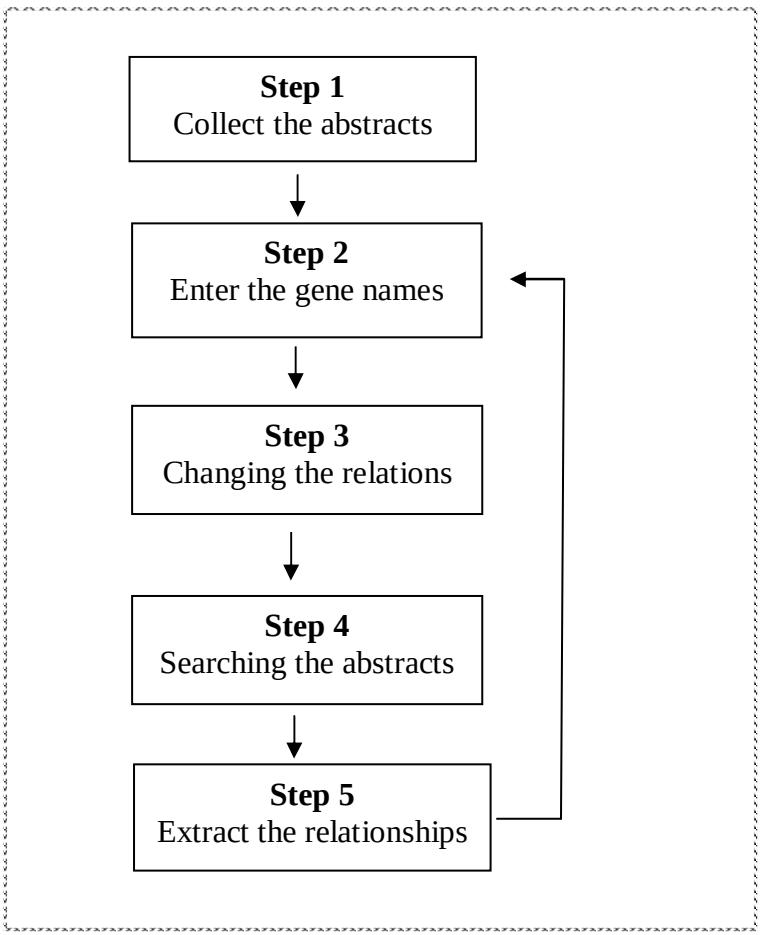

Figure 9. The different steps in the algorithm. A script further described in section 5.3.1 first collects the abstracts, and then the gene names are entered in the program. The abstracts are searched using different text processing commands and finally the relationships found in the abstracts are printed. If the result is unsatisfactory one can enter additional gene names or change the list of relationships.

In the following sections the different steps of the algorithm are described more in detail. Each of the specific Perl commands, found in the implementation, is described in appendix B.

5.3.1 Collect the abstracts

In the first step the abstracts are collected from the data source selected in section 5.1, this is done by using a simple Unix script, i.e. lynx:

```
lynx -dump –nolist http://www.ncbi.nlm.nih.gov:80/entrez/
query.fcgi?cmd=Search&db=PubMed&dispmax=50&term=$1&doptcmdl=Abstract
```

This script collects abstracts containing the genes over which a network should be created, from the database PubMed. The term displmax means that 50 abstracts are going to be collected, the search term $1, which refers to the search arguments i.e. gene names, is entered when the script is being executed and finally the term doptcmdl indicate the kind of document that will be collected from PubMed; in this example the kind of documents are Abstracts.

When executing the script the following line is printed in the UNIX command terminal:

```
script "(A | B )+C" > abstract.txt
```

Using this example of a script, all the abstracts that include the term C and either the term A or B are collected. In the script the terms are the different gene names for which a network should be created. The abstracts collected are then read to a text file named abstract.txt. When all the abstracts are collected the user executes the Perl program see below in section 5.3.2.

5.3.2 Enter gene names

In the second step the Perl program is executed. The program request the user for the gene names to be defined for which a network should be created. When all the gene names are printed the user prints the character q, which defines the end of the input part.

```
Enter a gene name or press q to continue:
```

The same gene names specified in the script in section 5.3.1 are printed. The gene names can be printed as either the whole gene name or as abbreviations (e.g. APETALA2 or AP2). Since the abbreviations can be spelled differently all the possible spellings should be printed. The program then puts the gene names into the array.

```
$number = 0;
while (1){

  print "Enter a gene name or press q to quit:";

  chomp ($j=<STDIN>);
  last if $j =~ /q/;
  push (@genarray, $j);
  $number=$number+1;
}

$numofgenes = $number;
```

5.3.3 Specifying the relationships

When all the gene names are printed, the algorithm reads a file which contains the relationships that the program is going to search, see table 1. A number is given each relation that indicates what kind of relation it is. The relationships are defined by a manual reading of the abstracts for some of the genes in the network by Mendoza et al. (1999). Since these relationships may not be the only relationships that can be found in an abstract between two genes, a possibility is given in the program to further add or delete these relationships.

Relation	Kind of relation		Relation	Kind of relation
inhibits	1		requires	2
regulates	1		relies	2
regulate	1		depends	2
regulating	1		regulated	2
regulator	1		suppressed	2
induce	1			
activate	1			
activity	1			
prevent	1			
require	1			
required	1			
maintain	1			

Table 1 An example of the list of keywords that the program searches. When running the program it is possible to add or delete these relationships. The number 1 stands for a direct regulation a gene while 2 stands for an indirect regulation,

1 stands for a direct regulation a gene while 2 stands for an indirect regulation, like in the sentence; geneA is regulated by geneB. To further explain the numbers on the right of the relationships, two examples are given below: The first sentence illustrates the relations that are defined as 1:

`We have found that AP1, in turn, can positively regulate LFY.`

Here it is clear that AP1 regulates LFY

The second sentence illustrates the relations that are defined as 2.

`AP1 expression in lateral meristems is activated by at least two independent pathways, one of which` *`is regulated by`* `LFY`

In this sentence LFY regulates one of the pathways that leads to the activation of AP1 expression i.e. LFY regulates AP1.

The list of relationships, see table 1, becomes an input to the program. While reading the list, the program creates an array in which each relation and its unique number is stored.

```
$numofrelation=0;
#read the file forhallandearray.txt and add the content to an array

$relations="forhallandearray.txt";
open RELATIONS, $relations;


while(<RELATIONS>){
  chomp;
  push (@relationarray, $_);
  $line = $line + 1;
}
$numofrelations = $line;
$numofrelationsb = $line;

close RELATIONS;
```

The program then prints the array and the user can read the relationships and if necessary, interactively with the program, make changes i.e. print additional relationships or remove relationships.

```
$num = 0;
foreach $relation (@relationarray){ #bc1
print "$num $relation\n";
$num=$num+1

}
```

This may be done when new relations are found which are not given at the relationship list. For specific implementation see Appendix 1. The program then prints the relations found in the relationship file together with a number of the relation. This number is necessary in the following step when the user is asked any changes should be done to the list of relationships. If the user prints the character y, which stands for yes, an option to add or delete a relation is given. To add a relation the user prints the name of a relation together with the number 1 or 2. If it is necessary to delete a relation, the user prints the number of the relation to be deleted.

The next step is to the program to read the file containing the abstracts into an abstract array. One position in the array is one of the sentences. The sentences are defined as the text between two dots.

```
my @array;
{#b1
  local $/ = '.';

open ABSTRACT, 'arabidopsistest.txt';
  @array = <ABSTRACT>;
  close ABSTRACT;
}#e1
```

5.3.3 Searching the abstracts

The abstracts are searched with the key words specified in section 5.3.1 in combination with the relationships defined in section 5.3.2. The program reads the whole array containing the abstracts. If a headline is found the program remembers the line and stores it into a headline array.

```
#if a heading appears it is put in an array
  if ($part =~ m/\n\s*(\[][_][]]\s[\d{*}].*)/){#b3
  $gen1=$1;
  $art=$art+1;
  push (@arti, $gen1);

}#e3
```

The program then iterates through each loop and extracts the results. A flowchart is presented in figure 10 that illustrates this step

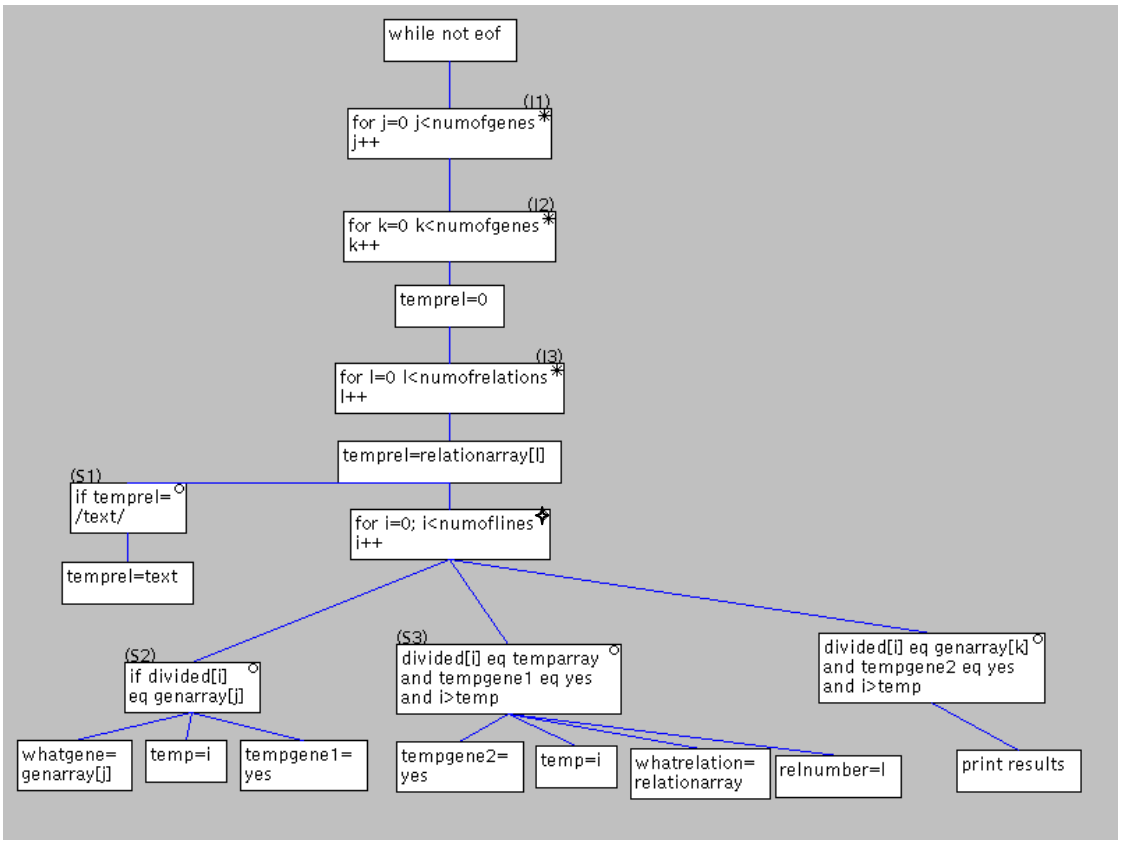

Figure 10 The flowchart over the loops in the implementation. Stars represent iterations, and circles represent choices.

