A NEW RELEVANCE ESTIMATOR FOR THE COMPILATION AND VISUALIZATION OF DISEASE PATTERNS AND POTENTIAL DRUG TARGETS

MODEST VON KORFF

Research Information Management, Actelion Pharmaceuticals Ltd., Gewerbestrasse 16
Allschwil, 4123, Switzerland
Email: modest.korff@actelion.com

TOBIAS FINK

Research Information Management, Actelion Pharmaceuticals Ltd., Gewerbestrasse 16
Allschwil, 4123, Switzerland
Email: tobias.fink@actelion.com

THOMAS SANDER

Research Information Management, Actelion Pharmaceuticals Ltd., Gewerbestrasse 16
Allschwil, 4123, Switzerland
Email: thomas.sander@actelion.com

A new computational method is presented to extract disease patterns from heterogeneous and text-based data. For this study, 22 million PubMed records were mined for co-occurrences of gene name synonyms and disease MeSH terms. The resulting publication counts were transferred into a matrix $\mathbf{M}_{data}$. In this matrix, a disease was represented by a row and a gene by a column. Each field in the matrix represented the publication count for a co-occurring disease–gene pair. A second matrix with identical dimensions $\mathbf{M}_{relevance}$ was derived from $\mathbf{M}_{data}$. To create $\mathbf{M}_{relevance}$ the values from $\mathbf{M}_{data}$ were normalized. The normalized values were multiplied by the column-wise calculated Gini coefficient. This multiplication resulted in a relevance estimator for every gene in relation to a disease. From $\mathbf{M}_{relevance}$ the similarities between all row vectors were calculated. The resulting similarity matrix $\mathbf{S}_{relevance}$ related 5,000 diseases by the relevance estimators calculated for 15,000 genes. Three diseases were analyzed in detail for the validation of the disease patterns and the relevant genes. Cytoscape was used to visualize and to analyze $\mathbf{M}_{relevance}$ and $\mathbf{S}_{relevance}$ together with the genes and diseases. Summarizing the results, it can be stated that the relevance estimator introduced here was able to detect valid disease patterns and to identify genes that encoded key proteins and potential targets for drug discovery projects.

I. INTRODUCTION

Many diseases coexist in a biological context with other diseases [1]. Patients often suffer from more than one disease. Furthermore, it is well known that some diseases cause secondary diseases. A well-established example for a disease with many co-occurring and second-order diseases is diabetes mellitus [2]. In addition, the context of a disease is of crucial importance in drug discovery. The ultimate goal of any new project in drug discovery is to treat or cure the disease in question. Realistically, however, for truly innovative drug discovery projects, in which the selected targets have been only recently identified, only sparse knowledge is available about the relationship between the chosen target and the potential disease. Knowledge about the context of the disease in which the target is involved may help to decide which constellation of diseases to take into account. Later in the drug discovery process, coexisting diseases should be considered for toxicity and DMPK studies. Drugs that are used for the treatment of pre-existing conditions may influence the metabolism of the patient and may result in an interaction with the drug being tested.

Whereas earlier research approaches for studying coexisting diseases were mainly based on phenotypic observations, recent technical advances have paved the way for using genomic and proteomic data. In this context, the "Online Mendelian Inheritance in Man"(OMIM) database was first published as a book and later as an electronic database [3]. This database related genotypes to phenotypes and inspired several research groups to develop computational tools to derive disease–disease associations. One of the first disease–disease association tools was reported by Goh et al. [4]. They developed a human disease network, connected by genes that were associated with two diseases. An enrichment of disease candidate genes via text mining of OMIM descriptions was implemented by van Driel et al. [5]. MeSH (Medical Subject Headings) annotations of MEDLINE articles were analyzed by Liu et al. to extract genetic and environmental factors associated with certain diseases [6]. An overview of the current research in disease associations was recently published by Sun et al. [7]. Hidalgo et al. derived the "Phenotypic Disease Network"(PDN) from 32 million patient records and received about six million co-morbidity relations [8]. This disease association network contains the co-morbidity data for more than 10,000 ICD9-encoded diseases and is one of the largest known so far. Taken together, a review of the existing literature suggests that multiple approaches exist to derive disease associations. However, MEDLINE represents the largest source of data, but it has not been used exhaustively for deriving disease–disease associations so far.

The approach presented here—the Disease–Disease Relevance Miner DDRelevanceMiner—annotated all records in MEDLINE with gene names and MeSH terms. Disease–disease associations were derived by comparing gene name based word vectors. These word vectors are histograms which are extracted from the records in PubMed by mining for co-occurrences of gene names and disease terms. This technique is known as second order co-occurrence and has been used by Schütze for word sense discrimination [9]. He exploited the fact that similar words are often accompanied by groups of identical words. Second-order co-occurrence has the advantage, as it allows calculating the similarity between two words that do not co-occur frequently, but that co-occur with the same neighboring words. Our method differs from Schütze's approach in two ways. The DDRelevanceMiner creates a word vector for a disease term from the complete text corpus and not from a single record. Word vectors, which represent a single text record, are often normalized and then weighted by the inverse document frequency. Weighing by the inverse document

frequency is not feasible for DDRelevanceMiner, because each word vector contains counts from all text records. To overcome this problem we introduced the relevance estimator.
In the next section, the calculation of the relevance estimator is explained in detail. Additionally, the data are described which were used to feed the algorithm and the diseases which were chosen for a thorough test of the results. The presentation and a discussion of the results follow at the end of the manuscript.

