Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Semantic Analysis of Genome Annotations using Weighting Schemes

Bogdan Done, Purvesh Khatri, Arina Done, Sorin Draghici
Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI, 48202

Abstract— **The correct interpretation of many molecular biology experiments depends in an essential way on the accuracy and consistency of the existing annotation databases. Such databases are meant to act as repositories for our biological knowledge as we acquire and refine it. Hence, by definition they are incomplete at any given time. In this paper we describe a technique that improves our previous method for extracting implicit semantic relationships between genes and functions. We added a number of weighting schemes to our previous latent semantic indexing approach. We used this technique to analyze the current annotations of the human genome. The predictions of 15 different weighting schemes were compared and evaluated. Out of the top 50 functional annotations predicted using the best performing weighting scheme, we found support in the literature for 82% of them. For 10% of our prediction we did not find any relevant publications, and 6% were actually contradicted by existing literature. This weighting scheme also outperformed the simple binary scheme used in our previous approach. Our method is independent of the organism and can be used to analyze and improve the quality of the data of any public or private annotation database.**

I. Introduction

Gene annotation databases capture the current biological knowledge allowing researchers to interpret the results of life sciences experiments. In spite of their indisputable importance, significant problems concerning the annotation databases still exist. One problem is that the annotations databases are currently incomplete. Organism specific annotation databases do not contain all the genes for these organisms, and even from the known genes only a subset are functionally annotated [28]. In addition to this, most of the annotations are introduced by curators who manually examine the literature. In this process it is possible that the annotations confirmed in publications might get overlooked [26]. Another problem is caused by the way these annotations are stored in the structure of Gene Ontology. There are, for instance, genes that are annotated for a particular molecular function but are not annotated for the corresponding biological process. This is not a problem for a database curator or a life scientist looking for the annotations of a specific gene, but this is not how such databases are used most of the time. In a more typical scenario, the researcher will try to interpret the results of a high throughput experiment using software that will query an annotation database in each of the three main branches of the GO graph and perform an automatic statistical significance analysis in order to uncover the biological processes that take place [12], [13], [25], [27], [2], [5], [22], [33], [40], [41]. This type of analysis will fail to correctly compute the statistical significance of the genes involved if they are not correctly annotated for each of the three GO categories. Lastly, a very important percentage of the annotations contained in these databases are relations inferred from electronic annotations (relations added without supervision of a human curator) [1]. Some of these annotations may also be incorrect [28], [38].

In order to address these problems, we proposed a method capable of finding gene-function associations that are not explicitly represented in the annotation databases [26]. For this purpose, our technique employs latent semantic indexing (LSI) on a organism specific annotations database. This method was demonstrated using the human genome annotations from the Onto-Tools database [12], [25], which includes all known annotations from the Gene Ontology Consortium (GO). The first implementation of our method used a binary representation of the relationships between genes and their functional annotations: if a gene i is found to be annotated with the function j, in the Gene Ontology graph, then the element at the intersection of row i and column j in the gene-function association matrix will have a value of 1; if this condition is not met the value of this matrix element is 0.

Clearly, this is limited in various ways including failing to properly capture the hierarchical relationship between various terms. Inspiration in how to eliminate these limitations can be found in information retrieval (IR) research. Previous work in this area shows that the use of a weighted representation, rather than a binary one, improves the quality of retrieval operations. Intuitively, IR term weighting attempts to exploit two simple observations: terms that appear repeatedly in a document are better suited to describe the topic of the document than terms that are rarely used, and infrequent terms across the document collection are better differentiators between documents than terms that appear in most or in all documents. Similar relationships might exist between genes and their annotations. For instance, genes that are annotated with only a few functions are most likely better suited to differentiate between functions in comparison to genes that are annotated with a large number of functions. Conversely, specific functions that are associated to very few genes, provide more information than generic functions that are associated with many genes. This paper explores the use of several weighting schemes in the context of a semantic analysis of biological annotations.

The technique described here is able to (i) discover potential inconsistencies in existing annotations and (ii) discover implicit gene-function relationships and propose them to the curators as novel annotations. Our approach applies latent semantic

1-4244-0710-9/07/$20.00 ©2007 IEEE

indexing (LSI) to the existing genome annotations databases to discover the missing functional annotations. LSI uses singular value decomposition (SVD) to find semantic relationships in the data that are not explicitly expressed (i.e. hidden) in the initial data. We present the results obtained with several weighting schemes on the annotations of the human genome stored in the Onto-Tools database [12], [25], which includes all known annotations from the GO Consortium.