5.3.4 Extraction of the relationships

The relation, the name of the article and the particular sentence that contains the keywords are printed if the relation involves one of the original relationships and two of the genes specified in section 5.3.1. After this step is made the user of the program is asked if the results from the search are satisfying. The user is also asked if changes to the gene names or relation names are necessary. If the results are considered as not satisfactory the user have three alternatives; first to change the gene names by the input of new gene names, second to change (add or delete) the relationships, or third to view the sentences containing the a gene name without any interactions involved. In the first and second choice the relation search of the program is executed again when the changes are made. If the user wished to view the gene names the program runs through the abstracts picking out those sentences that contains the gene names and print those on the screen together with the name of the abstracts that the particular

sentence were found in. Then an alternative to change the gene names or relationships is presented and the user can then run the program again.

5.4 Evaluation of algorithm

The network identified in section 5.2 is then verified using the algorithm. The abstracts are collected separately by searching for the genes of interests. The collection of abstracts was done using the following terms and the script described in section 5.3.1

```
EMF1, TFL1, AP1+LFY, AP1+AG, AP1+EMF, LUG, AG+Arabidopsis, LFY,
CAL+Arabidopsis, AP3+LFY, Ap3+SUP, AP3+PI, AP3+UFO, SUP+PI, UFO
```

Since all the interactions are known, searches were preformed for the above combinations of terms instead of only one term at a time. All the abstracts that are supposed to contain all the information are thereby extracted. In some cases a fusion of two term were made. This was done when to much computer space was demanded for a search on only one of the terms. A manual search of the abstracts was made in order to see if the gene names are found in the text. Then the program was executed using the gene names defined by Mendoza et al. (1998). While executing the Perl program different spellings of the gene name were added in those cases where an interaction was not found.

5.5 Creation of a new network using the algorithm

When the algorithm has been evaluated by validating the network identified in section 5.2 a new network is created using the algorithm. The genes that are used have known interactions but the type of interaction is not described. The genes come from the Yeast Proteome Database, which contains genes from the cell cycle. In table 2 are the genes that were used for creating a new network presented.

BUB1	CDC27
BUB3	CDC6
CDC14	CDC7
CDC15	CHH1
CDC20	MAD1
CDC23	MAD3

Table2 The genes from the Yeastproteom database, used in the algorithm to create a new network.

By using the script described in section 5.3.1, the abstracts are first collected. The abstract search resulted in a file that can be searched by the program. The program was executed and for those genes where an interaction was not found additional searches were made. Sometimes a gene name has several occurrences e.g. BUB1, BUB2,.. BUB5 can be written as BUB1-5. If not BUB1 or BUB2 are found the program is executed again with the occurrence BUB1-5. Since only the abbreviations are given there is no possibility to run the full names in the program.

5.6 Analysis of the results

The analysis of the results is made by studying the sentences found by the algorithm and determine if there is a correct interaction. Since the analysis of the first network was made when verifying the network of Mendoza et al. (1998), only the analysis of the new network is presented here. Consideration is taken if the interactions involve a protein instead of a gene because proteins might be difficult to distinguish in the text. Some of the sentences found by the algorithm can contain incorrect interactions; an example of such a relation is the relation between AP1 and PI in the example below:

```
AG, despite from AP1, is regulated by the gene PI
```

In this sentence AG, but not AP1 is regulated by the gene PI. These kinds of sentences are not correctly predicted by the algorithm. Other sentences demands a more close reading, with consideration of the context of the extracted sentence, to be able to determine if there is a correct or incorrect prediction. The number of correct predictions is presented in relation to the number of total connections found by the algorithm.

6 Results

In this chapter the results is presented. First the data source selected is described briefly. Then the network that is going to be evaluated is presented, the algorithm that has been developed is described and the results from the evaluation of the network are presented. Finally the new network is constructed and the results from the evaluation are presented.

6.1 Selecting data sources

The data source selected in this project is the PubMed database. PubMed has a comprehensive covering of several different fields by maintaining articles from MedLine and other reviewed life science journals. A lot of other data sources are also available on the Internet, some of these, like HUGO and GeneCards and are presented in section 2.2. These data sources give a more detailed description over the studied organism, but are often more difficult to access than PubMed. This because the abstract or the whole article often is connected with URL links from the page with the results one must therefore access another page to be able to read the abstract. This way a search is very time consuming when not all abstracts of interest can be gathered at one time. The PubMed database can be easily searched using the script described in section 5.3.1 and the search result can be entered in a text file and used by the program.

6.2 Identification of an interesting network

The gene network selected for this thesis work is the regulatory network of *Arabidopsis thaliana* flower morphogenesis. *Arabidopsis thaliana* is a well-studied plant and has a lot of available information, including a large amount of articles. Mendoza and Alvarez-Buylla derived the network in 1998 (Mendoza et al. (1998).

6.2.1 Regulatory network for *Arabidopsis thaliana*

In the work of Mendoza et al. (1998) published genetic and molecular data for 11 genes were used that participate in Arabidopsis Thaliana flower morphogenesis. Their network, illustrated in figure 11, was the first genetic regulatory network for a plant, or part of a plant. They analyzed the behavior of the network in order to determine if it was possible to recover the four gene activation states predicted by the ABC model of Arabidopsis thaliana flower morphogenesis. The ABC model is described below. The activation states were recovered and Mendoza et al. (1998) concluded that their network could be correct (Mendoza et al., 1998).

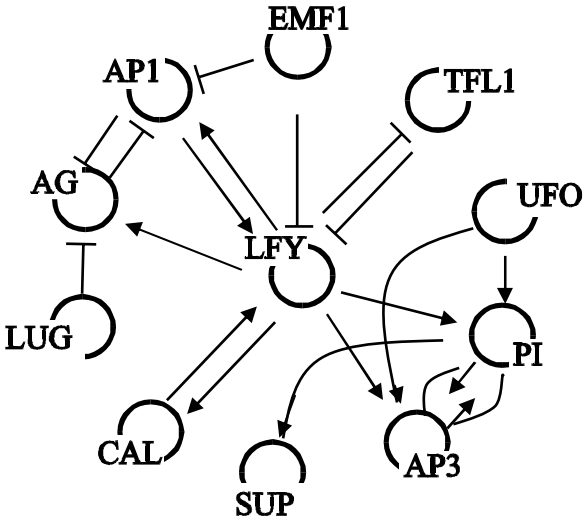

Figure 11 The genetic network over the genes that participate in *Arabidopsis thaliana* flower morphogenesis, derived by Mendoza et al. (1998). The circles in the figure represent genes, and the arrows and the lines represent activation and inhibition, respectively.

The ABC model is a model over the control functions of flower morphogenesis in plants. It is derived from studies of mutant flowers that appear if one of three activities, called A, B and C, is missing. The three activities control different specifications of organ identity in developing flowers. Different combinations of the activities also control different floral organs. In Mendoza's model of flower development in *Arabidopsis thaliana,* APETALA1 (AP1) is the only A function gene, APETALA3 (AP3) and PISTILLATA (PI) have B activity and AGAMOUS (AG) is the only C function gene (Mendoza et al., 1998)

6.3 Develop a text mining algorithm

Once the selection of databases is finished and an interesting network identified the implementation of the text mining algorithm is made. The programming part of the algorithm is presented in appendix 1. The algorithm was created using the programming language Perl. Perl has been chosen mainly because it has suitable text processing abilities, e.g. splitting up a text and the extraction of word from a text (O'reilly, 2002). In this project the text processing abilities in Perl were invaluable when extracting the article's headline, dividing the abstracts into meanings and the meanings into words, and finally matching words against defined genes and relations. The Perl program was used to search the abstract file and the specified relations between the defined genes were found. The Perl program then correctly prints the relations together with two of the gene names and the specific sentence together with the name of the abstract. An example of an output is illustrated below.

```
UFO ----->  PI
[_] 4: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,

          To understand the epistatic relationship among AP1, LFY, and
          UFO in regulating AP3 and PI expression, we generated two
          versions of AP1 that have strong transcriptional activation
          potential.
```

The relation is printed in bold, i.e. UFO is regulating the expression of PI on the first line. The name of the paper, the date of publishing, volume and page number is printed on the second line. Additional words e.g. *Related Articles* and *Books* are sometimes printed on the same line. These words are hyper links in the original paper and have no specific meaning, but are printed with the headline when using the text processing in Perl. Finally the sentence in which the gene names and the relations are found is printed. Multiple abstracts are printed if the same relation is found in several abstracts. After the first search is finished the program can view the sentences containing only one of the gene names, together with the name of the abstract. The program can then once again change the relations if necessary, or add additional relations and the search can be preformed again.

6.4 Evaluation of algorithm network

The network selected in section 6.2 was verified using the algorithm. The abstract search resulted in 470 abstracts from the year 1980-2002. A manual search stated that the collected abstracts contained the information necessary to reconstruct the network of Mendoza et al. (1998). The gene names, used to create the algorithm, are the same as Mendoza et al. (1998) used to construct their network, see section 6.2.1, and the result from the search is found in appendix C. In section 6.4.1 a closer description together with an analysis of the result is found. All but one connection were found and are described more closely in section 6.4.3. All the predictions in section 6.4.1 correspond to the predictions of Mendoza et al (1998). A network was created using the information derived by the text mining algorithm see figure 12.

