DIMA – TU Berlin

Exposé of a Master Thesis

Working Title:	**Design and Implementation of an ETL Process and Data Warehouse for Literature mining of genetic Mutations in Cell Lines and statistical analysis of their occurrence.**
Adviser:	**Prof. Markl, Prof. Leser, Dr. Groth**
Candidate:	**Martin Schenck (328993)**
Planned Period:	**August 15th, 2010 – February 15th, 2011**

Background

Genomic mutations may result in severe diseases and predict for response to drug treatment. Therefore, research needs access to data about diseases and their associated genomic mutations in order to develop highly anticipated treatments for cancer and other diseases effectively.

By a process called 'protein biosynthesis', proteins are synthesized from genes by transcription of DNA sequence to mRNA and translation into their amino acid sequence. It is by this mechanism that genes encode for proteins. Proteins are essential for every organism and participate in many processes in the cells. They form enzymes and are thus vital parts of the metabolism by catalyzing biochemical reactions. Furthermore, proteins have structural and mechanical functions, e.g. in muscles or in the cytoskeleton, maintaining the shape of cells. Proteins also play significant roles in cell signaling, immune response, cell adhesion, and in the cell cycle (Campbell and Reece 2008).

The genetic code defines which way the nucleotides of DNA are translated into amino acids of proteins. For example, the amino acid glutamate is encoded by the DNA triplets GAA and GAG (for more details see (JCBNCommission 1984)).

A single-nucleotide polymorphism (SNP; pronounced /snɪp/) is a substitution of one nucleotide in a DNA sequence. Depending on the type of this exchange, the consequence of a SNP can vary. If, for example, the DNA codon GAA would be changed to GAG, it would still encode for the same amino acid (Glutamic acid). This is a "silent mutation". However, if it would change to GAT, a different amino acid would be the result (Aspartic acid). This is called a "missense mutation". Finally, a "nonsense mutation" occurs when a codon mutates into a stop codon, for example TAA. Stop codons stop translation at the current position of the genome.

Cell lines are used as a model for a certain disease state of a tissue sample. They are grown under controlled conditions and are used e.g. to analyze responses to treatments. Drug efficacy can be highly dependent on mutations in genes in cell lines.

As of now, only manually maintained databases exist, which contain information on the triplet of mutations in genes in cell lines. Human annotators have lost overview over the vast number of medical papers including the sought-after information. Therefore, creation of an algorithm that automatically mines the available medical literature and populates a database with information on mutations, genes, cell lines and their relation to one another will make a significantly larger dataset available.

Central Hypothesis

If automatic extraction instead of manual annotation is used to associate mutations, genes and diseases, then the resulting information will be considerably more comprehensive leading to better research conditions that may lead to novel therapeutic approaches.

Goal

Different tagger and verification methods and their results will be evaluated and compared to one another and to existing manually annotated databases. Furthermore, it will be investigated whether adjustment of an existing algorithm will lead to better results. The goal is to find the algorithm with the highest precision while having a reasonable recall, and to find out, whether an algorithm can be developed, which can generate more information faster than the manual approach.

Materials and Methods

To have a corpus the algorithm can extract information from, the Medline database can be used. The Medical Literature Analysis and Retrieval System Online (Medline) is a bibliographic database of biomedical literature from the life sciences. It currently contains approximately 18 million records[1]. The database can be accessed e.g. via PubMed (Wheeler, Barrett et al. 2006). To most of these entries, an abstract is freely available. A subset of these abstracts contains information on mutations, genes and diseases. Therefore they will be used to test the newly developed algorithm.

To find mutations, the algorithm will contain at least one existing mutation mining system. MutationFinder[2] finds and normalizes SNPs within free text (Caporaso, Baumgartner et al. 2007). MutationFinder also tries to find all different possibilities of mutation nomenclature (Antonarakis and McKusick 1994) and normalize them. Whether adjusting the regular expressions MutationFinder is using and its algorithm will improve the results, must be investigated.

For the purpose of gene/protein extraction, integration of named entity recognition algorithms for genes is indispensable. GNAT[3] searches text for mentions of genes and maps each gene to Entrez Gene identifiers (Hakenberg, Plake et al. 2008). Another tool to extract and normalize gene mentions is Moara (Neves, Carazo et al. 2010).[4] It is possible to include GNAT within Moara. Furthermore, the National Center for Integrative Biomedical Informatics[5] (NCIBI) provides an online name tagger, which tags genes in PubMed abstracts[6].

OSIRISv1.2[7] is a named entity recognition system for sequence variants of genes in biomedical literature (Furlong, Dach et al. 2008). It maps mutations and genes found in free text to dbSNP database entries. It is an alternative to e.g. MutationFinder in combination with GNAT.

To locate diseases, either the development of a disease term extractor based on a dictionary and regular expressions, or the incorporation of ABNER (Settles 2005) is feasible. ABNER (A Biomedical Named Entity Recognizer) is a machine learning system using conditional random fields. Among others, it extracts cell lines but the inclusion of ABNER would also enable another protein tagging possibility.

After the entities have been found, SNPs, genes and mutations must be associated to one another. Therefore, the following resources will be investigated to validate putative associations:

The Single Nucleotide Polymorphism Database (dbSNP) (Wheeler, Barrett et al. 2006) contains almost all known SNPs and additional related information. The SNP with the reference number rs28933988[8], for example, belongs to a certain organism (human) and to a

[1] http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update (accessed May 19th, 2010)
[2] http://mutationfinder.sourceforge.net/
[3] http://cbioc.eas.asu.edu/gnat/
[4] http://moara.dacya.ucm.es/
[5] http://portal.ncibi.org/gateway/
[6] http://nlp.ncibi.org/about.html
[7] http://ibi.imim.es/OSIRISv1.2.html
[8] http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=28933988

specific protein (Erythrocyte membrane protein band 4.2). Furthermore, it occurs at a specific location (142/112 in the protein) and has a specific substitution (A=>T). In addition, the database stores information on integrated maps, submitter records, the sequence in FASTA format (Pearson and Lipman 1988) and more. A SNP-gene-association found in medical literature can be checked against dbSNP. If it exists in dbSNP, it is very likely to be true positive. If the wildtype of SNP and protein match, the association can be assumed to be true positive.