Vector Space Model (VSM) [7], [6], [17] has been used previously to cluster genes by creating a vector space of genes and MEDLINE abstracts of papers discussing those particular genes [18]. The similarity between genes was assessed by computing a distance between the vectors that were representing them. It was found that weighted vectors improved the results significantly over boolean vectors (term-matching) [18]. VSM was also used to compute the similarity between Gene Ontology terms and the results were compared with other two non-lexical methods of analyzing the GO graph [8]. Latent Semantic Indexing (LSI) [7], [6], [10] was utilized in recent years for genome-wide expression data analysis [3]. LSI was also employed to identify relations between genes by creating a vector space of genes and MEDLINE abstracts [21]. Earlier Information Retrieval research has shown that LSI is 30% more effective than word matching methods [10]. The technique we are proposing is a novel, organism-independent approach that analyzes the entire body of annotations for a given organisms. It applies LSI on a weighted matrix of genes and GO terms. We used the human genome annotations but the same technique can be applied on annotation databases constructed for any organism.

Other approaches able to predict functional annotations for a given gene do exist. The most commonly used approach for function prediction uses sequence similarity. This approach is based on the hypothesis that a function can be transferred between similar sequences in different organisms since such similarity has been conserved over long periods of evolution [11]. This method of annotation transfer can result in incorrect function predictions due to reasons such as divergence of function within homologous proteins. Furthermore, this type of inference can also be incorrect because the annotations are only transferred from the closest homolog [24]. In order to overcome these problems, approaches combining sequence similarity data with structural information have been proposed [15], [36]. The guilt by association (GBA) approach [31], [37], [42], based on the observation that functionally related genes tend to share similar mRNA expression profiles, has also been widely applied to predict gene functions [9], [14], [23], [34], [39]. This approach clusters the genes based on their expression profiles in order to predict the gene functions. The GBA approaches are affected by issues such as data transformation [16], [30] and filtering intended to boost the signal-to-noise ratio [20]. An alternative approach uses sequence similarity and protein domain data in order to predict functional annotations [35]. Raychaudhuri et. al. [32] proposed a natural language processing approach for automatically extracting gene-function associations from the literature abstracts.

II. Methods

The technique described in this paper uses the annotations specific to *Homo sapiens* contained in GO [4]. GO maintains an organism-independent ontology of functional annotations that has a directed acyclic graph (DAG) structure. Each node in this graph represents a functional category and groups a number of genes annotated with that category. Researchers and curators endeavor to annotate the genes with the most specific functional category available in each case. For instance, if a gene is known to regulate the cell growth by extracellular stimulus, it is annotated with the specific category *"regulation of cell growth by extracellular stimulus (GO:0001560)"*, instead of a higher level, more general category such as *"regulation of cell growth (GO:0001558)"* or *"cell growth (GO:0016049)"*. However, a gene involved in *regulation of cell growth by extracellular stimulus* is actually involved in *regulation of cell growth* which is indeed part of the *cell growth* phenomenon. Because of this, we will consider that a gene annotated with a specific function f is also associated with the more general functional categories represented by the ancestors of f. In order to represent this in our data, we create a gene-function matrix GF as follows:

$$GF = \{gf_{ij}\} = \begin{cases} 1, \text{ if gene } g_i \text{ is known to be} \\ \quad \text{involved in function } f_j \text{ or} \\ \quad \text{any of its subcategories} \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

The rows of this matrix correspond to genes, while its columns correspond to functions. The i-th row of the matrix GF will represent all functions known to be associated with gene g_i either directly, as found in the literature, or through its descendants. Similarly, the j-th column of the matrix GF will represent all genes known to be associated with the function f_j, or any of its descendants.

Functional categories such as "unknown biological process" are used in GO in order to ensure a consistency of annotations. However, such terms lack any semantic content since they can be seen to group completely unrelated genes. Since our goal is to construct a model of the semantic relationships between genes and functions, such terms lacking semantic content must be removed from the analysis. Similarly, the top level nodes, "gene ontology" (GO:0003673), "biological process" (GO:0008150), "molecular function" (GO:0003674) and "cellular component" (GO:0005575) also lack a specific semantic content since all genes will appear other each of these terms. For these reasons, these GO terms are removed from the GF matrix.