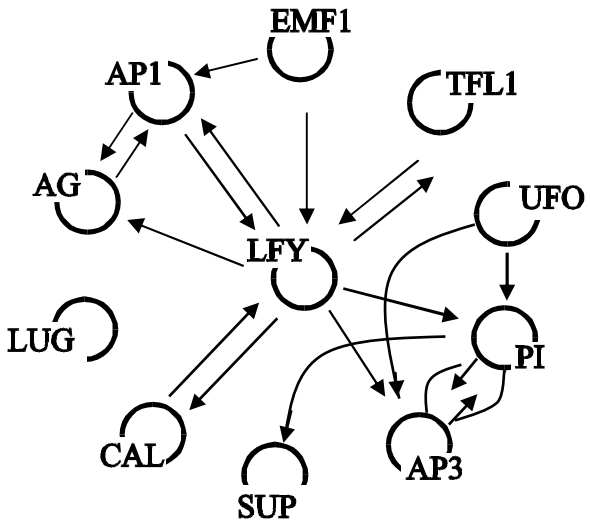

Figure 12. The network developed by the text mining method Notice that it is exactly the same as for Mendoza et al. (1998) except for one relation, the one between LUG and AG. The circles in the figure represent genes, and the arrows and the lines represent activation and inhibition, respectively.

6.4.1 Correct predictions

Below are conclusions of the predictions, by the algorithm presented that corresponds to the interactions found by Mendoza et al. (1998), a closer analysis of the results is found in appendix D.

prediction	relation	Num. of sent.
LFY----> AG	LFY activates AG	2
LFY -----> AP1	LFY is a regulator of AP1	2
AP1 -----> LFY	AP1 can positively regulate LFY.	1
AP1------> AG	Expression of AG relies on functions of LFY.	1
AG -----> AP1	AG inhibits expression of AP1	1
EMF -----> AP1	Expression of AP1 are dependent on the activity of EMF.	1
EMF-----> LFY,	Expression of LFY are dependent on the activity of EMF.	1
LFY -----> PI	PI requires LFY	3
UFO -----> AP3	AP3 requires UFO	6
LFY -----> AP3	LFY regulates AP3.	4
UFO -----> PI,	UFO regulates PI	3
PI ------> AP3	PI activates AP3.	2
AP3 -----> PI,	AP3 regulate PI.	1
PI -----> SUP	PI are required for expression of SUP	1
AP3 -----> SUP	AP3 are required for expression of SUP	1
LFY -----> TFL1	LFY regulates TFL1	1
LFY -----> TFL1	TFL1 regulates LFY	1

Table 3. The correct predictions made by the algorithm. Num. of sent. is the number of sentences in which the relation were found.

6.4.2 Incorrect results

An incorrect prediction was made between AG and the two genes LFY and AP1. This prediction was stated in the following section:

```
Here, we provide evidence that AG function
          is required for the final definition of floral meristem
          identity and that constitutive AG function can promote,
          independent of LFY and AP1 functions, the determinate floral
          state in the center of reproductive meristems.
```

Incorrect predictions are also made when the gene names can be confused with chemical substances, e.g. *DL-AP3* and *DL-2-Amino-3-phosphonopropionic acid* respectively. Two sentences containing this information were extracted:

```
[_] 37: J Neurochem 1992 Nov;59(5):1893-904 Related Articles, Books,
         These studies
          suggest that either L-A beta HA and DL-AP3 bind to a site on
          the receptor and irreversibly block activation of the receptor,
          or that these inhibitors act via a distinct site that
          specifically regulates EAA receptors coupled to PI hydrolysis.

[_] 27: Neurochem Int 1995 Jan;26(1):77-83 Related Articles, Books,
         The
          glutamate metabotropic receptor antagonist
          2-amino-3-phosphonopropionic acid (AP3), the ionotropic
          non-NMDA receptor antagonist
          6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) and the NMDA
          channel blocker dizolcipine (MK-801) failed to prevent the PI
          response to ACPD (1000 microM).
```

At one occurrence a sentence was extracted with the information that the gene regulates itself. The authors have reported that variations in AG (like mutants, in vitro constructed AG proteins) have an impact on the expression on AG. The regulation of AG controlling AG is not happening in nature. An example of such an abstract is found below

```
ag ----->  AG

[_] 58: Plant Cell 1996 May;8(5):831-45 Related Articles, Books,
         In addition,
          transformants with a 35S-AG construct encoding an AG protein
          lacking the C-terminal region produced ag-like flowers,
          indicating that this truncated AG protein inhibits normal AG
```

6.4.4 Additional relations already known

In some cases the algorithm has found relations where A ---- > C when it in the network of Mendoza et al. (1998) is stated that A ------>B and B----- > C. In the result the implicit information is that A---- > C. So the relation found by the algorithm is true correct not additional. This is the case for the relation between AP1 and PI and for the relation between AP3 and PI.

6.4.3 Sentences not found by the algorithm

In the verification of the network of Mendoza et al (1998), one of the relevant abstracts was not found. This abstract is the only abstract containing the information necessary to extract the relation between AG and LUG. When this abstract was not found by the algorithm the hypothesis of this project was not verified. Below is the abstract with the sentences containing the information necessary.

1: Plant Cell 2000 Oct;12(10):1799-810

Separable whorl-specific expression and negative regulation by enhancer elements within the AGAMOUS second intron.

Deyholos MK, Sieburth LE.

Biology Department, McGill University, Montreal, Quebec, Canada H3A 1B1.

We analyzed the 4-kb intragenic control region of the AGAMOUS (AG) gene to gain insight into the mechanisms controlling its expression during early flower development. We identified three major expression patterns conferred by 19 AG::reporter gene constructs: the normal AG pattern, a stamen-specific pattern, and a predominantly carpel pattern. To determine whether these three expression patterns were under negative control by APETALA2 (AP2) or LEUNIG (LUG), we analyzed beta-glucuronidase staining patterns in Arabidopsis plants homozygous for strong ap2 and lug mutations. Our results indicated that the stamen-specific pattern was independent of AP2 but dependent on LUG; conversely, the carpel-specific pattern was independent of LUG but dependent on AP2. These results lead to a model of control of AG expression such that expression in each of the two inner whorls is under independent positive and negative control.

PMID: 11041877 [PubMed - indexed for MEDLINE]

The result of Deynholds et al. (2000) showed that AG is controlled by the gene LUG. This information is implicit in the sentence:

Our results indicated that the stamen-specific pattern was independent of AP2 but dependent on LUG.

In the abstract the authors have defined stamen-specific as expression patterns conferred by 19 AG::reporter gene constructs.

Another relation that is hard to find is a relation were a gene is not defined, e.g. all the EMF genes instead of EMF1. Here one must use the abbreviation EMF in order to find the relation. This can be done by the view function.

6.5 Creation of a new network using the algorithm

When running the program, information for the genes, proteins and their homologues was collected. The program was executed with the genes described in section 5.5: BUB1, BUB3, CDC14, CDC15, CDC20, CDC23, CDC27, CDC6, CDC7, CHH1, MAD1, MAD3. The result from the search is given in appendix E. The connections between the genes are represented in the networks illustrated below. In figure 13 are all the interactions found by the algorithm presented. Those connections that are illustrated with dotted lines are considered as incorrect.

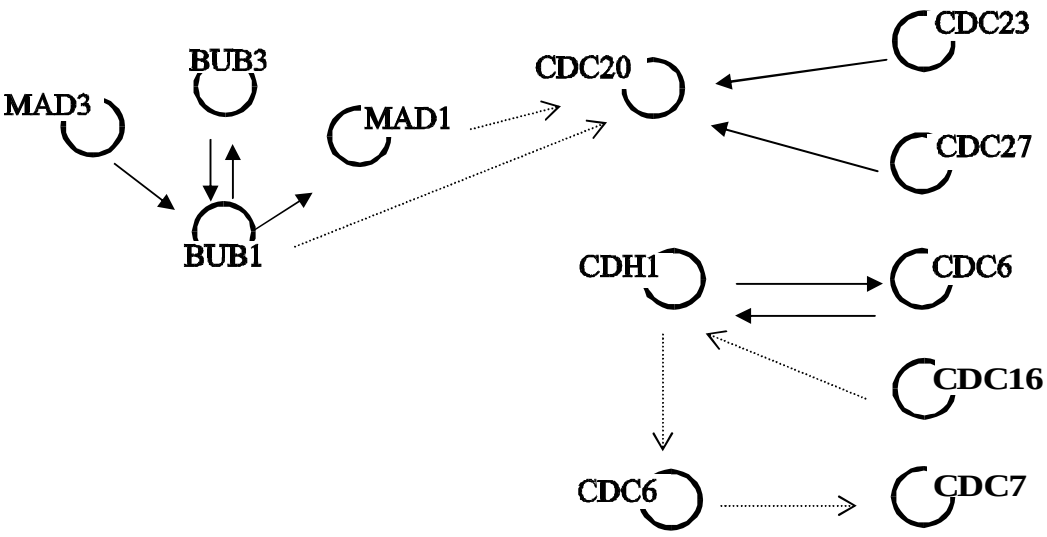

Figure 13.The networks derived by the algorithm. The connections are both between proteins and genes. The circles in the figure represent genes, and the arrows represent an activation or inhibition. The dotted lines are those interactions, found by the program, that are incorrect.

In figure 14 are the same networks presented, with all the incorrect connections between the genes removed.

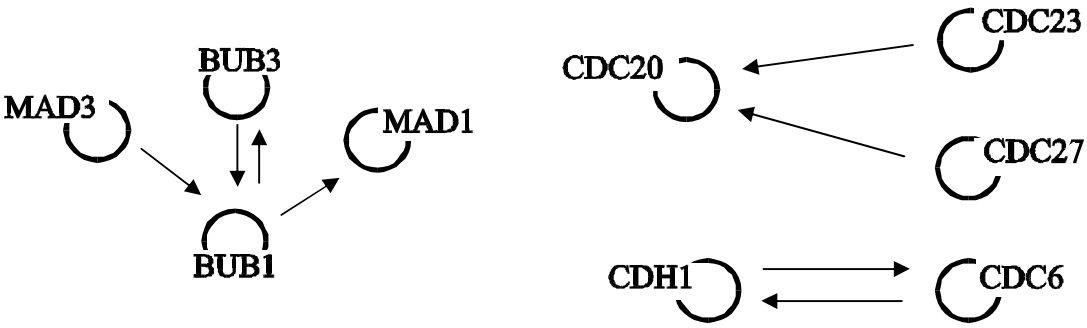

Figure 14. The correct networks derived by the algorithm. Observe like in the previous picture that some of the connections are between proteins and not genes. The program could not find any correct connections between the three networks. The circles in the figure represent genes, and the arrows represent an activation or inhibition.

