Mapping Genes to Diseases With Translational Data Mining

M. von Korff[1], A. Klenk[1], and T. Sander[1]
[1]Research Information Management, Actelion Pharmaceuticals Ltd., Allschwil, Switzerland

Abstract - *A new software tool (Gene2diseaseMapper) is presented that takes the HUGO symbol of a gene as input and delivers a ranked list of diseases, considering microarray experiments. Gene name synonym expansion was used to generate a query for MEDLINE. Retrieved article records were filtered by a disease stoplist which was created from Medical Subject Headings (MeSH). From the ArrayExpress microarray database, all experiments were retrieved in which the gene was differentially expressed. The experiments were searched for disease terms extracted from the MEDLINE articles. A similarity function was developed to compare the MeSH terms and the terms in ArrayExpress. A scoring function was implemented which ranked the disease MeSH terms according similarity and frequency. The method was explored with 12 genes for whose corresponding protein a drug was approved or is under development. Gene2diseaseMapper was able to find the diseases of the approved drugs in 11 out of 12 cases.*

Keywords: Bioinformatics, Translational medicine, Data mining, Microarray, Medline, Medical subject headings

1 Introduction

Many diseases are related to a change in the expression of proteins. The largest changes in protein expression can be found in cancer related diseases. Inflammation-related diseases also cause large changes in the protein expression pattern. A prominent example is rheumatoid arthritis, which affects millions of people worldwide. Neurodegenerative diseases such as Alzheimer, Parkinson and Huntington also show different protein expression patterns when compared to healthy control groups. Changes in protein expression can also be used to detect endogenous biomarkers; i.e. molecules which indicate a disease state [1]. For several years, microarrays have enabled expression profile analysis of the whole genome. The expression profiles of tens of thousands of genes can be explored in one experiment. In combination with information about experimental conditions, control groups, test groups and treatment these profiles are a valuable source for data mining. Because many journals require submission of the microarray data to one of the public repositories such as Gene Expression Omnibus [2] (GEO) or ArrayExpress [3] together with a publication, an enormous and quickly growing resource of medical information is available. The Gene Expression Atlas of the ArrayExpress database contains curated data from more than 5600 experiments. The differential expression of the genes is already calculated and the database is programmatically accessible. A p-value is given for every calculated differential expression value, indicating its reliability. The experiments are described and annotated with the Experimental Factor Ontology (EFO) [4].

A central point in a drug discovery program is the determination of the protein that will be targeted by the drug. This is often done at the beginning of a project, when a gene is chosen for cloning and expression to establish a biological screening assay. It is also possible to start a drug discovery program with a phenotype-based approach, but a drug will hardly be approved without a defined target protein. This target protein has to fulfill manifold requirements. It has to be related to a disease with a certain chance to cure it or, if not possible, to palliate symptoms. The gene encoding the target protein has to be known in order to develop biological screening assays.

Searching the medical literature for the relation between target and disease is the starting point in drug discovery. The huge expenditures for pharmaceutical research during the last decades resulted in a plethora of publications. Most of the medical information is not published in open access journals and is therefore not freely accessible. But almost all relevant biomedical literature is indexed in MEDLINE [5]. With more than 17 million bibliographic records MEDLINE is the largest repository for biomedical literature. Interfaces like PubMed enable human and programmatic access. What makes MEDLINE interesting for drug discovery is not only the specialization in life science-related subjects, but the hierarchical indexing system used to categorize the collected publications. The medical subject headings (MeSH) thesaurus consists of a controlled vocabulary, the MeSH descriptors, supplementary concepts and entry points [6]. The hierarchical structure of the MeSH thesaurus can be mapped onto a tree with general concepts close to the root and specific concepts in the leaves. Each node in the tree contains a unique node name, a MeSH descriptor, related concepts and entry points. There is no MeSH descriptor for the root node. Articles in MEDLINE are indexed with MeSH descriptor terms by searching the articles for entry points.