II. METHODS

A detailed description of the algorithms used by the precursor of DDRelevanceMiner—DDMiner—has been recently published [10]. Here, we describe the new algorithm which significantly improved the results of DDMiner. The new algorithm calculates a relevance estimator for a gene in dependency to a disease. How can one assess the merits of the relevance estimators? We assumed that ranking genes by their relevance estimators should help identifying potential drug targets. We also assumed that calculating similarities between diseases based on the relevance estimators should group these diseases in a meaningful way. Finally, if the relevance estimators are applied to a well-studied disease, it should be possible to prove the importance of the top-ranked genes and the disease patterns by literature. The relevance estimator is loosely related to the scoring scheme that was recently published by Mørk et al. [11].

A. Description of DDRelevanceMiner

DDRelevanceMiner used for analysis all available gene names from a table provided by the HUGO Gene Nomenclature Committee (HGNC) [12]. Gene name synonyms were taken from the HUGO table and other public available sources [13;14]. Every synonym was checked for ambiguity. Every synonym that passed the check was used to form a query for PubMed. A successful query retrieved a number of PubMed records. Index, title, and abstract of the record were searched for disease MeSH terms. All found disease MeSH terms were labeled with the approved gene symbol that was linked to the PubMed record. A detailed description of the search algorithm for gene and disease terms has been previously described in [10].
Querying PubMed with all gene name synonyms and parsing all retrieved records with all disease MeSH terms resulted in the central data matrix $\mathbf{M}_{\text{data}}$. A row in the matrix stood for a disease MeSH term and a column – for an approved gene symbol. Each field in the matrix, indicated by a row and a column, contained an integer number indicating how often a disease MeSH term occurred together with a gene. A row m from $\mathbf{M}_{\text{data}}$ shows which genes were studied together with the disease m. Vice versa, a column n shows which diseases were reported together with gene n.
However, the pure count of disease–gene co-occurrences is only of limited benefit. Genes with a high frequency of occurrence in the medical literature are often studied in relation to many diseases. But pharmaceutical research is mostly interested in genes that are specific for the disease of interest. Most interesting are genes that are specifically mentioned together with a disease of interest and not together with other genes. A gene with a high number of occurrences with one disease and no mentioning together with other diseases could be assumed to have a high relevance for the disease. Column n was extracted from $\mathbf{M}_{\text{data}}$ to calculate the relevance estimator $r_{\text{m,n}}$ for a disease with index m and a gene with index n. From column n, only fields with a publication count >0 were considered. For all fields in the column, their rank fraction $f_{\text{m,n}}$ was calculated. The rank ρ of a disease m for gene n is the position of the disease after sorting column n according to the number of publications. Diseases with identical publication counts were assigned the same rank.

Consequently, the number of ranks can be smaller than the number of diseases that were mentioned together with gene *n*. The rank fraction equaled one minus the rank divided by the total number of ranks $f_{m,n}= 1-\rho/\theta$, with θ for the total number of ranks. A fraction of publications $p_{m,n}$ was calculated by dividing the number of publications for disease *m* found in column *n* by the sum of all publications for gene *n*. The fraction of publications for disease *m* was weighted with the relative rank by $w_{m,n}= p_{m,n} f_{m,n}$. Finally, the relevance estimator $r_{m,n}$ was calculated by multiplying $w_{m,n}$ by the Gini coefficient g_n. The Gini coefficient describes the statistical dispersion for a group of values [15]. A Gini coefficient close to one indicates that all values except one in the group are zero. A Gini coefficient of zero indicates that all values in the group are equal. A relevance estimator $r_{m,n}$ of one is obtained if all three factors in the equation $r_{m,n}= p_{m,n} f_{m,n} g_n$ are equal to one. This means that all publications for gene *n* refer only to disease *m*. The calculation of the relevance estimator was done for all fields in the matrix $\mathbf{M}_{data}$. As result, a new matrix $\mathbf{M}_{relevance}$ was obtained, with the same dimensions as the input matrix. This matrix contained the relevance estimators; they covered a range between zero and one. This normalization enabled the meaningful comparison of matrix rows. To reduce the risk of rounding errors and to cut off the influence of very small values, the relevance estimators were multiplied by a factor of 1,000 and converted into integer numbers.
Each two $\mathbf{M}_{relevance}$ matrix rows were compared by calculating the generalized Jaccard similarity coefficient. Comparison of two matrix rows gave a similarity value between two diseases. The resulting similarity matrix $\mathbf{S}_{relevance}$ contained the similarity between all diseases.