Between the three networks the program did not find any correct connections, this may mean that they are missing connections or that there is a connection that is not described in an abstract. Three of the correct connections found were already known to have an interaction but not of what kind.

6.6 Analysis of the result

The analysis of the results is presented in table 4 and a closer analysis of the results is made in appendix F. Some of the relationship words have a double meaning, an example of this is the word *required*. Either you can say “A is required for the expression of B” which means that the gene A is necessary for the expression of B, or you can say “A required an expression of B” which means the opposite, namely B is required for the expression of A. This will though be detected when reading the extracted sentences. The number of extracted connections was 13 from these 8 were “correct “, this is a lower number than that from the verification of Mendoza et al. (1998) where only 1 sentence were incorrect. 9 of the 12 genes the algorithm found a connection to another gene.

prediction	relation	Num. of sent.	correct
Bub1 ----- > Mad1	Bub1 is required for kin. localization of Mad1	1	yes
Bub1 -----> Bub3	Bub1 requires Bub3	4	yes
Bub3 -----> Bub1	Bub1 depends upon Bub3	3	yes
Bub1 -----Cdc20	BUB1 homologues inhibit CDC20 proteolysis	1	no
Mad3 ------> Bub1	BUB1 was dependent upon Mad3	1	yes
Bub3 -----> Mad3	some indications that BUB3 regulates MAD3	1	no
Mad1 -----> Cdc20	homologue of MAD1 regulates the CDC20	1	no
Cdc23 ---- > Cdc20	Cdc20 depends on the activity of Cdc23	3	yes
Cdc27---- > Cdc20	Cdc20 depends on the activity of Cdc27	3	yes
Cdc15 ----- > Cdc14	Cdc14 is regulated by proteins including Cdc15	1	no
Cdh1 -----> Cdc6	CDH1 is required for CDC6 proteolysis	2	yes
Cdc16 -----> Cdh1	APC-CDH1 dependent proteolysis of CDC6	1	no
Cdc6 -----> Cdc7	CDC6 is required for the activities of CDC7	1	no

Table 4. The predictions the algorithm made for the new set of genes. Num. of sent. are the number of sentences in which the relation occurred.

7 Discussion

The amount of information on the Internet is today abundant and will probably continue to grow during the next years. A search through this amount of information for some specific text of interests can be made by several ways, often with the aid of search engines like AltaVista, Evreka or Google. Scientists' looking for specific information about genes or proteins may search databases like PubMed or Genbank containing scientific information. This search, using different search engines is often time consuming and results in a lot of text reading necessary to extract the valuable information.

The use of PubMed was in part limiting because PubMed consists only of abstracts and not full text documents. The full text documents are available through links from the abstracts and then the advantages using PubMed for text mining is reduced. If another datasource had been used this would have been a problem when collecting the data since most of the datasources available on internet has the articles available through URL links. The algorithm extracts information necessary for constructing a genetic network without the use of search engines.

The text mining algorithm was partly developed in the program Perl. By this means that all the text processing abilities included in Perl are available. When finding the specific genes in the texts and extracting the relevant parts from the abstracts in PubMed these text processing tools were used.

It was not possible for the algorithm to collect all the interesting abstracts from PubMed at one time for the network of Mendoza et al. (1998), because of the limitations in data capacity. Therefore only the connections between the different genes that made up the network can be found. By examine the result from the verification of the network of Mendoza et al. (1998) other genes were found that are additional to the initial ones. These genes did not contradict the result from Mendoza et al., (1998) but can be used to find additional information about relations. Since the algorithm did not find all of the interactions in the Mendoza et al (1998) network it is not a perfect tool for finding a genetic network, although it may be a very helpful tool on the way to construction networks.

The algorithm was tested on a new set of genes, this resulted in three networks containing not only genes but proteins as well. Even if the network created by the

algorithm did not consist of only genes, it is a valuable source for finding information about the gene printed. The algorithm could be useful when looking for connections between the proteins, aminoacids, and chemical substances. One can also find information about a specific gene or its homologue and if it is part in some process.
A conclusion from this project is that it is not always possible to have a distinct genetic network or a metabolic network. Instead a genetic network must be created that includes both genes and proteins. It is not always so that the scientists are clear in their abstract what kind of substance the author mean. Sometimes the gene name is written when the author in fact mean the name of the protein coded by the gene.

It is not always possible to use the algorithm for developing a network over the defined genes in a list. In some specific cases the scientists have only reported the chemical substances or light dependencies by which the specific genes are regulated. One can in such cases create a list over the regulatory substances that regulate the genes. When executing the program it is necessary to know ahead what kind of substances that are be able to regulate gene expressions. These can be found by using the view function in the program, all the meanings containing the gene name of interest are then viewed despite if it contains a relation or not.

The approach to find genetic network using a text mining algorithm is very good in some cases and not so great in other cases. This can depend on the amount of information in the abstracts. Sometimes there is a lot of information in the abstracts and the algorithm performs well. In other cases, only searching the abstracts is a too narrow approach and no information is given about the relations between genes. A better result will probably be given if entire articles can be searched. If such a search through the entire article on the internet is possible a problem remains on how much one can rely on the information available. The articles can be found on the internet without having being reviewed by a scientific magazine, this means that the article must be examined and a decision must be made on whether the article is reliable or not. In this work this is made by supplying the sentences and a reference to a scientific journal. A possibility should also be given to find the sentences that contradict to earlier results. There are constantly upcoming new results, which can change earlier findings. The connections in the network can be trusted at different levels. Some of the connections are based on sentences found in several abstracts and are thereby considered as more correct than those based on only one abstract.

8 Conclusions

The aim of this project has been to develop a method for deriving a genetic network using text mining methods. The algorithm first collects information from PubMed and then searches through this information in order to find the specific relations between two genes. When verifying the network derived by Mendoza et al. (1998), a manual reading first stated that the abstracts, collected from PubMed, were sufficient to create a network. When the Perl program then was executed all but one connection was found, i.e. the relation between AG and LUG. The abstract containing this relation had implicit information and the time disposed for this project was not sufficient to solve this problem. The problem definition of this work was that it is possible to derive a genetic network using a text mining algorithm and that the algorithm extracts all the information necessary to derive the network from public data sources. Since not all interactions were found by the algorithm it is not considered to correctly predict the network.

However the method may be a helpful tool in this kind of research and a not complete but correct network model is still a result.

8.1 Experiences from the project

An asset in this project when analyzing the results is the previous knowledge in molecular biology. Sentences like “the gene BUB1 had also kinas activity which could autophosporhylate and catalyze the phosphorylation of Bub3” had otherwise required a lot of extra reading. Knowledge gained during the project is how to implement a Perl program. Even though it is not particularly different from the programming language Pascal, it is a lot simpler, and has a pattern matching ability called regular expressions invaluable to this project. These regular expressions were tremendously difficult to learn and are one of the few difficult parts in the Perl language. Lots of information are as fortunate available on the Internet on how to write both a Perl program and about the specific parts like pattern matching.

8.2 Future work

In this project a text mining method has been developed that can be used when creating a network. The incorrect predictions made by the program should, if time had not been limiting, been eliminated. To solve this problem one could include a list of words that should not be included in the sentence, like for example the words *independent* and *not*, and then the program is executed, searching for these words in the extracted sentences. Another problem that would be dealt with, if more time were given, is the problem about sentences that included over a longer passage, this was not a problem in the verification of Mendoza et al. (1998) network, since there were no relations not found by the program that included such a passage. But in other networks there may be relations that are not covered by another sentence and where such relations in that case not are found. When solving the problem one could include in the problem a variable that remembers the sentence before. If the next sentence begins with certain words like *both* or *these*, the gene names included in the sentence before are considered as belonging to the sentence after.

In the future a saving function may be included so that a search for a particular gene only is made one time. The next time the user wants to have some information about the gene the result is already saved and perhaps indexed. This makes the system faster and specific information about a gene can be found without any considerable computational cost.

9 References

Altman R., Genetics Networks, (2000), *Representations and Algorithms for computional molecular biology*, Stanford instructional television network, Available through Internet: http://s-star.bic.nus.edu.sg/downloads/lecture5/asf/slides/sld004.htm [collected 02-04-02].

Altman R. B. and Raychaudhuri S., (2001), Whole-genome expression analysis: challenges beyond clustering, *Current opinions in Structural Biology*, 11, 340-347.

Benfey P. N., Weigel D., (2001), Transcriptional Networks Controlling Plant Development, *Plant Physiology*, 125, 109-111

Blaschke C., Andrade M. A, Ouzounis c. and Valencia A., (1999), Automatic extraction of biological information from scientific text: protein-protein interactions, *Intelligent Systems for Molecular Biology,* Heidelberg, Germany - AAAI Press, pp 60-67.

Brand E. and Gerritsen R., (1998), *Data Mining and Knowledge Discovery*, Miller Freeman, Inc., available through internet: http://www.dbmsmag.com/9807m01.html [collected 02-05-06].

Brylawski B., (2002), *OMIM™ Online Mendelian Inheritance in Man*, availiable through internet: http://www.ncbi.nlm.nih.gov/Omim/, [collected 02-04-27]

Chang J. T., (2001) Towards Incorporating Scientific Literature into Biological Algoritms, Presented at *9th International Conference on Intellegent Systems for Molecular Biology* Copenhagen, Denmark, 21 July 2001, 125-130.

9 References

Dci, (1999), Text mining not DataMining, *DataWarehouse report*, available through internet: http://datawarehouse.dci.com/Articles/990316mining.htm [collected 02-05-11]

Dickerson J. A., Berleant D., Cox Z. and Qi W., (2001), Creating Metabolic Network Models using Text Mining and Expert Knowledge, *Atlantic Symposium on Computational Biology and Genome Information Systems & Technology*, Durham, N.C., USA, 15 mars 2001.