As a working hypothesis for this examination, it was assumed that the vocabulary from ArrayExpress and the standardized vocabulary from the medical subject headings can be used to detect overlapping information. Starting with the approved symbol for a gene it should be possible to detect information about related diseases in MEDLINE. Searching a microarray database with a gene symbol and related disease information should retrieve evidence on differential expression of that gene in a disease.

2 Methods

2.1 Stoplists

Stoplists are lists of node names that are used to activate or inactivate branches in the MeSH tree [7]. Only active MeSH terms are used for searching corresponding expressions in the microarray data. A simple stoplist for diseases activates the whole branch 'C' in the MeSH tree. Without sub-branch C22, which contains MeSH terms related to animal diseases, branch C contains 10782 nodes with 4466 unique MeSH headings (stoplist disease). Branch C04 in the MeSH tree was deactivated while searching microarray datasets for diseases which are not related to any form of cancer (neoplasms). Branch C04 contains only cancer-related MeSH terms. In addition, nodes in other branches were deactivated if their heading was equal to one of the headings in the C04 branch. After applying the disease stoplist omitting cancer, 8905 MeSH nodes with 3807 unique descriptor headings remained (stoplist disease, no cancer). Separating between neoplasm and other diseases is necessary, because neoplasia causes so many changes in gene expression that the relations between gene expression and other diseases would not be recognized. Of course this kind of restriction can also be applied to other diseases; e.g., inflammation-related diseases causes manifold changes in gene expression.

2.2 Programmatic access to PubMed

The MEDLINE databases can be accessed programmatically via the Entrez tools [8]. A query, containing a search term, submitted to MEDLINE via the PubMed interface returns a list of identifiers (PMID) which is used to obtain the publication records **R**. These records contain bibliographic information, often an abstract and the MeSH term headings which were used to index these articles.

2.3 Gene names and synonyms

A table with Human Genome Organization (HUGO) ids, gene names, approved symbols and synonyms was retrieved from HGNC (HUGO Gene Nomenclature Committee) [9]. The HUGO Gene Nomenclature Committee is located at the European Bioinformatics Institute and works under suspicion of the Human Genome Organization. From HGNC a table with gene names and their synonyms was retrieved. The MEDLINE database Gene also delivered HUGO ids, gene names and synonyms. [10] There is not a complete overlap between the synonyms in the two databases.

2.4 Searching PubMed records with gene names

To generate the query for searching the PubMed database, the approved symbol from the HUGO Gene Nomenclature Committee (HGNC) was used to find the synonyms from PubMed Gene and genenames.org. The synonyms were combined in a string by using 'OR' and sent as a query to PubMed. Without any further specification all fields in the PubMed database were searched. Depending on the gene symbol, a few records to the extent of several ten thousand were retrieved. All records which did not contain at least one active MeSH descriptor of the disease branch were skipped (stoplist disease, no cancer). The result was a dataset $\mathbf{R}_{\text{Gene}}$ for each gene. For the genes TNFSF11 and TPPP the branch C04, containing cancer-related diseases, was also activated (stoplist disease).

2.5 Searching ArrayExpress database

The Gene Expression Atlas of the ArrayExpress database contains curated and re-annotated microarray datasets [3]. This database was queried with the HUGO symbol for the gene under consideration. All experiments $\mathbf{MA}_{\text{Gene}}$ in which this gene was differentially expressed were retrieved. A gene experiment record contains the identifier of a microarray experiment in which the gene is up or down-regulated. Connected with the microarray experiment identifier is a record containing the experiment title, a description, the sample attribute values and the experimental factor values.