B. Data

1) Genes and disease MeSH terms

A total of 39,410 approved gene and protein symbols were retrieved from the HUGO table. At least one disease MeSH term was found for 15,203 approved gene symbols. This number defined the number of columns in $\mathbf{M}_{data}$ and $\mathbf{M}_{relevance}$. A number of 5,256 unique MeSH descriptors defined the number of rows in $\mathbf{M}_{data}$ and $\mathbf{M}_{relevance}$.

2) Example diseases to assess the quality of the relevance estimators

Three example diseases, type 2 diabetes mellitus (T2DM), melanoma, and vitiligo were chosen for a detailed analysis of the disease–gene and the disease–disease associations. For each of the three diseases, five genes with the highest relevance estimators—the top genes—and the equal number of genes with the highest publication counts were analyzed. Additionally, for each of the three example diseases, ten most similar diseases were evaluated. Based on the working hypotheses described at the beginning of the Methods section, the following success criteria were defined. The usage of the relevance estimator can be regarded as a success if the top five genes include disease relevant genes. These genes should not be related to numerous other diseases. A relation of an example disease to a similar disease was regarded as valid if there was evidence found in literature.
T2DM is a subtype of diabetes and a complex metabolic disease. It is a complex disease because it involves environmental factors and multiple genes [16;17]. Despite the fact that many anti-diabetic drugs are on the market, the need for new anti-diabetic drugs is still high [18]. Obesity is a dominant risk factor for T2DM, whereas hypertension is one of the main co-occurring diseases. Therefore, we expected to see at least these two disease MeSH terms among the results of the similarity analysis. A number of genetic-driven studies were done for T2DM. Specific genes were found that increase the risk for this disease. Will the relevance estimator be able to identify some of these genes?

Melanoma is a malignant neoplasm (cancer) of the skin and the leading cause of death due to skin disease [19]. It is considered a highly immunogenic tumor [20]. Consequently, we expected to see association between melanoma and the genes that are tumor related but also with the genes that have relevance for the immune response.
Vitiligo is a disease where parts of the skin lose their pigment. Vitiligo is a frequently occurring disease, seen in 0.2%–2% of the population. However, its cause is still unknown [21]. A reason to choose vitiligo as an example disease was the relatively small number of related publications, compared to T2DM and melanoma. Another reason was the existing link between melanoma and vitiligo [22]. Will the disease similarity analysis based on the relevance estimator be able to find the link between these two diseases?

III. RESULTS

After querying PubMed with the synonyms from 39,410 approved gene and protein symbols, 2.7 million unique PubMed records were retrieved. Each of these records contained at least one disease MeSH term together with an unambiguous gene name synonym. The number of rows in $\mathbf{M}_{data}$ and $\mathbf{M}_{relevance}$ were defined by 5,256 unique disease MeSH terms found in the above publications, and the number of columns – by the 15,203 approved symbols. The similarity matrix $\mathbf{S}_{relevance}$ was calculated as described above in Methods. All results described in the following paragraphs were taken from $\mathbf{M}_{data}$, $\mathbf{M}_{relevance}$ and $\mathbf{S}_{relevance}$. The results for all genes and diseases are freely accessible at http://gene2disease.org.

A. Results for the three example diseases

Table 1 summarizes the disease–gene associations for the three example diseases. The table contains the disease name, the number of publications for the disease, the total number of genes found in the publications, and the top ten approved gene symbols. In the fourth column, the approved gene symbols are sorted by the relevance estimator from $\mathbf{M}_{relevance}$. In the fifth column, gene symbols are sorted by their publication counts from $\mathbf{M}_{data}$.
For every disease, the ten most similar diseases were extracted from $\mathbf{S}_{relevance}$. Cytoscape was used to visualize the results [23]. A Cytoscape network was created for every example disease (Figs. 1–3). A Cytoscape sketch for a disease contains the results from the corresponding row of Table 1 and from the corresponding table with the similar diseases. Every gene in a sketch is connected to the example disease. A disease is never directly connected to another disease and a gene is never connected directly to another gene. The example disease is marked by white text on the label. Other diseases are depicted as rectangles with black text on the label. The background color of the disease label corresponds to the similarity with the example disease. Similar diseases have a more intense background color than less similar ones. Genes are represented by ellipsoids or diamond shapes. The width of each shape corresponds to the number of connected diseases. A diamond shape symbolizes a gene that is in the top-five list of the genes sorted by relevance in Table 1. An ellipsoid indicates a gene that has similar relevance for two or more diseases, including an example disease, and is among the top five genes for one of the similar diseases.