The representation we used for gene-function relationships up to this point is binary. Previous work in information retrieval (IR) has shown that the performance of a system can be improved in terms of both precision and recall by using a more sophisticated representation [17]. Such representations can weight differently the associations between specific genes and functions using a vector space model (VSM). The weighting

schemes used in this paper will be denoted by three letter codes: the first letter refers to the local weight, the second letter to the global weight and the third letter is the normalization method used for the annotation vector. In our context we define gene frequency (gf) as the number of times a gene is associated with a function in the GO graph, and inverse annotation frequency (iaf) as *ln (total number of annotations / number of annotations that the particular gene considered has)*.

The local weight uses gene frequency. A gene frequency larger than 1 might indicate a stronger relation between the gene and a particular functional annotation. Inverse annotation frequency, used by the global weight, can be employed to favor genes that have only a few functional associations, because they are better differentiators between annotations. A normalization factor can be useful to penalize the annotations that are common to many genes. For instance, the annotations at the root of the graph are associated with almost all the genes in the gene-function matrix, therefore we need to normalize the weight so that functions close to the root will not overwhelm the specific functions found close to the leaves of GO DAG. Depth can be another useful factor for computing gene weights. Depth can be used with both local and global weights, for example relations close to the root, or relations that are contained in the gene-function matrix but are not found in the original relationship database (relations between the gene and an ancestor of the initial function), are not sufficiently specific and can be penalized.

Because the weighting schemes codes have not been used consistently in publications, we are giving here the definitions of the weighting schemes factors, the way they are employed in this paper. For local weights: n (none) means simple gene frequency, m (max) means *gene frequency* divided by maximum *gene frequency* in each annotation vector, a (augmented) equals *(0.5 + 0.5 * (gf / (max gf in the annotation vector))*, l (logarithmic) equals $1 + ln(gf)$; for global weights: t refers to *inverse annotation frequency*, which is equal to *ln (total number of annotations / number of annotations that the gene has)*; for normalization factor: n (none) indicates that normalization is not used, m (max) equals *weight / (maximum weight in the annotation vector)*, s (sum) equals *weight / (sum of weights in the annotation vector)*, c (cosine) equals *weight / (sqrt (sum of squared weights in the annotation vector))*, where weight equals *local weight * global weight for each element in the gene-function matrix*. Maximum and augmented local weights are employed to compensate for high gene frequencies; cosine normalization can be used to compensate for annotations common to a large number of genes. Based on these codes, the following 8 weighting schemes were tested in a first stage: ntn, ntm, ntc, mtn, atn, atm, atc and lts.

A depth correction was applied to both local and global weights. The weight of an indirect relationship between a gene g_i and a GO term t_j should diminish with the increase of the distance between the terms t_i and t_j (g_i is in a direct relationship with t_i, i.e. g_i is annotated with t_i in the GO database): if t_j is the parent of t_i then the relationship between g_i and t_j is most likely strong; however, if g_i is a gene annotated with a function t_i, found on a leaf of the GO tree, then the relationship between g_i and the annotations at the top of the GO tree (e.g. GO:0008150 - *biological process*) are not informative. Therefore, we would like to decrease the weight of an indirect relationship depending on the depth of the indirect relationship, which in the example above is the distance between t_i and t_j. For this purpose the local weight was multiplied with $1.7^{(-distance\ to\ the\ direct\ GO\ relationship)}$. Similarly, the global weight was multiplied with $(1 - 1.7^{(-distance\ to\ the\ root)})$, a factor that diminishes the global weight by almost half with every level up in the tree, to the root of the GO DAG. In-depth descriptions of VSM weighting schemes and the reasons behind them can be found in [17].

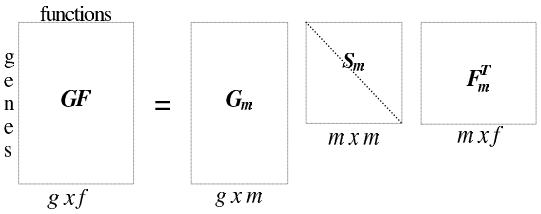

Fig. 1. Singular value decomposition of the gene-function association matrix GF. There are g genes and f functions. S_m is a diagonal matrix such that $S_{ij} = 0$, if $i \neq j$ and $S_{ij} \geq 0$, if $i = j$.