2.6 Searching microarray experiments with disease MeSH terms

The microarray experiments $\mathbf{MA}_{\text{Gene}}$ were searched for matching disease MeSH terms from $\mathbf{R}_{\text{Gene}}$. Each disease term was compared with the title, the description, the sample attribute values and the experimental factor values. Each of the resulting similarities s_{Title}, $s_{\text{Description}}$, s_{Sample} and s_{ExpFac} was multiplied by s_{Disease}, the frequency of occurrence of the disease term in the retrieved publication records. The highest scores from all microarray experiments $\mathbf{MA}_{\text{Gene,i}}$ were summed up. Because the terms which are used to annotate the experiments in ArrayExpress differ from the medical subject headings, a similarity function was needed to find similar terms. Each term, medical subject header $\mathbf{t}_{\text{MeSH}}$ or from a microarray experiment $\mathbf{t}_{\text{MA}}$, was decomposed into a list of unique words $\mathbf{u}_{\text{MeSH}}$ and $\mathbf{u}_{\text{MA}}$. A complete similarity matrix between these two lists was calculated by single word comparison using the Levenshtein similarity function [11]. From this matrix the optimum list of similarity pairs was

derived and their median taken as total similarity score for sim($\mathbf{u}_{MeSH}$, $\mathbf{u}_{MA}$). If at least one word from $\mathbf{u}_{MeSH}$ did not fit with a similarity >= 0.8 to any word in $\mathbf{u}_{MA}$ the similarity sim($\mathbf{u}_{MeSH}$, $\mathbf{u}_{MA}$) was set to 0. The score for a disease term $s_{MA,MeSH}$ was computed as the sum of all products of the frequency of occurrence of this term in the publication records multiplied by the maximum similarity of this term with a corresponding term in the microarray annotation.

3 Experiments

3.1 Targets with approved drugs

Seven targets for which a recently approved drug was available were chosen from the literature ($GeneSet_{Mature}$) (Table 1) [12]. With the HUGO approved gene symbol and the found synonyms a query string was generated and the PubMed records were retrieved as described in the paragraph "Searching PubMed records with gene names". The retrieved records were filtered with the disease filters and the remaining records underwent a first evaluation. From the MeSH headings a simple histogram was generated with the most frequent MeSH terms at the top. An example is given for gene HTR1A in Table 2. The rank of the indication equal to the indication of the approved drug was taken as a figure of merit for the applied algorithm. One rank score was obtained for the PubMed record derived MeSH term histogram and one for the sorted scores $s_{MA,MeSH}$ of the microarray experiment to MeSH term comparison.

3.2 Targets with drugs in development

Another set of five genes ($GeneSet_{New}$) was selected for which a drug was in development for the encoded protein [12]. The genes in $GeneSet_{New}$ are much less well explored than the genes in $GeneSet_{Mature}$, as can be seen from the number of retrieved PubMed records (Table 3). "Neuroinflammatory disease" is the indication of the corresponding drug for gene ALCAM. Because there is no MeSH term "Neuroinflammation" the indication was set to inflammation.

Table 2. MeSH term histogram for HTR1A with expanded query. Applied stoplist: disease, no cancer. "Frequency" is the frequency of occurrence of the MeSH headings in the 17617 PubMed records.

Rank	Disease MeSH heading	Frequency
1	Depression	206
2	Schizophrenia	178
3	Pain	162
4	Body Weight	152
5	Inflammation	133
6	Genetic Predisposition to Disease	123
7	Hypertension	118
8	Hypothermia	112
9	Heart Failure	100
10	Catalepsy	95

4 Results and conclusions

For dataset $GeneSet_{Mature}$ all approved drug indications were found by the MeSH term histograms (Table 4, index 1-7) and all found indications had a histogram rank below five, except for TPPP which was at rank 42. In dataset $GeneSet_{New}$, containing less explored genes (Table 4, index 8-12) also all drug indications were found. Two outliers were observed with the gene ALCAM and SLC6A7 on rank 27 and 29 respectively. In 11 of 12 cases the indications were confirmed by microarray experiments. For F13A1 (Factor XIII deficiency) no matching sample or condition was found in Gene Expression Atlas. Table 5 shows that the number of gene name synonyms ranges from 7-22. Comparing the retrieved number of articles for the HUGO symbol only and the query containing the gene name synonyms demonstrates a huge increase in retrieved articles by using synonyms (Table 1 and 5).