1) Type 2 diabetes mellitus

DDRelevanceMiner found about 31,000 publication records relevant to T2DM which collectively mentioned synonyms for 2,273 genes (Table 1). All five genes with the highest relevance estimators have demonstrated relevance to T2DM. Namely, TCF7L2 is a diabetes susceptibility gene of substantial importance [24], and a potential target for anti-diabetic drugs [25]. GCK

(glucokinase) plays an important role in the carbohydrate metabolism and is an attractive target for new drugs [18], whereas GCK mutations are known to cause diabetes. CAPN10, SLC5A2, and SLC30A8 have high relevance for T2DM and are potential drug targets.

Ranking of the same set of genes by publication count differed completely from that by relevance estimators. Four out of the five genes with the highest publication counts, GCG (glucagon), INS (insulin), HBA1 (hemoglobin) and DIANPH, had low relevance estimators. This means that they were mentioned together with many other diseases, i.e., are not specific for T2DM. Low relevance of INS to T2DM was expected, because T2DM is a non–insulin-dependent disease [26]. In contrast, the fifth gene, DPP4 (dipeptidyl peptidase-4), had both a high publication count and a relatively high relevance estimator and is, indeed, a proven drug target for treating T2DM [27].

Table 1. Three selected example diseases and top relevant genes.

Disease	Publications	Gene count	Approved gene symbol (relevance estimator, publications) Top five genes sorted by	
			relevance	publication count
Type 2 diabetes mellitus	31,024	2,273	TCF7L2 (0.140, 386) GCK (0.138, 767) CAPN10 (0.132, 106) SLC5A2 (0.123, 210) SLC30A8 (0.120, 99)	GCG (0.053, 3761) INS (0.036, 3465) HBA1 (0.090, 2374) DIANPH (0.040, 2352) DPP4 (0.113, 1937)
Melanoma	27,000	3,271	MAGEA11(0.236, 14) TYR (0.209, 2357) PMEL (0.206, 23) MIA (0.202, 138) DCT (0.196, 253)	TYR (0.209, 2357) IL2 (0.018, 2191) IFNA1 (0.010, 2007) IFNG (0.010, 1867) IFNA2 (0.010, 1261)
Vitiligo	1,608	380	PCBD1 (0.033, 3) DCT (0.017, 28) PMEL (0.016, 3) LRR1 (0.015, 4) TYR (0.013, 170)	TYR (0.013, 170) CAT (0.0004, 70) IFNG (0, 52) IL2 (0, 49) IFNA1 (0, 46)

Table 2 lists the disease MeSH terms most similar to T2DM based on $\mathbf{S}_{\text{relevance}}$. All ten MeSH terms are known as major co-occurring diseases with diabetes, symptoms of diabetes, or, in case of obesity and body weight, known risk factors for the disease [28]. The size of the common gene set linking each disease with T2DM ranged from 72 to 517 genes. The 'top five genes' column of Table 2 shows genes whose relevance estimators calculated for the disease/MeSH term were most similar to those calculated for T2DM. For this reason, the genes and their order is different from Table 1 showing genes most relevant to T2DM only.

In Table 2, the disease MeSH terms most similar to T2DM are listed. The similarity values were taken from $\mathbf{S}_{\text{relevance}}$. All ten MeSH terms are major co-occurring diseases with diabetes, symptoms of diabetes, or, in case of obesity and body weight, known risk factors for the disease [28]. The size of the common gene set linking each disease with T2DM ranges from 72 to 517 genes. The top five

genes are the genes with the relevance estimators that are most similar to those calculated for T2DM. For this reason, there is a difference in the sorting order by the relevance estimators compared with Table 1, where the top genes were those most specific for one disease. Genes that connect more than one disease with T2DM are, e.g., SLC2A4, GCK, PLTP, and APOB. These genes may be regarded as having a key role in the disease pattern. The importance of these key genes is visualized in Fig. 1 where the width of the gene nodes corresponds to the number of connections.

Table 2. Diseases most similar to type 2 diabetes mellitus.

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Insulin Resistance	0.81	484	SLC2A4, IRS1, INSR, CAPN10, IRS2
Hyperglycemia	0.79	253	SLC5A2, GCK, PDX1, GCGR, HBA1
Obesity	0.72	517	ADIPOQ, GIP, ADRB3, FTO, GLP1R
Body Weight	0.70	286	GCK, IAPP, SLC2A4, GLP1R, UCP1
Hyperinsulinism	0.66	145	ABCC8, KCNJ11, GCK, SLC2A4, INSR
Hypoglycemia	0.61	72	SLC5A2, DPP4, GLP1R, ABCC8, HBA1
Atherosclerosis	0.58	243	PLTP, PON2, CETP, APOB, APOA1
Dyslipidemias	0.57	122	CETP, APOB, PPARA, PLTP, HMGCR
Hyperlipidemias	0.56	122	VLDLR, APOB, LPL, LIPC, PLTP
Hypertension	0.53	281	DIANPH, HBA1, ACE, ADRB3, REN