The next step is to decompose the matrix of weights GF as follows:

$$GF = G_m \times S_m \times F_m^T \tag{2}$$

In the equation above, G_m and F_m are matrices of the left and the right singular vectors. G_m and F_m have orthonormal columns, i.e. $G_m^T G_m = F_m^T F_m = I$, columns that are eigenvectors of the square matrices $(GF)(GF)^T$ and $(GF)^T(GF)$, respectively [19]. S_m is an $m \times m$ diagonal matrix. The elements of S_m are the singular values of GF and m is the rank of GF (i.e., the number of linearly independent rows or columns). The matrices G_m and F_m^T are the basis sets of size $g \times m$ and $m \times f$, respectively. All non-zero values in the i-th column of matrix G_m represent the genes known to be involved in the i-th function of matrix F_m. Similarly, all non-zero values in the i-th row of matrix F_m^T lists all functional categories the i-th gene of matrix G_m is known to be involved in. The decomposition of the matrix GF is represented in Fig. 1.

In Vector Space Models it is assumed that terms are independent, an assumption that is often false, because many terms are semantically related, or even equivalent. Term independence presupposition makes VSM easy to implement and conceptualize, but the accuracy of retrieval is negatively affected by it. However, there are IR techniques that can take advantage of the term dependence, LSI being one of them.

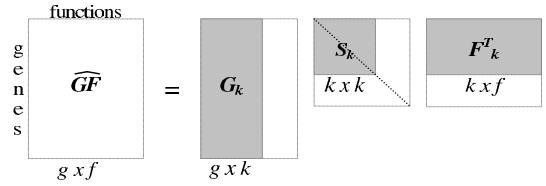

Fig. 2. The dimensionality reduction from m to k produces an approximation matrix $\widehat{GF}$ of the original matrix GF. By reducing the dimensionality we force the new matrix to capture the latent semantics and filter out the noise. This essentially will capture those interactions that are strongly represented in the data.

In our case, genes are not independent - groups of genes can be involved in more than one function. When this situation occurs it can be viewed as an indication that the group of genes might be semantically related. SVD projects co-occurring genes onto the same dimension and independent genes onto different dimensions. Such dependencies help us cluster the related genes and functions, but also provide us with the opportunity to approximate the same data in a space with fewer dimensions.

SVD works by rotating the m dimensional vector space and projecting the data into a new vector space where the highest variation of the data is found along the first dimension, the second highest variation along the second dimension and so forth. A useful SVD property is that we can reduce the dimensionality of the vector space. This is achieved by selecting only the k largest singular values of S_m and their corresponding vectors in G_m and F_m matrices, creating the matrices S_k, G_k and F_k. The product of these, $\widehat{GF}$, is a matrix which is the closest rank k approximation of GF in the least squares sense (see Fig. 2):

$$\widehat{GF} = G_k \times S_k \times F_k^T \tag{3}$$

We reduce the dimensionality, by selecting the largest k singular values (i.e., the k largest independent linear components) from S_m, in order to construct a model of the relationships between the genes and the functions. This model eliminates much of the noise, and also allows us to extract implicit gene-function relationships from the data. The matrix $\widehat{GF}$ only contains the associations that are strongly represented in the data. We should note that such strong relationships will be present in $\widehat{GF}$, even if they were not included in the original annotation database.

Once the matrix $\widehat{GF}$ is computed, we can investigate the semantic relationships between genes and their functional annotations, by selecting a threshold T. A value of $\widehat{gf}_{ij}$ greater than the threshold T might indicate that gene i has function j. Gene-function relationships which had $gf_{ij} = 0$ in the original GF matrix and now have $\widehat{gf}_{ij} > T$ correspond to newly discovered associations between genes and functions. Gene-function relationships which had $gf_{ij} \neq 0$ in the original GF matrix and now have $\widehat{gf}_{ij} \leq T$ in our projection space correspond to known functional annotations that have weak semantic support in the data, according to the method. Nevertheless, we should not conclude that these relationships are incorrect, novel phenomena may appear in contradiction with the rest of the annotations just because there are not enough annotations related to them at the time the investigation is made.