Table 1. Seven drug targets with at least one approved drug on the market (GeneSetMature). "HUGO" is the approved symbol for the target protein encoding gene. "Articles" is the number of articles retrieved from PubMed for the expanded gene query. "Articles, no expansion" is the number of filtered articles querying PubMed with the HUGO symbol only.

Index	HUGO	Drug	Indication	Articles	Articles, no expansion
1	HTR1A	Vilazodone	Depression	17617	103
2	TNFSF11	Xgeva	Bone metastases	5956	1
3	TPPP	Eribulin	Breast neoplasm	8796	42
4	GHRH	Tesamorelin	Obesity HIV patients (Obesity)	9705	2891
5	GLP1R	Victoza	Diabetes mellitus	866	36
6	PDE4	Roflumilast	Chronic obstructive pulmonary disease	2835	158
7	F13A1	Corifact	Factor XIII deficiency	3179	68

Table 3. Five drug targets with a drug in development (GeneSetNew). For explanations see Table 1.

Index	HUGO	Drug	Indication	Articles
8	ALCAM	AT-002 (CD166)	Neuroinflammatory disease (Inflammation)	507
9	APOC3	(Isis pharmaceuticals)	Cardiovascular disease	1544
10	SLC6A7		Alzheimer	1896
11	MAPKAPK5	GLPG-0259	Rheumatoid arthritis	247
12	CXCL16		Inflammation	241

The ratio between unique MeSH terms and the number of retrieved article records shows a roughly tenfold reduction taking the median of all values. CXCL16 is an interesting outlier, because with 238 retrieved records 114 MeSH terms were found. Remember that these are MeSH terms from the diseases stoplist that excluded cancer-related terms. This indicates that this target is active in a multitude of disease processes. The microarray experiments gave additional information. The number of experiments in which the gene under consideration was differentially expressed ranged from 233 for APOC3 to 741 for ALCAM. Interestingly, ALCAM is the gene with the second lowest number of matches between MeSH terms and microarray experiments. This means that many sample values and conditions from the microarray experiments did not match any one of the 77 disease MeSH terms which were used to index the literature containing one of the ALCAM gene name synonyms.

In conclusion, the proposed method summarizes up to thousands of MEDLINE publication records and relates the indexing MeSH terms to hundreds of microarray experiments in the Gene Expression Atlas. After sorting disease related MeSH term lists according to their score $s_{MA,MeSH}$, indications for approved drugs and drugs under development were at the top of the lists. Disease stoplists as filters for the indexing MeSH terms together with publicly available microarray data were successfully applied to targets of approved drugs and drugs under development. This demonstrates that Gene2diseaseMapper implements a new data mining method which prioritizes indications for targets in drug discovery programs.

Table 4. Result table for GeneSetMature (index 1-7) and GeneSetNew (index 8-12). "Indication" is the indication given for the drug targeting the protein encoded by "Gene". "Rank PubMed histogram" is the rank of the indication in the histogram of the MeSH terms which were derived from the PubMed query with the corresponding gene names. In "Rank PubMed-microarray" the rank of the indication according to the scored microarray experiments "MA score" is given. "MA score" is the resulting score from the evaluation of the microarray experiments with the disease MeSH terms. a No experiment with Factor XIII deficiency was found in the Atlas DB.