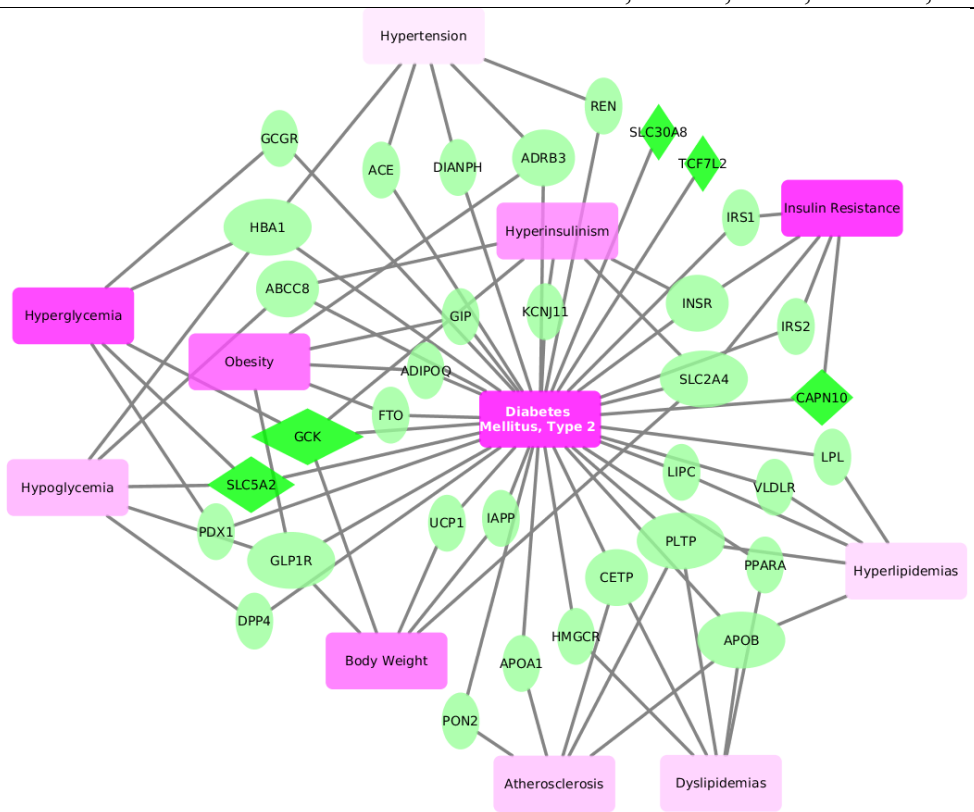

Fig. 1. Disease and gene pattern for type 2 diabetes mellitus

Analysis of gene-disease pattern for T2DM showed that SLC5A2 gene connects T2DM with two out of ten diseases with the highest similarity to T2DM (Fig. 1), which makes SLC5A2 an interesting drug target [29]. CAPN10 connects T2DM with insulin resistance, which is one of its known pre-conditions [30]. Other genes that connect T2DM with more than one disease (e.g., SLC2A4, GCK, PLTP, APOB), may be regarded as having a key role in the gene-disease pattern. The importance of these key genes is highlighted in Fig. 1 by the width of the gene nodes that is proportional to the number of disease connections.

2) Melanoma

Melanoma is a malignant cancer of the skin. The development of cancer includes many genes for cell growth and proliferation. Tyrosinase (TYR) is the gene with the highest publication count, and the top second rank according to the relevance estimator calculated for melanoma. Tyrosinase plays a central role in the process of skin pigmentation. Next four genes with high publication counts, interleukin 2 (IL2) and three interferons, are important for the immune response against cancer but are not specific for melanoma. The lack of specificity is the reason why these genes have low relevance estimators. MAGEA11 is the gene with the highest relevance estimator in the complete examination. Indeed, it is a melanoma antigen. Other genes with high relevance estimators, PMEL, MIA, and DCT, are highly specific for melanoma and are in the focus of ongoing research [31] [32] [33].