The choice of VSM weighting schemes is driven by analysis of results on various data sets. For each of the eight weighting schemes the first 50 best scoring relationships were evaluated manually by an expert. The scheme that performed the best, ntn, was altered in another six different weighting schemes in order to identify the terms in the weighting scheme that helped it achieve the best performance. We defined one new local weight and two new global weights: the local weight, called $n2$, had the same local depth factor but the *gene frequency* was not used (i.e. it has value 1 for all the genes); the global weight nt had the same global depth factor but the *inverse annotation frequency* was not used; for the global weight $nt2$ both the global depth factor and the *inverse annotation frequency* were not used. The six new weighting schemes derived from ntn were: $nntn$, $n2tn$, $n2tm$, $n2ntn$, $n2ntm$ and $n2nt2n$. The weighting terms, other than the three new terms defined here, have the same meaning as before.

III. Results and Discussion

The method described above was used to examine the existing functional annotations for the human genome. We were interested in analyzing the GO annotation graph in order to find relationships between genes and functions that are captured in the semantic layer of the graph, but are missing from the GO database. For the purpose of improving the performance of the method over our previous effort [26] we applied vector space model weighting schemes. VSM weighting schemes are considered better for retrieval purposes than a simple binary representation of the relationships between genes and their annotations. The performance of eight widely used weighting schemes was investigated on our dataset. The best performing weighting scheme was selected for a second round, where it was modified into six slightly different schemes, for fine tuning, and tested again.

The gene-function matrix GF was built using the human annotations contained in the Gene Ontology database, released in May 2003. The initial GF matrix contained 10,078 genes and 4,693 functional annotations, for a total of 300,204 relations between genes and functions. Relations that involved the annotations at the root of the Gene Ontology graph were not included in the GF matrix, to prevent these annotations from overwhelming the others: GO:0003673 (Gene Ontology), GO:0003674 (molecular function), GO:0008150 (biological process), GO:0005575 (cellular component). GO:0005554 (molecular function unknown), GO:0000004 (biological process unknown), GO:0008372 (cellular component unknown) were also excluded from the matrix, because they cluster unrelated genes. The genes and GO terms that had no associations were not included in the GF matrix, because they do not add semantic information.

TABLE I

AUTOMATIC ASSESSMENT RESULTS: EACH NUMBERED ROW CONTAINS THE NUMBER OF CONFIRMED PREDICTIONS IN MAY 2006 GO DATABASE FOR A PARTICULAR WEIGHTING SCHEME; atm, atn, ntn AND ntm PERFORMED BEST, BUT THE RESULTS WERE NOT CONCLUSIVE.

number of relations	atn	atm	atc	mtn	ntn	ntm	ntc	lts
25	2	3	0	1	1	1	0	2
50	6	5	1	1	1	3	0	2
100	7	8	1	1	4	3	3	2
200	7	13	1	1	15	12	3	3
400	18	22	13	1	23	20	13	3
800	31	32	21	10	38	28	22	9
1600	53	39	29	16	57	47	31	20
3200	-	-	57	-	93	84	60	25
6400	-	-	105	-	149	153	100	-
12800	-	-	175	-	-	228	176	-
25600	-	-	296	-	-	255	304	-
confirmed relations	53	39	331	16	149	255	339	25
relations above threshold	1552	1035	29552	1576	6040	14418	29033	2790
threshold used	0.17	0.34	0.04	0.04	0.02	0.12	0.04	0.04

TABLE II

THE RESULTS OF THE MANUAL ASSESSMENT IN THE FIRST STAGE: THE MOST SUCCESSFUL WEIGHTING SCHEME, ntn, OUTPERFORMED THE SIMPLE BINARY REPRESENTATION SCHEME, bin.

	atn	atm	atc	ntn	ntm	ntc	mtn	lts	bin	prev
2	18	21	12	35	26	19	3	4	18	-
1	4	9	7	5	6	5	5	7	9	-
0	16	17	15	7	12	20	19	34	18	7
-1	0	0	0	0	0	0	0	0	1	-
-2	10	1	6	3	6	6	5	4	2	-
obsolete	2	0	-	0	0	0	1	1	2	0

TABLE III

THE RESULTS OF THE MANUAL ASSESSMENT IN THE SECOND STAGE: THE WEIGHTING SCHEMES ANALYZED IN THE SECOND STAGE DID NOT SHOW IMPROVED RESULTS OVER THE EARLIER BEST PERFORMING SCHEME, ntn, BUT SOME OF THEM WERE COMPUTATIONALLY LESS DEMANDING.