Elmasri R. and Navathe S., (2000), *Fundamentals of Database systems*, 3ed, Addison-wesley

Ekberg R., Nilsson J. and Hermansson M., (2000), Bättre beslut med data mining? *Metoder - Tekniker – Processer*, Högskolan Trollhättan/Uddevalla, availiable through internet: http://www.exjobb.udd.htu.se/2000/sv3/17/ [collected 02-05-19]

Elseth G. D. and Baumgardner K. D., (1995), *Priciples of Modern Genetics*, West Publishing Company.

Fukuda K., Tsunoda T., Tamura A. and Takagi T., (1998), Toward Information Extraction: Identifying protein names from biological papers, *Proc. of Pacific Symposium on Biocomputing*, Maui, Hawaii, 4-9 January 1998, 3:705-716.

HUGO, (2002), Human Genome Organisation, *U.S. Department of Energy*, available at: http://www.ornl.gov/hgmis/, collected [02-05-05]

Kurhekar, M. P., Adak S., Jhunjhunwala S. and Raghupathy K., (2002), Genome wide pathway analysis and visualization using gene expression data, *Pacific Symposium on Biocomputing*, Lihue, Hawaii, 3-7 January 2002, 7:462-473.

9 References

Loomis W. F., Sternberg P. W. (1995), Genetic networks, *Science*, vol.269, p649

Mendoza L. and Alvarez-Buylla E. R., (1998) Dynamics of the Genetic Regulatory Network for *Arabidopsis thaliana* Flower Morphogenesis, *Journal of theoretical biology*, 193, 307-319

NCBI, (2002), *National Center for Biotechnology Information*, available through internet: http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html [collected 02-04-22]

Noveen A., Hartenstein V. and Chuong C-M, (1998), Gene Networks and Supernetworks: Evolutionarily Conserved Gene Interactions, *Molecular Basis of Epithelial Appendage Morphogenesis*, 9-15-98 editors: Cheng-Ming Chuong

O'reilly, (2002), *Perl.com the source for Pearl*, O'Reilly & Associates, Inc, available through internet: http://www.Perl.com/ [collected 02-05-06]

PubMed, (2002), *Overview*, U.S. Government, available through internet: http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html [collected 02-04-03]

Rebhan M., (2002), *GeneCards*, Weizmann Institute of Science, available through internet: http://bioinformatics.weizmann.ac.il/cards/, [collected 02-04-27].

Tanabe l., Scher U., Smith L. H., Lee. J. K., Hunter L. and Weinstein J. N., (1999), MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling, *BioTechniques*, 27, 1210-1217.

Weaver, D., C., Workman C., T. and Stormao, G. D. (1999), Modeling regulatory networks with weight matrices, *Pacific Symposium on Biocomputing*, 4-9 January 1999, Hawaii 4:4, AA3.

Weaver R. F. and Hedrick P. W., (1995), *Basic Genetics*, 2ed, US, Wm. C. Brown Publishers.

Weizmann Institute of Science, (2002), *The Weizmann Institute at a Glance*, Rehovot, Israel, available through internet: http://wis-wander.weizmann.ac.il/, [collected 02-04-27]

Yeates S., (2002), Text mining, availiable through internet : http://www.cs.waikato.ac.nz/~nzdl/text mining/ [collected 02-05-11].

Appendix

In the following chapters the appendix are presented.

Appendix A Implementation in perl

Presents the implementation of the Perl program.

Appendix B Implementation details

Presents some of the specific implementation details.

Appendix C Verification results

Present the results from the verification of Mendoza et al. (1998) network.

Appendix D Analysis of the results

Presents an analysis of the results seen in appendix C.

Appendix E new network results

Presents the results from creating a new network.

Appendix F Analysis of the second results

Presents an analysis of the new network creation.

Appendix A Implementation in perl

This appendix presents the implementation of the Perl program.

```
#!/usr/bin/perl
   use Term::ANSIColor;
$number = 0;
while (1){ #ba1
  print "Enter a genename or press q to quit:";

  chomp ($j=<STDIN>);
  last if $j =~ /q/;
  push (@genarray, $j);
  $number=$number+1;
} #sa1
$numofgenes = $number;
$numofrelation=0;
#read the file forhallandearray.txt and add the content to an array
$relations="forhallandearray.txt";
open RELATIONS, $relations;

while(<RELATIONS>){ #bb1
  chomp;
  push (@relationarray, $_);
  $line = $line + 1;
} #sb1

$numofrelations = $line;
$numofrelationsb = $line;
close RELATIONS;
$num = 0;
foreach $relation (@relationarray){ #bc1
print "$num $relation\n";
$num=$num+1

} #sc1

while (1){ #b1

  print "\nIs there any changes to the list of relationships you
would like to do? y/n ";
  chomp ($answer=<STDIN>);

  if ($answer =~ /y/){#b2
    while (1){ #b3
      print "\n Would you like to add an relation or delete? a/d or
press q to quit ";
      chomp ($answer=<STDIN>);

      if ($answer =~ /a/){#b4
      print "Enter the changes or press q to quit:";
      chomp ($answer2=<STDIN>);
      last if $answer2 =~ /q/;
      push (@relationarray, $answer2);
      $line=$line+1;
      }#e4
      if ($answer =~ /d/){#b4
      print "Enter the number of the relation to remove or press q to
quit:";
```

```
      chomp ($answer2=<STDIN>);
      last if $answer2 =~ /q/;
      splice (@relationarray, $answer2,1);
      $line=$line-1;
      }#e4
last if $answer =~ /q/;
$num = 0;
foreach $relation (@relationarray){#b4
print "$num $relation\n";
$num=$num+1
}#e4
}#e3
}#e2
last if ($answer =~ /n/);
}#e1

$numofrelations = $line;


#open a readable file
$out = " out.txt";
open OUT, ">$out";;

$art=-1;

#Read arabidopsis.txt and split when a dot
#print "Enter the name of the text file:";
#chomp ($filsvar=<STDIN>);
@arti=();
my @array;
{#b1
  local $/ = '.';

open ABSTRACT, 'absract.txt';
  @array = <ABSTRACT>;
  close ABSTRACT;
}#e1

while(1){ #b1
#a loop that is made for each part
foreach $part (@array){#b2
$tempgene1="0";
$tempgene2="0";
$temprel="0";

#if a heading appears it is put in an array
  if ($part =~ m/\n\s*(\[][_][]]\s[\d{*}].*)/){#b3
  $gen1=$1;
  $art=$art+1;
  push (@arti, $gen1);

}#e3

#for each sentence extract the words and count the words
@divided = split(/\W/,$part);
$numoflines= scalar(@divided);



for ($j = 0; $j < $numofgenes; $j++){ #b3
  for ($k = 0; $k < $numofgenes;$k ++){ #b4
    $temprel=0;
```

```
    for ($l = 0; $l < $numofrelations; $l++){#b5
      $temprel = $relationarray[$l];
      if ($temprel =~ /(\w*)\s\d*/){#b6
      $temprel=$1;
      }#e6
      $temp=0;
      $tempgene1="no";
      $tempgene2="no";
      for ($i=0; $i< $numoflines; $i++){#b6

      if (($divided[$i]) eq ($genarray[$j])){#b7
        $whatgene = $genarray[$j];
        $temp=$i;
        $tempgene1="yes";
      }#e7

      elsif ((($divided[$i]) eq $temprel) and ($tempgene1 eq "yes")
and ($i>$temp)){#b7

        $tempgene2="yes";
        $temp=$i;
        $whatrelation=$relationarray[$l];
        $relnumber=$l;
      }#e7

      elsif ((($divided[$i]) eq ($genarray[$k])) and ($tempgene1 eq
"yes") and ($tempgene2 eq "yes") and ($i>$temp)) {#b7


        if ($whatrelation =~ 1) { #b8
          print color 'bold';
          print  "\n$whatgene ----->  ", $genarray[$k] ;
          print color 'reset';
          print "\n", $arti[$art], "\n";
          print"\t$part\n\n\n  ";
        }#e8

       elsif ($whatrelation =~ 2) {#b8
          print color 'bold';
          print  "\n$whatgene <------ ", $genarray[$k];
          print color 'reset';
          print "\n", $arti[$art], "\n";
            print"\t$part\n\n\n";
        }#e8
        $relnumber=0;
      }#e7
      }#e6
    }#e5
  } #e4
} #e3
}#e2

print "\nIs the result satisfactory y/n ";
chomp ($answer3=<STDIN>);
last if ($answer3 =~ /y/);

print" \nWould you like to change the genes or the relationships or
view the sentences including a special gene the relationships
involved? (gene/rel/view): ";
chomp ($answer4=<STDIN>);
```

```
 if ($answer4 =~ /view1/){#b2


  foreach $part (@array){#b3

    if ($part =~ m/\n\s*(\[[]_[]]\s[\d{*}].*)/){#b4
    $gen1=$1;
    $artag=$artag+1;
    push (@artii, $gen1);
  }#e4


  for ($j = 0; $j < $numofgenes; $j++){ #b4
    if ($part =~ /($genarray[$j])/){#b5
      print "printing", $genarray[$j], " to file out.txt\n";
      print OUT color 'bold';
      print OUT $genarray[$j];
      print OUT color 'reset';
      print OUT $arti[$art], "1\n";
      print OUT "$part\n\n";
  }#e5
}#e4
}#e3



print" \nWould you like to change the genes or the relationships or
view the sentences including a special gene the relationships
involved? (gene/rel/view): ";
chomp ($answer4=<STDIN>);

}#e2


if ($answer4 =~ /view/){#b2
 print "\nEnter a genename or press q to quit: ";
    chomp ($j=<STDIN>);
 foreach $part (@array){#b3

    if ($part =~ m/\n\s*(\[[]_[]]\s[\d{*}].*)/){#b4
    $gen1=$1;
    $artag=$artag+1;
    push (@artii, $gen1);
  }#e4

    if ($part =~ /$j/){#b5

      print color 'bold';
      print "$j";
      print color 'reset';
      print $arti[$art], "1\n";
      print "$part\n\n";
  }#e5

}#e3
}


if ($answer4 =~ /gene/){#b2
  while (1){ #b3
    print "\nEnter a genename or press q to quit: ";
    chomp ($j=<STDIN>);
```

```
    last if $j =~ /q/;
    push (@genarray, $j);
    $number=$number+1;
  }#e3
  $numofgenes = $number;
}#e2

elsif($answer4 =~ /rel/){#b2
  $num = 0;
  foreach $relation (@relationarray){#b3
    print "$num $relation\n";
    $num=$num+1
  }#e3
  while (1){#b3
    print "\n Would you like to add an relation or delete? a/d or
press q to run ";
    chomp ($answer5=<STDIN>);

    if ($answer5 =~ /a/){#b4
      print "Enter the changes or press q to quit:";
      chomp ($answer6=<STDIN>);
      last if $answer6 =~ /q/;
      push (@relationarray, $answer6);
      $line=$line+1;
    }#e4
    if ($answer5 =~ /d/){#b4
      print "Enter the number of the relation to remove or press q to
quit:";
      chomp ($answer6=<STDIN>);
      last if $answer6 =~ /q/;
      splice (@relationarray, $answer6,1);
      $line=$line-1;
    }#e4
    last if $answer6 =~ /q/;
    $num = 0;
    foreach $relation (@relationarray){#b4
      print "$num $relation\n";
      $num=$num+1
    }#e4
  }#e3
  $numofrelations = $line;

}#e2

}#e1
```