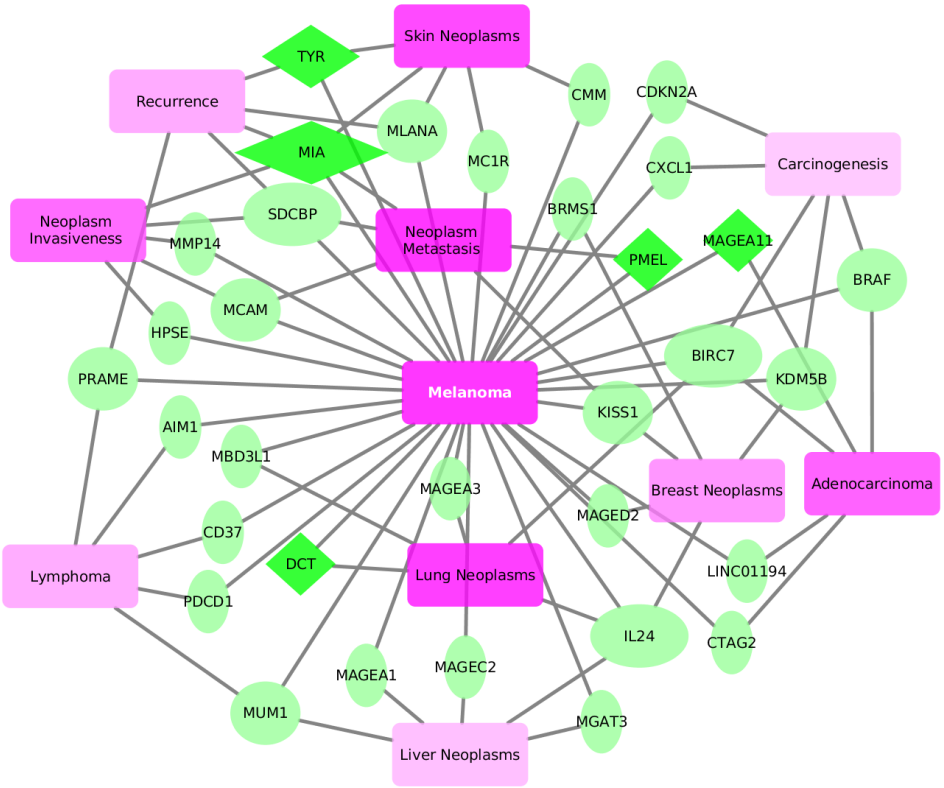

Fig. 2. Disease and gene pattern for melanoma.

The list of similar diseases in Table 3 is determined by diseases generally related to skin cancer and other cancer types. As already mentioned, the number of genes involved in cancer is higher than for other diseases. In the similarity list with the top five genes MIA the only gene that is represented

four times. MIA encodes the melanoma-derived growth regulatory protein. The combination of high relevance estimator and the connection of three melanoma-related MeSH terms suggest that MIA is a strong candidate for a drug discovery project.

Table 3. Diseases most similar to melanoma.

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Neoplasm Metastasis	0.65	1028	SDCBP, MIA, MCAM, KISS1, PMEL
Lung Neoplasms	0.65	622	MAGEA3, IL24, DCT, BIRC7, MBD3L1
Skin Neoplasms	0.65	366	MC1R, MIA, CMM, MLANA, TYR
Adenocarcinoma	0.63	728	LINC01194, BIRC7, MAGEA11, CTAG2, BRAF
Neoplasm Invasiveness	0.63	508	SDCBP, MIA, MCAM, HPSE, MMP14
Breast Neoplasms	0.60	788	BRMS1, IL24, KDM5B, MAGED2, KISS1
Recurrence	0.59	541	PRAME, MIA, MLANA, SDCBP, TYR
Lymphoma	0.59	581	CD37, PRAME, MUM1, AIM1, PDCD1
Liver Neoplasms	0.58	548	MAGEC2, MGAT3, MUM1, MAGEA1, IL24
Carcinogenesis	0.58	954	KDM5B, CDKN2A, CXCL1, BIRC7, BRAF

3) Vitiligo

Much less is known about vitiligo than for the other two example diseases. The absence of any gene with a high relevance estimator for vitiligo indicates a comparative lack of research.

Table 4 shows the most similar diseases to vitiligo. Coinciding with the low number of publication counts is the small size of the common gene sets. Nevertheless, some genes show multiple connections in the disease–gene network shown in Fig. 3. Tyrosinase has the most connections by linking nine diseases. Dopachrome tautomerase (DCT) connects seven diseases and is one of the most relevant genes for vitiligo. Also seven diseases are connected by the MITF gene encoding melanogenesis associated transcription factor, but this gene is not part of the top relevance genes. PMEL and TYRP1 genes connect six and five diseases, respectively. Fig. 3 shows that all four genes with the most connections (TYR, DCT, MITF, PMEL) relate vitiligo to the same three disease MeSH terms within the skin cancer complex: hypopigmentation, hyperpigmentation and skin neoplasms.

Table 4. Diseases most similar to vitiligo.

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Hyperpigmentation	0.72	7	TYR, DCT, MITF, PMEL, TYRP1
Melanosis	0.69	5	TYR, ASIP, LGI3, MITF, MC1R
Melanoma, Experimental	0.63	8	DCT, TYR, PMEL, MITF, TYRP1
Microphthalmos	0.51	10	DCT, TYR, MITF, PMEL, ASIP
Hypopigmentation	0.49	3	TYR, DCT, MITF

Disease/MeSH term	Similarity	Size of the common gene set	Top five genes
Melanoma, Amelanotic	0.49	3	TYR, DCT, MLANA
Epilepsy, Partial, Sensory	0.48	1	LGI3
Skin Neoplasms	0.43	10	TYR, DCT, PMEL, STX17, MITF
Arthritis, Juvenile	0.42	3	PTPN22, NLRP1, PTPN2
Albinism, Oculocutaneous	0.42	5	TYR, PMEL, TYRP1, GCHFR, MC1R