The universal protein resource (UniProt) (Apweiler, Bairoch et al. 2004) is a resource of proteins and additional information about every protein, e.g. amino acid sequence, annotation and known mutations. Additionally, it references other databases like dbSNP. Therefore, it is possible to either use dbSNP or UniProt to evaluate found genes and mutations.

After associating genes and mutations, it is necessary to match those to disease models in form of cell lines or to disease terms from MeSH, in this work limited to neoplasms. If the algorithm extracts only one cell line or neoplasm from the text, associating the SNP-gene-pair to that one disease is reasonable. However, in case the algorithm finds multiple diseases, it is important to distinguish between right and wrong associations. Gene expression in cell lines is a strong indicator for mutations (Achan 1999; Jiang, Zhou et al. 2002; Langerod, Zhao et al. 2007). It is therefore most likely, that the gene and its mutations are associated to the cell line with a change in gene expression. The ArrayExpress Archive is a database of functional genomics experiments including gene expression[9]. Using ArrayExpress and comparing the gene expressions in different cell lines will help identify the right association. It must be investigated, whether an inclusion of the shortest word distance algorithm will improve results.

COSMIC (Bamford, Dawson et al. 2004) is an online database, containing 90,000 individual mutations for 3,423 genes in almost 370,000 tumors, derived from 7,797 Medline articles. Besides trying to develop a larger dataset than COSMIC, it is possible to verify findings using its entries. If a SNP-gene-disease-triplet is listed in COSMIC, it is considered as true positive.

Evaluation of the different available algorithms is the basis for comparing them to each other.

To evaluate a text mining algorithm, the following principles apply:

Precision measures the quality of a system. It is the fraction of true positives over all positively labeled elements. Recall measures the completeness. Recall is the fraction of true positives over all positives in the source.

The F_β-Score is a measurement for the quality of a text mine algorithm. Depending on β, precision and recall weigh differently.

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot precision \cdot recall}{\beta^2 \cdot precision \cdot recall}$$

Usually, one uses the F_1-Score. Precision and recall weigh equally.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

[9] http://www.ebi.ac.uk/microarray-as/ae/

Technical Realization

The entire tool will be implemented in Java. Most of the existing tools that can be incorporated also have a Java version available (MutationFinder, GNAT, Moara). To compare different approaches, the user will be able to choose one of the implemented text mining algorithms at runtime or upon execution.
The results will be written into an Oracle database. The database will be the foundation of the evaluation. It will be compared either to COSMIC or to a manually annotated gold standard.

Timeline

Main deadlines:
Mai 1st – August 15th: Preliminary research to set up the correct work space, tools, etc.:

- Integration of MutationFinder, GNAT and additional software
- Use new algorithm to mine PubMed for SNPs, genes and cell lines and their relation.

August 15th: Master Thesis Application
September 15th: Database creation and storing of mined data from Medline

- By then the database is hopefully filled with information gathered from Medline.

October 15th: Evaluation of the algorithm, project conclusion

- By then the algorithm shall have been evaluated.

December - February: Writing the Master Thesis
February15th: Master Thesis submission

Literature

Achan, V. (1999). "An introduction to molecular biology: gene structure, expression, and mutation." Pediatr Cardiol **20**(2): 94-96.

Antonarakis, S. E. and V. A. McKusick (1994). "Discussion on mutation nomenclature." Hum Mutat **4**(2): 166.

Apweiler, R., A. Bairoch, et al. (2004). "UniProt: the Universal Protein knowledgebase." Nucleic Acids Res **32**(Database issue): D115-119.

Bamford, S., E. Dawson, et al. (2004). "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website." Br J Cancer **91**(2): 355-358.

Campbell, N. and J. Reece (2008). Biology: International Version, Pearson Education.

Caporaso, J. G., W. A. Baumgartner, Jr., et al. (2007). "MutationFinder: a high-performance system for extracting point mutation mentions from text." Bioinformatics **23**(14): 1862-1865.

Furlong, L. I., H. Dach, et al. (2008). "OSIRISv1.2: a named entity recognition system for sequence variants of genes in biomedical literature." BMC Bioinformatics **9**: 84.

Hakenberg, J., C. Plake, et al. (2008). "Inter-species normalization of gene mentions with GNAT." Bioinformatics **24**(16): i126-132.

JCBNCommission (1984). "IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983." Eur J Biochem **138**(1): 9-37.

Jiang, Y., X. D. Zhou, et al. (2002). "Association of hTcf-4 gene expression and mutation with clinicopathological characteristics of hepatocellular carcinoma." World J Gastroenterol **8**(5): 804-807.

Langerod, A., H. Zhao, et al. (2007). "TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer." Breast Cancer Res **9**(3): R30.

Neves, M. L., J. M. Carazo, et al. (2010). "Moara: a Java library for extracting and normalizing gene and protein mentions." BMC Bioinformatics **11**: 157.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A **85**(8): 2444-2448.

Settles, B. (2005). "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text." Bioinformatics **21**(14): 3191-3192.

Wheeler, D. L., T. Barrett, et al. (2006). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **34**(Database issue): D173-180.