Appendix B Implementation details

if ($part =~ /text/);	A text processing ability, if the part is the same as the text between / and / something happens
($divided[$i])	array example
last if	ends a while loop
use term: ANSIColor;	Necessary to be able to use colors.
open FILE, 'file.txt';	Opends a file to be read, puts the file into a variable FILE that can be processed.
print "line";	prints the word line on the screen
push (@array, $j);	puts the value in $j to the last element in an array
chomp	Removes line endings from all elements in the list
foreach $part (@array){	something is done for each element in @array, $part is refering to the element just beeing processed
local $/ = '.';	splits the text into sentences
print OUT	the answer is printed to a file
($answer=<STDIN>);	The answer that the user prints is added into an variable
#!/usr/bin/perl	The first line in a perl program. It tells the shell that is executing this script where to find the perl binary used to run the program
while (1){	the loop goes on forever
while(<FILE>){	The loop goes through the whole file.

Special characters	
\n	Newline
\r	carriage return
\t	tab
\d	digit (the same as [0-9])
\D	non digit (the same as [^0-9])
\w	a word character (same as [0-9, a-z, A-Z])
\W	a non word character
\s	any whitespace character (\t, \n, \r, \f)
\S	any nonwhitespace character
*	zero or more occurencies of the character
+	one or mor occurencies of the character
.	any character
?	zer or one occurencies of the character
^	the word mathces at the beginning of the line
/[^kalle]/	not the word kalle
/[abc]d/	matches any of the characters a,b or c together with the character d
|	alternation
$	matches the end of a string
\b	matches a word boundary
\B	matches a non word bounary
\g	matches globally (for the entire string)

Appendix C Verification results

The following connections are found:

LFY -----> AG

LFY -----> AP1

AP1 -----> LFY

AP1------> AG

AG -----> AP1

EMF -----> AP1

LFY ----- TFL1

LFY -----> PI

LFY -----> AP3

UFO -----> PI

UFO -----> AP3

PI -----> AP3

AP3 -----> PI

PI -----> SUP

AP3 -----> SUP

CAL------LFY

A closer description over these connections are given in the following pages.

LFY -----> AG

[_] 8: Cell 2001 Jun 15;105(6):793-803

We show that the floral identity protein LEAFY (LFY), a transcription factor expressed throughout the flower, cooperates with the homeodomain protein WUSCHEL (WUS) to activate AG in the center of flowers.

[_] 9: Plant Cell 1997 Mar;9(3):393-408

Expression of AG mRNA in the central region of floral meristems relies on the partially overlapping functions of the LEAFY (LFY) and APETALA1 (AP1) genes, which promote initial floral meristem identity.

LFY -----> AP1

[_] 10: Plant Cell 1999 Jun;11(6):1007-18 Related
Articles, Books,

We present evidence here that AP1 expression in lateral meristems is activated by at least two independent pathways, one of which is regulated by LFY.

[_] 10: Plant Cell 1999 Jun;11(6):1007-18 Related
Articles, Books,

In lfy mutants, the onset of AP1 expression is delayed, indicating that LFY is formally a positive regulator of AP1.

AP1 -----> LFY

[_] 10: Plant Cell 1999 Jun;11(6):1007-18 Related
Articles, Books,

We have found that AP1, in turn, can positively regulate LFY, because LFY is expressed prematurely in the converted floral meristems of plants constitutively expressing AP1.

AP1------> AG

[_] 11: Plant Cell 1997 Mar;9(3):393-408 Related
Articles, Books

Expression of AG mRNA in the central region of floral meristems relies on the partially overlapping functions of the LEAFY (LFY) and APETALA1 (AP1) genes, which promote initial floral meristem identity.

AG -----> AP1

[_] 45: Plant Mol Biol 1995 Aug;28(5):767-84

Finally, since AG inhibits the expression of another floral regulatory gene AP1, we examined AP1 expression in antisense AG flowers, and found that AP1 is expressed at a relatively high level in the center of type II flowers, but very little or below detectable levels in the inner whorls of type III flowers.

EMF -----> AP1

[_] 5: Plant Cell Physiol 2001 May;42(5):499-507 Related Articles,
Our results indicate that the precocious expression of AP1 and LFY is dependent not only on the low EMF and FWA activities but also on the expression of most of the late-flowering genes such as FT, FCA, FE, CO and GI.

LFY TFL1

[_] 7: Plant Cell 1999 Jun;11(6):1007-18
Therefore, the normally sharp phase transition between the production of leaves with associated shoots and formation of the flowers, which occurs upon floral induction, is promoted by positive feedback interactions between LFY and AP1, together with negative interactions of these two genes with TFL1.

LFY -----> PI

[_] 4: Plant Cell 2001 Apr;13(4):739-53 Related Articles,
Localized expression of AP3 and PI requires the activities of at least three genes: APETALA1 (AP1), LEAFY (LFY), and UNUSUAL FLORAL ORGANS (UFO).

[_] 14: Development 2001 Jul;128(14):2735-46
AP3 and PI expression are positively regulated by the LEAFY (LFY) and UNUSUAL FLORAL ORGANS (UFO) genes.

[_] 20: Plant Cell 2001 Apr;13(4):739-53
To understand the epistatic relationship among AP1, LFY, and UFO in regulating AP3 and PI expression, we generated two versions of AP1 that have strong transcriptional activation potential.

LFY -----> AP3

[_] 1: Development 2002 May 1;129(9):2079-2086 Related
Here we show that the floral meristem identity genes LEAFY (LFY) and APETALA1 (AP1) are required for the activation of AP3.

[_] 3: Development 2002 May 1;129(9):2079-2086 Related Articles,

The LFY transcription factor binds to a sequence, with dyad symmetry, that lies within a region of the AP3 promoter required for early expression of AP3.

[_] 1: Development 2002 May 1;129(9):2079-2086 Related Articles,

Experiments using a steroid-inducible form of LFY show that, in contrast to its direct transcriptional activation of other floral homeotic genes, LFY acts in both a direct and an indirect manner to regulate AP3 expression.

[_] 2: Plant Cell 2001 Apr;13(4):739-53

To understand the epistatic relationship among AP1, LFY, and UFO in regulating AP3 and PI expression, we generated two versions of AP1 that have strong transcriptional activation potential.

UFO -----> PI

[_] 4: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,

Localized expression of AP3 and PI requires the activities of at least three genes: APETALA1 (AP1), LEAFY (LFY), and UNUSUAL FLORAL ORGANS (UFO).

[_] 4: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,

To understand the epistatic relationship among AP1, LFY, and UFO in regulating AP3 and PI expression, we generated two versions of AP1 that have strong transcriptional activation potential.

[_] 14: Development 2001 Jul;128(14):2735-46 Related Articles, Books,

These results support the idea that UFO and ASK1 together positively regulate AP3 and PI expression.

UFO -----> AP3

[_] 4: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,
Localized expression of AP3 and PI requires the
activities of at least three genes: APETALA1 (AP1), LEAFY
(LFY), and UNUSUAL FLORAL ORGANS (UFO).

[_] 14: Development 2001 Jul;128(14):2735-46 Related Articles, Books,
AP3 and PI expression are positively regulated by the
LEAFY (LFY) and UNUSUAL FLORAL ORGANS (UFO) genes.

[_] 20: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,
To understand the epistatic relationship among AP1, LFY, and
UFO in regulating AP3 and PI expression, we generated two
versions of AP1 that have strong transcriptional activation
potential.

[_] 14: Development 2001 Jul;128(14):2735-46 Related Articles, Books,
These results support the idea that UFO and ASK1
together positively regulate AP3 and PI expression.

[_] 77: Curr Biol 1997 Feb 1;7(2):95-104 Related Articles, Books,
In 35S::UFO flowers, AP3 was expressed precociously
and ectopically, confirming that UFO is an upstream
regulator of AP3.

[_] 1: Plant Cell Physiol 2002 Jan;43(1):52-7 Related
Articles, Books,
The expression of the B function homeotic gene APETALA3 (AP3)
and its regulator UNUSUAL FLORAL ORGANS (UFO) were delayed
and reduced in AP1::SUP flowers.

PI -----> AP3

[_] 19: Plant Cell 1997 Apr;9(4):559-70 Related Articles, Books,
Like AP3, all aspects of DEF function in Arabidopsis
required a functional PI protein.

[_] 32: Cell 1994 Feb 25;76(4):703-16 Related Articles, Books, LinkOut
AP3 and PI also activate an AP3 promoter-reporter gene fusion, demonstrating that AP3 positively autoregulates.

AP3 -----> PI (APETALA3 -----> PI)

[_] 31: Genes Dev 1994 Jul 1;8(13):1548-60 Related Articles,
The PI and APETALA3 proteins specifically associate in solution and so may act together in regulating PI and other genes.

PI -----> SUP

Plant Cell 2000 Sep;12(9):1607-18
AP3, PI, and another homeotic gene, AGAMOUS (AG), are further required for SUP expression in the later maintenance phase.

AP3 -----> SUP:

Plant Cell 2000 Sep;12(9):1607-18
AP3, PI, and another homeotic gene, AGAMOUS (AG), are further required for SUP expression in the later maintenance phase.

CAL LFY

Percept Mot Skills 1971 Jun;32(3):994
The closely related APETALA1 (AP1) and CAULIFLOWER (CAL) meristem identity genes are also important for flower initiation, in part because of their roles in upregulating LFY expression.