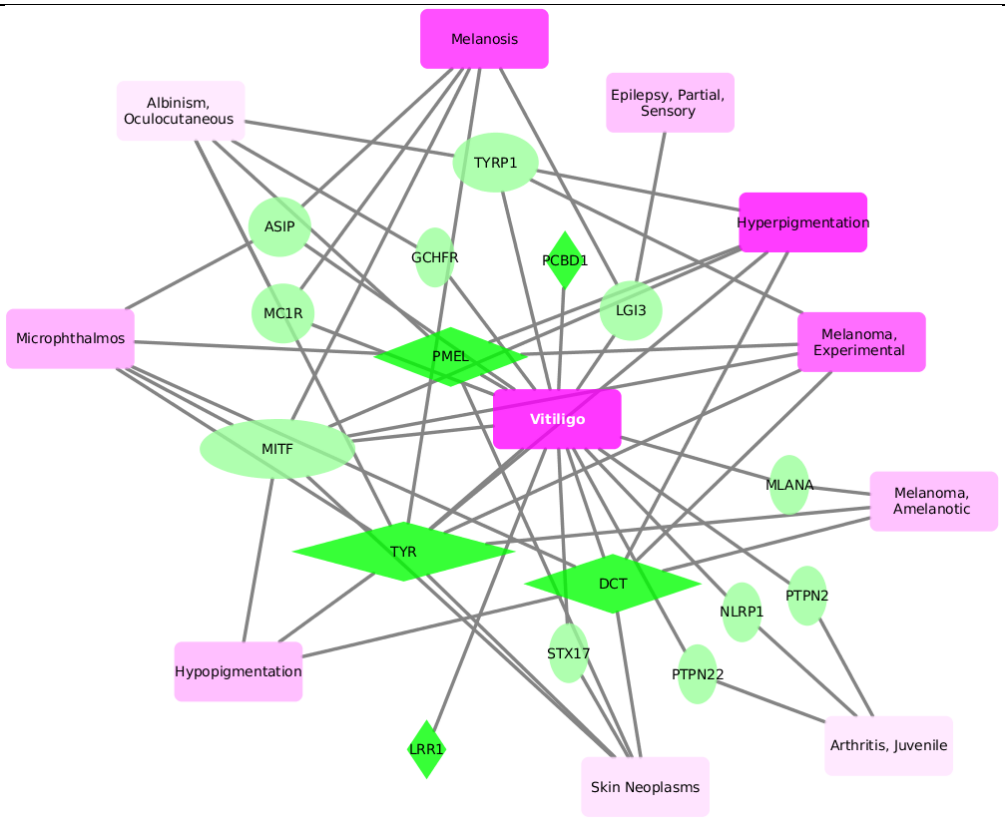

Fig. 3. Disease and gene pattern for vitiligo.

IV. Discussion

The majority of scientific publications contain information as heterogeneous data. And scientific publications are the main source for medical information. The NIH collected abstract records for millions of scientific publications in the PubMed database. It was our goal to extract and to visualize meaningful disease–gene and disease–disease patterns from this plethora of unstructured information. For this purpose, we introduced the relevance estimator described in this study. Three example diseases were chosen to examine this new figure of merit. For each disease, the relation to the most relevant genes was confirmed by literature. Furthermore, ten disease MeSH terms with disease–gene relationships most similar to one of the example diseases were identified and

analyzed. Cytoscape was used to visualize the most relevant genes, the similar diseases and the genes which connect the diseases with the example disease.

The most relevant genes for T2DM and melanoma were found to be highly specific for each disease. The ability of the relevance estimator to link MeSH terms with highly disease-specific genes that may only affect small patient groups makes it interesting for personalized medicine. The gene with the highest relevance estimator, TCF7L2, has been identified as a potential anti-diabetic drug target [25]. Obesity and hypertension were among the top ten disease MeSH terms with highest similarity to T2DM. This finding was pre-defined as a success criterion for the use of the relevance estimator.

For melanoma, a different disease–gene relationship pattern was obtained than for T2DM. All top genes are connecting at least one additional disease MeSH term with melanoma. Immune system relevant genes were listed as top genes by publication counts. However, these immune regulatory genes had low relevance estimators for melanoma. An example is interleukin 2, the protein product of the IL2 gene. This protein is used as a drug in the treatment of melanoma and is known to cause adverse side effects [34]. No genes with a high relevance estimator were found for vitiligo. Here, the combination of low publication counts and low relevance estimators emphasized that vitiligo is a disease with unknown genetic causes. Regardless of the low relevance score, three of the top five genes for vitiligo connected vitiligo to other skin-related disease MeSH terms. Thus, the earlier mentioned link between vitiligo and melanoma was confirmed using relevance estimators.