	nntn	n2tn	n2tm	n2ntn	n2ntm	n2nt2n	ntn	bin
2	34	34	29	28	27	25	35	18
1	5	7	6	8	4	7	5	9
0	7	5	12	7	13	9	7	18
-1	0	1	0	1	1	1	0	1
-2	4	2	2	4	3	5	3	2
obsolete	0	1	1	2	2	3	0	2

We decompose the matrix GF using SVD and reduce its dimensionality to the largest 500 eigenvalues. The $\widehat{GF}$ matrix is constructed as in Fig. 2, by multiplying the reduced matrices resulted after SVD. We compute the value for the threshold T mentioned above, in following manner. We assume that the annotation database used to construct matrix GF contains mostly correct but also some incorrect gene-function relationships. For the purpose of defining a threshold, we assume that the true gene-function associations are those indicated by the LSI, i.e., those captured by $\widehat{gf}_{ij}$. In this hypothesis, the gene-function associations for which $gf_{ij} \neq 0$ in the original GF matrix and $\widehat{gf}_{ij} > T$ are true positives (TP). The gene-function relationships for which $gf_{ij} = 0$ in the original GF matrix and $\widehat{gf}_{ij} > T$ are false negatives (FN). In the same hypothesis, gene-function relationships for which $gf_{ij} \neq 0$ in the original GF matrix and $\widehat{gf}_{ij} \leq T$ are false positives (FP) and the associations that were not in the database initially ($gf_{ij} = 0$), and are also not revealed by the LSI ($\widehat{gf}_{ij} \leq T$) are true negatives (TN). A threshold close to the maximum value of $\widehat{gf}_{ij}$ will fail to discover many new functional annotations, but it also would imply that the database used has many FP relations (note that the maximum value of $\widehat{gf}_{ij}$ is greater than 1 for the weighting schemes that do not use the normalization factor, like ntn for instance). Clearly, this cannot be the case since most relationships are verified experimentally and known to be true. Similarly, for a threshold close to zero, the algorithm associates many genes with many functions which would imply that the original data set had many FNs. Using a criterion analogous to Occam's razor, we chose the value of the threshold T that corresponds to the assumption that the initial data set has the minimum amount of errors ($FP + FN$). This method preserves the large majority of the known relationships, but

several relations not previously found in the database will score higher than the threshold, indicating that they might be valid associations between genes and functions.

Initially we tried to automatically evaluate the accuracy of the weighting schemes by counting the number of confirmed relationships in the annotation database released three years after the data used for input. That is the new associations obtained from the annotation data released in May 2003 were compared with the data released in May 2006. The thresholds, the number of gene-function relationships that scored above the threshold and the number of confirmed relationships for each of the weighting schemes investigated in the first stage can be found in Table I. While ntn, ntm, atn and atm performed better than the other schemes, in terms of discovery rate, this automatic assessment could not differentiate well enough between them. In order to overcome this difficulty a biologist evaluated manually the first 50 highest scoring relations for each of the weighting schemes. The results of this examination can be seen in Table II. The rows of the table represent the scores given to the relationships. A score of 2 means that at least two papers were found proving that the relationship is correct, or that the relationship is already included in the annotation databases. A score of 1 was given when papers or tests suggest that the relationship is correct. Relationships for which no support was found in the literature able to confirm or contradict them, were given a score of 0. A score of -1 was given when papers were found suggesting that the relationship is not correct, and a score of -2 was given when strong literature support was found to prove that the relationship is not correct. The manual evaluation clearly showed that ntn was the best performing scheme: among its 50 relationships that were evaluated 35 are strongly supported in the literature, another 5 are confirmed by various research, about 7 of them there is nothing published yet, and only 3 were contradicted in the literature. In the second stage $n2tn$ performed as good as ntn, in spite of its lower computational overhead (Table III). Out of the 50 manually evaluated functional annotations predicted using $n2tn$, we found support in the literature for 82% of them. For 6% of the annotations we found evidence to the contrary and for 10% we did not found any relevant publications.