“improvements”

AP3 -----> AP1

1: Development 2002 May 1;129(9):2079-2086 Related Articles,
This LFY-induced expression of AP3 depends in part on the function of the APETALA1 (**AP1**) floral homeotic gene, since mutations in AP1 reduce LFY-dependent induction of AP3 expression.

AP1 -----> AP3

[_] 1: Development 2002 May 1;129(9):2079-2086 Related Articles

Here we show that the floral meristem identity genes LEAFY (LFY) and APETALA1 (AP1) are required for the activation of AP3.

: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,

To understand the epistatic relationship among AP1, LFY, and UFO in regulating AP3 and PI expression, we generated two versions of AP1 that have strong transcriptional activation potential.

AP3 -----> PI

[_] 19: Plant Cell 1997 Apr;9(4):559-70 Related Articles, Books,

Like AP3, all aspects of DEF function in Arabidopsis required a functional PI protein.

AP1 -----> PI

4: Plant Cell 2001 Apr;13(4):739-53 Related Articles, Books,

To understand the epistatic relationship among AP1, LFY, and UFO in regulating AP3 and PI expression, we generated two versions of AP1 that have strong transcriptional activation potential.[Genetic and molecular analyses of transgenic plants expressing these activated AP1 proteins show that the endogenous AP1 protein acts largely as a transcriptional activator in vivo and that AP1 specifies petals by regulating the spatial domains of AP3 and PI expression through UFO.]

AP1 ------> TFL1

3: Plant Cell 1999 Jun;11(6):1007-18

We show here that this negative regulation can be mutual because TFL1 expression is downregulated in plants constitutively expressing AP1

EMF1 ------> AP1

5: Plant Cell 1997 Nov;9(11):2011-24

We found that APETALA1 and AGAMOUS promoters were activated in germinating emf seedlings, suggesting that these genes may normally be suppressed in wild-type seedlings in which EMF activities are high.

Incorrect

AG -----> LFY

17: Plant Cell 1997 Mar;9(3):393-408 Related Articles,

Here, we provide evidence that AG function is required for the final definition of floral meristem identity and that constitutive AG function can promote, independent of LFY and AP1 functions, the determinate floral state in the center of reproductive meristems.

[_] 37: J Neurochem 1992 Nov;59(5):1893-904 Related Articles, Books,

These studies suggest that either L-A beta HA and DL-AP3 bind to a site on the receptor and irreversibly block activation of the receptor, or that these inhibitors act via a distinct site that specifically regulates EAA receptors coupled to PI hydrolysis.

[_] 27: Neurochem Int 1995 Jan;26(1):77-83 Related Articles, Books,

The glutamate metabotropic receptor antagonist 2-amino-3-phosphonopropionic acid (AP3), the ionotropic non-NMDA receptor antagonist 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX) and the NMDA channel blocker dizolcipine (MK-801) failed to prevent the PI response to ACPD (1000 microM).

AP3 -----> LFY

_] 1: Development 2002 May 1;129(9):2079-2086 Related Articles,

This LFY-induced expression of AP3 depends in part on the function of the APETALA1 (AP1) floral homeotic gene, since mutations in AP1 reduce LFY-dependent induction of AP3 expression.

AG -----> ag

[_] 5: Plant Cell 2001 Aug;13(8):1719-34 Related Articles, Books,
Interestingly, the homeotic conversion was not dependent on AG activity, because it was maintained in the ag-1 ap2-5 double mutant.

AG -----> ag

[_] 26: Mol Cell Biol 1999 Dec;19(12):8505-12 Related Articles, Books,
We prepared a series of AG genomic constructs in which these codons are mutated and assayed their activity in phenotypic rescue experiments by introducing them as transgenes into ag mutant plants.

ag -----> AG

[_] 32: Genes Dev 1999 Apr 15;13(8):1002-14 Related Articles,
In contrast to the outer two floral organs in sap mutant flowers, normal sepals and petals develop in ag/sap double mutants, indicating that SAP negatively regulates AG expression in the perianth whorls.

AGAMOUS -----> AG

[_] 43: Dev Biol 1997 Sep 15;189(2):311-21 Related Articles, Books,
In this report we demonstrate that manifestation of the fdh-1 phenotype does not require the product of the AGAMOUS gene, indicating that the phenotype is either independent of the carpel development program or that fdh-1 mutations activate a carpel-specific developmental program downstream of the AG gene.

ag <------ AG

[_] 56: Plant J 1996 Aug;10(2):343-53 Related Articles, Books, LinkOut
In the floral homeotic mutants ag-1, ap3-3 and ap2-2, AGL1 mRNA is expressed in an organ-dependent manner, suggesting that AGL1 is a carpel-specific gene and as such ultimately depends upon the carpel identity gene AG for proper gene expression.

ag <------ AG

[_] 56: Plant J 1996 Aug;10(2):343-53 Related Articles, Books, LinkOut
In the floral homeotic mutants ag-1, ap3-3 and ap2-2, AGL1 mRNA is expressed in an organ-dependent manner, suggesting that AGL1 is a carpel-specific gene and as such ultimately depends upon the carpel identity gene AG for proper gene expression.

ag -----> AG

[_] 58: Plant Cell 1996 May;8(5):831-45 Related Articles, Books,
In addition, transformants with a 35S-AG construct encoding an AG proteinlacking the C-terminal region produced ag-like flowers,indicating that this truncated AG protein inhibits normal AG

Appendix D Analysis of the results

- *LFY*----> *AG,* LFY activates AG. The program extracts two different sentences in which first the gene and the then the protein LFY activate AG. There were no incorrect predictions between these two.
- *LFY* -----> *AP1,* LFY is a regulator of AP1. By two sentences this conclusion was drawn. In one of the sentences is also included the prediction that AP1 may be regulated by LFY together with at least one other pathway. This is not shown in the headline but can be found by reading the generated sentences.
- *AP1* -----> *LFY* , AP1 can positively regulate LFY. The sentence extracted by the program includes both the information that the regulation of AP1 on LFY is positive i.e. that an increased expression of AP1 gives an increased expression of LFY, and how the author of the article has done this conclusion.
- *AP1*------> *AG,* Expression of AG relies on functions of LFY. Included in the sentence is also the statement that expression of AG relies on the function of AP1 and what kinds of genes LFY and AP1 are (they promote the development of initial floral meristem). Mendoza et al (1998) concluded in their network that AP1 represses the expression of AG. This is not seen in the result from the algorithm, but relies on could both include activates and represses so no specific answer is given here.
- *AG* -----> *AP1,* AG inhibits expression of AP1. Included in the text is also what kind of gene AP1 is (an floral regulatory gene) and which examination preformed that has given this conclusion.
- *EMF* -----> *AP1, EMF*-----> *LFY,* Expression of AP1 and LFY are dependent on the activity of EMF. These genes are not only dependent on EMF but also on other genes not included in the network. This information can be used when creating a new network se future work. The algorithm did not exclusively found the specific gene EMF1 but assumptions are made that these to genes are the same.

- *LFY* -----> *PI, UFO* -----> *AP3,* PI requires LFY and AP3 requires UFO. This is shown in the same sentences extracted by the algorithm. In the sentences are also concluded that the regulation is positive. The AP3 requires UFO findings are also found in three additional sentences, additional information about UFO being an upstream regulator of AP3.
- *LFY* -----> *AP3,* LFY regulates AP3. Two of the sentences extracted show that LFY is necessary to give an expression of AP3, the third sentence extracted also points out that LFY acts both in a direct and an indirect manner to regulate AP3 expression.
- *UFO* -----> *PI,* UFO regulates PI. In one of the sentences it is also stated that UFO is required for the activation of PI, in the same sentence it is found that other genes despite from UFO is necessary to regulate the activity namely AP1 and LFY. In one of the other sentences it is stated that UFO together with a gene not found in this network, ASK1, positively regulates the expression of PI. Since the abstracts, in which the sentences are found, are published the same year only three months apart, it is difficult to assess which abstract that is correct. Both could also be correct, by the means that AP3 requires the genes AP1, LFY and UFO to be expressed, but if the gene ASK1 also positively regulates it, it will get a different expression. ASK1 were not one of the genes included in the network of Mendoza et al. (1998) see future work.
- PI ------> AP3, PI activates AP3. In the sentence it stated that AP3 is required for the expression of PI. The result where both the protein and the gene are activating a gene is also shown in the relation between PI and AP3 In the second sentence it is also shown that PI activates a promoter-reporter gene fusion. In the second sentence it is also given that AP3 auto regulates which conclusion also were drawn by Mendoza et al. (1998).
- *AP3* -----> *PI,* AP3 regulate PI. This result was first shown when not using the abbreviation of AP3 but instead using the full name APETALA 3. The sentence also includes the result that AP3 acts together with PI to regulate PI, which was the same conclusion that Mendoza et al. (1998) drew in their network.

- *PI -----> SUP, AP3 -----> SUP,* PI and AP3 are required for expression of SUP. In the sentence is also concluded that AG is required. The sentence also concludes when the expression of SUP is necessary. These findings are in line with the ones of Mendoza et al. (1998).
- *LFY TFL1*, The algorithm did not first find the interaction between TFL1 and LFY, this may be because of the sentence it belongs in is negative. Although the sentence where LFY regulates TFL1 and TFL1 regulates LFY were found using the view function.

Appendix E New network results

This appendix includes the results from the developing of a new network.

Bub1 and Mad1

Bub1 -----> Mad1
[_] 81: J Cell Biol 2001 Jun 11;153(6):1239-50 Related Articles,
Books, LinkOut
Spindle checkpoint protein Bub1 is required for
kinetochore localization of Mad1, Mad2, Bub3, and
CENP-E, independently of its kinase activity.

Mad1 -----> mad
[_] 221: J Cell Biol 1999 May 31;145(5):979-91 Related Articles,
The following observations indicate that Bub2 and Mad1,
2 probably activate the checkpoint via different pathways:
(a) unlike the other Mad and Bub proteins, Bub2 localizes
at the spindle pole body (SPB) throughout the cell cycle;
(b) the effect of concomitant lack of Mad1 or Mad2 and
Bub2 is additive, since nocodazole-treated mad1 bub2 and
mad2 bub2 double mutants rereplicate DNA more rapidly and
efficiently than either single mutant; (c) cell cycle
progression of bub2 cells in the presence of nocodazole
requires the Cdc26 APC subunit, which, conversely, is not required
for mad2 cells in the same conditions.