Index	Gene	Indication	Rank PubMed histogram	Frequency	Rank PubMed-microarray	MA score
1	HTR1A	Depression	1	189	1	2057
2	TNFSF11	Bone metastases	3	259	6	798
3	TPPP	Breast neoplasm	42	27	6	504
4	GHRH	Obesity HIV patients (Obesity)	4	190	1	784
5	GLP1R	Diabetes mellitus	1	195	1	14
6	PDE4	Chronic obstructive pulmonary disease	3	51	3	104
7	F13A1	Factor XIII deficiency	1	139	a	0
8	ALCAM	Neuroinflammatory disease (Inflammation)	27	3	13	48
9	APOC3	Cardiovascular disease	8	46	10	50
10	SLC6A7	Alzheimer	29	4	34	4
11	MAPKAPK5	Rheumatoid arthritis	6	2	10	4
12	CXCL16	Inflammation	1	14	1	240

Table 5. Details of the search for the genes in GeneSetMature (index 1-7) and GeneSetNew (index 8-12). "Synonyms" is the number of gene name synonyms that were used to query MEDLINE. "Articles" contains the number of unique MEDLINE article records that were retrieved by the query. "Unique MeSH terms" is the number of unique MeSH terms found in the indexing section of the articles. Column six gives the ratio between unique MeSH terms and the number of unique articles. Column seven contains the number of microarray experiments where the gene was found to be differentially expressed. Column eight indicates how many disease MeSH terms matched on at least one microarray experiment.

Index	Gene	Synonyms	Articles	Unique MeSH terms	Ratio MeSH/Articles	MA Experiments with differentially expressed genes	Matching MeSH on MA Experiments
1	HTR1A	16	17617	5044	0.29	258	279
2	TNFSF11	22	5956	5438	0.91	326	212
3	TPPP	16	8796	9070	1.03	313	315
4	GHRH	14	9705	4933	0.51	250	288
5	GLP1R	7	866	597	0.69	293	52
6	PDE4	11	2835	859	0.30	474	171
7	F13A1	16	3179	2068	0.65	429	282
8	ALCAM	12	507	298	0.59	741	75
9	APOC3	9	1544	1549	1.00	233	87
10	SLC6A7	7	1896	582	0.31	243	140
11	MAPKAPK5	9	247	69	0.28	451	33
12	CXCL16	13	247	69	0.28	404	77

5 Acknowledgement

We thank Susan Flores for editorial assistance.

6 References

[1] Y. Bauer, P. Hess, C. Qiu, A. Klenk, B. Renault, D. Wanner, R. Studer, N. Killer, A. K. Stalder, M. Stritt, D. S. Strasser, H. Farine, K. Kauser, M. Clozel, W. Fischli, and O. Nayler. "Identification of Cathepsin L as a Potential Sex-Specific Biomarker for Renal Damage"; Hypertension, 57, 4, 795-801, Feb, 2011.

[2] http://www.ncbi.nlm.nih.gov/geo/, accessed March 7 2011.

[3] http://www.ebi.ac.uk/arrayexpress/, accessed Feb 13 2011.

[4] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. "Modeling sample variables with an Experimental Factor Ontology"; Bioinformatics, 26, 8, 1112-1118, Apr, 2010.

[5] http://www.nlm.nih.gov/pubs/factsheets/medline.html, accessed Feb 8 2011.

[6] H. J. Lowe, and G. O. Barnett. "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches"; JAMA, 271, 14, 1103-1108, Apr, 1994.

[7] Don R. Swanson, Neil R. Smalheiser, and Vetle I. Torvik. "Ranking indirect connections in literature-based discovery: The role of medical subject headings"; Journal of the American Society for Information Science and Technology 57, 11, 1427-1439, 2006.

[8] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. "Database resources of the National Center for Biotechnology Information"; Nucleic Acids Res, 34, Database issue, D173-180, Jan, 2006.

[9] http://www.genenames.org, accessed Mar 29 2011.

[10] http://www.ncbi.nlm.nih.gov/gene, accessed Feb 10 2011.

[11] Fred J. Damerau. "A technique for computer detection and correction of spelling errors"; Communications of the ACM, 7, 3, 171-176, 1964.

[12] http://www.pharmaprojects.com/, accessed Mar 30 2011.