The evidence provided by the relevance estimators can be summarized as follows:

1. A high relevance estimator together with a low publication count indicates potential drug targets.
2. A low relevance estimator together with a high publication count indicate non-disease-specific genes.
3. A high relevance estimator and a high publication count mark a well-studied gene that is highly specific for the related disease.
4. Genes with low relevance estimators for a certain disease and high connectivity between multiple disease MeSH terms are likely to encode key proteins in a biochemical or signaling pathway.

Concluding, the relevance estimator is a valuable tool to extract disease-gene relation patterns from very large and heterogeneous data sets. Yet, the nature and importance of these patterns can only be evaluated by a scientist.

References

1 M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, *J Chronic Dis* **40**, (1987).

2 K. G. Alberti, and P. Z. Zimmet, *Diabet Med* **15**, (1998).

3 A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, *Nucleic Acids Res* **30**, (2002).

4 K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabasi, *Proc Natl Acad Sci USA* **104**, (2007).

5 M. A. van Driel, and H. G. Brunner, *Hum Genomics* **2**, (2006).

6 Y. I. Liu, P. H. Wise, and A. J. Butte, *BMC Bioinformatics* **10 Suppl 2**, (2009).

7 K. Sun, J. P. Goncalves, C. Larminie, and N. Przulj, *BMC Bioinformatics* **15**, (2014).

8 C. A. Hidalgo, N. Blumm, A. L. Barabasi, and N. A. Christakis, *PLoS Comput Biol* **5**, (2009).

9 H. Schütze, *Computational linguistics* **24**, (1998).
10 M. Von Korff, B. Deffarges, and T. Sander (2015). *In* "Computational Intelligence, 2015 IEEE Symposium Series on", p. 314. IEEE.
11 S. Mork, S. Pletscher-Frankild, A. Palleja Caro, J. Gorodkin, and L. J. Jensen, *Bioinformatics* **30**, (2014).
12 http://www.genenames.org
13 D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, *Nucleic Acids Res* **33**, (2005).
14 http://www.ncbi.nlm.nih.gov/gene
15 C. Gini, *Colorado College Publication, General Series* **208**, (1936).
16 L. Chen, D. J. Magliano, and P. Z. Zimmet, *Nat Rev Endocrinol* **8**, (2012).
17 R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos, and P. Froguel, *Nature* **445**, (2007).
18 P. Gaitonde, P. Garhyan, C. Link, J. Y. Chien, M. N. Trame, and S. Schmidt, *Clin Pharmacokinet* **55**, (2016).
19 M. A. Papadakis, S. J. McPhee, and M. W. Rabow (2015). *In* "Current Medical Diagnosis & Treatment 2015", p. 101. McGraw-Hill Education.
20 T. H. Nguyen, *Clin Dermatol* **22**, (2004).
21 A. Alkhateeb, P. R. Fain, A. Thody, D. C. Bennett, and R. A. Spritz, *Pigment Cell Res* **16**, (2003).
22 K. U. Schallreuter, C. Levenig, and J. Berger, *Dermatologica* **183**, (1991).
23 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, *Genome Res* **13**, (2003).
24 C. J. Groves, E. Zeggini, J. Minton, T. M. Frayling, M. N. Weedon, N. W. Rayner, G. A. Hitman, M. Walker, S. Wiltshire, A. T. Hattersley, and M. I. McCarthy, *Diabetes* **55**, (2006).
25 M. Ridderstrale, and L. Groop, *Mol Cell Endocrinol* **297**, (2009).
26 C. N. Hales, *Br Med Bull* **53**, (1997).
27 D. J. Drucker, and M. A. Nauck, *Lancet* **368**, (2006).
28 S. E. Kahn, R. L. Hull, and K. M. Utzschneider, *Nature* **444**, (2006).
29 A. Cesar-Razquin, B. Snijder, T. Frappier-Brinton, R. Isserlin, G. Gyimesi, X. Bai, R. A. Reithmeier, D. Hepworth, M. A. Hediger, A. M. Edwards, and G. Superti-Furga, *Cell* **162**, (2015).
30 L. J. Baier, P. A. Permana, X. Yang, R. E. Pratley, R. L. Hanson, G. Q. Shen, D. Mott, W. C. Knowler, N. J. Cox, Y. Horikawa, N. Oda, G. I. Bell, and C. Bogardus, *J Clin Invest* **106**, (2000).
31 F. Shi, Z. Xu, H. Chen, X. Wang, J. Cui, P. Zhang, and X. Xie, *Monoclon Antib Immunodiagn Immunother* **33**, (2014).
32 K. T. Yip, X. Y. Zhong, N. Seibel, S. Putz, J. Autzen, R. Gasper, E. Hofmann, J. Scherkenbeck, and R. Stoll, *Sci Rep* **6**, (2016).
33 S. A. Ainger, X. L. Yong, S. S. Wong, D. Skalamera, B. Gabrielli, J. H. Leonard, and R. A. Sturm, *Exp Dermatol* **23**, (2014).
34 C. Ma, and A. W. Armstrong, *J Dermatolog Treat* **25**, (2014).