Mad1 and Bub14

Bub1 -----> Mad1
[_] 81: J Cell Biol 2001 Jun 11;153(6):1239-50 Related Articles,
Books, LinkOut
Spindle checkpoint protein Bub1 is required for kinetochore
localization of Mad1, Mad2, Bub3, and CENP-E, independently
of its kinase activity.

Bub1 and Bub3

Bub1 -----> Bub3

[_] 81: J Cell Biol 2001 Jun 11;153(6):1239-50 Related Articles, Books, LinkOut

Spindle checkpoint protein Bub1 is required for kinetochore localization of Mad1, Mad2, Bub3, and CENP-E, independently of its kinase activity.

Bub3 <------ Bub1

[_] 240: Chromosoma 1998 Dec;107(6-7):376-85 Related Articles, Nucleotide, Protein, Books, LinkOut

Localization of the Drosophila checkpoint control protein Bub3 to the kinetochore requires Bub1 but not Zw10 or Rod.

Bub1 <------ Bub3

[_] 240: Chromosoma 1998 Dec;107(6-7):376-85 Related Articles,

Combined with recent findings showing that the kinetochore localization of Bub1 conversely depends upon Bub3, these results support the hypothesis that the spindle assembly checkpoint proteins exist as a multiprotein complex recruited as a unit to the kinetochore.

Bub3 -----> Bub1

[_] 260: J Cell Biol 1998 Jul 13;142(1):1-11 Related Articles, Nucleotide, OMIM, Protein, Books, LinkOut

The human homologue of Bub3 is required for kinetochore localization of Bub1 and a Mad3/Bub1-related protein kinase.

Bub1 -----> Bub3

[_] 260: J Cell Biol 1998 Jul 13;142(1):1-11 Related Articles,

Deletion mapping was used to identify the domain of Bub1 required for binding Bub3.

Bub1 -----> Bub3

[_] 360: Mol Cell Biol 1994 Dec;14(12):8282-91 Related Articles,

In vitro experiments confirmed that Bub1 possesses kinase activity; Bub1 was able to autophosphorylate and to catalyze phosphorylation of Bub3.

Bub3 and Mad3

Bub3 -----> Mad3

[_] 260: J Cell Biol 1998 Jul 13;142(1):1-11 Related Articles, Nucleotide, OMIM, Protein, Books, LinkOut

The human homologue of Bub3 is required for kinetochore localization of Bub1 and a Mad3/Bub1-related protein kinase.

Cdc20 and Mad1

Mad1 -----> Cdc20

[_] 221: J Cell Biol 1999 May 31;145(5):979-91 Related Articles,

Vertebrate homologues of Mad1, 2, 3, and Bub1, 3 bind to unattached kinetochores and prevent progression through mitosis by inhibiting Cdc20/APC-mediated proteolysis of anaphase inhibitors, like Pds1 and B-type cyclins.

Cdc20 and BUB1

Bub1 -----> Cdc20

[_] 221: J Cell Biol 1999 May 31;145(5):979-91 Related Articles,

Vertebrate homologues of Mad1, 2, 3, and Bub1, 3 bind to unattached kinetochores and prevent progression through mitosis by inhibiting Cdc20/APC-mediated proteolysis of anaphase inhibitors, like Pds1 and B-type cyclins.

Cdc14 and Cdc15

Cdc14 <------ Cdc15

[_] 70: Proc Natl Acad Sci U S A 2001 Jun 19;98(13):7325-30 Related

Cdc14 activity, in turn, is regulated by a group of proteins, the mitotic exit network (MEN), which includes Lte1, Tem1, Cdc5, Cdc15, Dbf2/Dbf20, and Mob1

Appendix

CDH1 and CDC6

CDC6 <------ CDH1

[_] 363: Genes Dev 2000 Sep 15;14(18):2330-43 Related Articles, Books,

Cell cycle- and cell growth-regulated proteolysis of mammalian CDC6 is dependent on APC-CDH1.

CDH1 -----> CDC6

[_] 363: Genes Dev 2000 Sep 15;14(18):2330-43 Related Articles, Books,

Furthermore, APC, in association with CDH1, ubiquitinates CDC6 in vitro, and both APC and CDH1 are required and limiting for CDC6 proteolysis in vivo.

CDH1 <------ CDC6

[_] 363: Genes Dev 2000 Sep 15;14(18):2330-43 Related Articles, Books,
The APC-CDH1-dependent proteolysis of CDC6 in early G(1) and in quiescent cells suggests that this process is part of a mechanism that ensures the timely licensing of replication origins during G(1).

CDC6 and CDC7

Cdc6 -----> Cdc7

[_] 709: Mutat Res 1999 Jan;436(1):1-9 Related Articles, Books,
Evidence replication to be a two-step process: the origin recognition complex, Cdc6 and Mcm proteins are required for establishingthe prereplicative complex and the activities of Cdks and of Cdc7 kinase then trigger the G1-S transition.

CDC7 <------ Cdc7

[_] 763: J Biol Chem 1998 Sep 4;273(36):23248-57 Related Articles,
We previously reported human and Xenopus cDNAs encoding CDC7-related kinases and suggested the possibility that chromosomal replication of higher eukaryotes may be regulated through conserved mechanisms involving Cdc7-related kinases.

Cdc7 -----> CDC7

[_] 1086: Mutat Res 1995 Jul;329(2):143-52 Related Articles, Books,
Because Cdc28 protein kinase and Dbf4 protein, a Cdc7 kinase regulator, are also important for induced mutagenesis and the CDC7 promoter is not induced in response to DNA damage, Cdc7 protein kinase may be regulated post-translationally following DNA damage, in the same manner as it is regulated during the cell cycle.

CDC27, CDC20 and CDC23

CDC27[_] 1240: Cell 1975 Aug;5(4):423-8 Related Articles, Books, LinkOut1
The instability of Cdc20 depends on CDC23 and CDC27, which encode components of the APC.

CDC27[_] 1240: Cell 1975 Aug;5(4):423-8 Related Articles, Books, LinkOut1
During the G1 phase, a destruction box within Cdc20 mediates its instability, but during S phase and mitosis, although Cdc20 destruction is still dependent on CDC23 and CDC27, it does not depend on the Cdc20 destruction box.

CDC27[_] 600: Curr Biol 1998 Jun 18;8(13):750-60 Related Articles, Books, LinkOut
The regulation of Cdc20 proteolysis reveals a role for APC components Cdc23 and Cdc27 during S phase and early mitosis.

Appendix F Analysis of the new networks relations

In this appendix an analysis of the result in appendix E is given.

- **BUB1 ----- > MAD1**

Probably correct. The checkpoint protein Bub1 is required for kinetochore localization of Mad1. By this means that during the chromosome segregation the protein BUB1 is required for the kinetochore to localize the correct localisation of Mad1. One could say that BUB1 is regulating MAD1 because if not BUB1 exists there is no transcription of MAD1. Here is thought the problem of the protein BUB1 that regulates the gene MAD1.

- **Bub1 -----> Bub3**

Bub1 requires Bub3. In one of the sentences extracted both Bub1 and Bub3 are genes, but in the other sentence the relation is between the gene Bub1 and the protein Bub3. The conclusion of the sentences is that the Bub3 requires the gene Bub1. In another sentence it is clear that the gene Bub3 have an effect on Bub1 because Bub1 has a domain used for binding Bub3. The gene Bub1 also have a kinas activity, which could autophosporylate and catalyze the phosphorylation of Bub3. The conclusion from the analysis is that the algorithm makes a correct prediction.

- **Bub3 -----> Bub1**

The kinetochore localization of Bub1 depends upon Bub3. In these findings a false connection were also stated; that the human homologue of Bub3 is required for localisation of BUB1. This doesn't mean that the gene Bub3 is required but the homologue, and the relation would therefore be incorrect if only this sentence were considered.

In the second of the sentences of BUB1 and BUB3 in appendix B, the algorithm does not pick out the relation between BUB1 and BUB3 although there is a correct relation. The sentence in this case contains implicit information; by autophosphorylate and to catalyse phosphorylation means that the gene is regulated by another gene, but the program does not understand this.

- **Bub1 -----Cdc20**

This sentence contains the information that it is not the specific BUB1 gene but instead it's homologues that inhibit CDC20 proteolysis. Since it is the homologues of Bub1, one cannot say anything specific between these two genes.

- **Mad3 ------> Bub1**

In this sentence we see that the delay in BUB1 was dependent upon other checkpoint genes including Mad3. This sentence should therefore be considered as a correct prediction by the algorithm.

- **Bub3 ----- Mad3**

Since it is the homologues of BUB1 that is concluded in the sentence in appendix, one cannot say anything specific between these two genes. But there are some indications that BUB3 regulates MAD3 but one needs more predictions to be able to include the relation in a network.

- **Mad1 -----> Cdc20**

The same thing is concluded for the gene MAD1 as for BUB1 above, it is the homologue that regulates the CDC20 mediated proteolysis. And the program therefore gives an inaccurate prediction.

- **CDC23 ---- > Cdc20 CDC27---- > Cdc20**

In the sentence it is stated that Cdc20 depends on the activity of Cdc23 and Cdc27. But in the second sentence the authors are talking about the destruction of Cdc20, this makes it is possible that they mean the protein Cdc20 and not the gene. The CDC20 is here meanwhile considered as a gene. In the third sentence it is the proteolysis of CDC20 that is regulated by the APC components CDC23 and CDC27. These relations were considered as correct.

- **Cdc15 ----- > Cdc14**

In this sentence the Cdc14 activity is regulated by a group of proteins which among others includes Cdc15. By this means that it is not the gene CDC15 but the protein CDC15 that is regulating the activity and the prediction is therefore incorrect.

- **CDH1 -----> CDC16**

In the first sentence extracted by the program is stated that CDH1 is required for CDC6 proteolysis, this sentence is correct found by the program. The second sentence involves the statement it is an APC-CDH1 dependent proteolysis of CDC6. The program has extracted this incorrectly as CDH1 is dependent of CDC6. The reverse is therefore incorrect.

- **CDC6 -----> CDC7**

Here the protein CDC6 together with another protein MCM is required for establishing the activities of CDC7 and CDKS. The connection is consequently not between the genes CDC6 and CDC7. An incorrect sentence is also given, and it is a sentence where CDC7 appears to be self-regulated.