Two such predictions made using the $n2tn$ scheme are the "SLC2A10 - glucose transporter activity" and "SLC2A9 - glucose transporter activity". The human gene SLC2A10 is the solute carrier family 2 (facilitated glucose transporter), member 10 (a validated, well-documented structure). Obviously, SLC2A10 has "glucose transporter activity". Yet, it is not annotated for this biological category. It is annotated for the biological process "glucose transport", but not for the molecular function "glucose transporter activity". The annotations for the molecular function are far less specific than "glucose transporter activity": "sugar porter activity" and "transporter activity". The human gene SLC2A9, the solute carrier family 2 (facilitated glucose transporter), member 9, is in the same situation.

Another possible application of our predictions is to help increase the specificity of the annotations. For example, we predicted the relationship "AQP1 - water channel activity". AQP1, the human gene aquaporin 1 (Colton blood group), appears annotated for "porin activity", "transporter activity", and "water transporter activity". In spite of the fact that these annotations do not offer the user the precious information that aquaporin forms a channel for the water molecules (research awarded with the Nobel Prize in Chemistry in 2003) [29], not any other type of transporter, but a water channel.

IV. Conclusion

Gene annotation databases represent an essential resource for modern research in genetics. Such databases are used on a daily basis by thousands of researchers worldwide. However, it is well known that these annotations are incomplete and it is likely that some annotations are also incorrect. In this paper, we compared 15 weighting schemes that can be used to perform a global semantic analysis of the contents of such databases. As shown in Table II and Table III, use of gene frequency, inverse annotation frequency, and local and global depth provide better results than the binary approach (i.e., *bin*) that we used before [26]. Out of the top 50 functional annotations predicted using the best performing weighting scheme, we found support in the literature for 82% of them. For 10% of our prediction we did not find any relevant publications, and 6% were actually contradicted by existing literature. In addition, Table II shows that any normalization applied to the local weights deteriorates the accuracy. This technique is able to predict novel functional annotations for known genes, and is independent of the organism. It can be used to analyze and improve the quality of the data of any public or private annotation database.

Acknowledgment

This work has been supported by the following grants: NSF DBI-0234806, DOD DAMD 17-03-02-0035, NIH(NCRR) 1S10 RR017857-01, MLSC MEDC-538 and MEDC GR-352, NIH 1R21 CA10074001, 1R21 EB00990-01 and 1R01 NS045207-01.

References

[1] GO Consortium, Current Annotations. http://www.geneontology.org/GO.current.annotations.shtml, 2004.

[2] Fatima Al-Shahrour, Ramon Diaz-Uriarte, and Joaquin Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[3] O. Alter, P.O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97(18):10101–10106, 2000.

[4] Michael Ashburner et al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, May 2000.

[5] T Beissbarth and TP. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics.*, 20:1464–1465, June 2004.

[6] M. W. Berry, Z. Drmac, and E. R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335–362, 1999.

[7] Michael W. Berry, Susan T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM: Review*, 37(4):573–595, 1995.

[8] O. Bodenreider, M. Aubry, and A. Burgun. Evaluation of the vector space representation in text-based gene clustering. In *Pacific Symposium on Biocomputing*, pages 91–102, 2005.

[9] Michael P. S. Brown, William Boble Grundy, David Lin, Nello Cristianini, Charles Waish Sugnet, Terrence S. Furgey, Manuel Ares Manuel, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97(1):262–267, 2000.

[10] Scott Deerwester, Susan T. Dumais, G. W. Furnas, Thomas K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[11] Damien Devos and Alfonso Valencia. Practical limits of function prediction. *PROTEINS: Structure, Function, and Genetics*, 41:98–107, 2000.

[12] Sorin Drăghici, Purvesh Khatri, Pratik Bhavsar, Abhik Shah, Stephen A. Krawetz, and Michael A. Tainsky. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, 31(13):3775–81, July 2003.

[13] Sorin Drăghici, Purvesh Khatri, Rui P. Martins, G. Charles Ostermeier, and Stephen A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, February 2003.

[14] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 95(25):14863–14868, December 1998.

[15] Jacquelyn S. Fetrow, Naomi Siew, Jeannine A Di Gennaro, Maria Martinez-Yamout, H. Jane Dyson, and Jeffrey Skolnick. Genomic-scale comparison of sequence- and structure-baed methods of function prediction: Does structure provide additional insight? *Protein Science*, 10:1005–1014, 2001.

[16] Sue C. Geller, Jeff P. Gregg, Paul Hagerman, and David M. Rocke. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, 19:1817–1823, 2003.

[17] Salton Gerard. *Introduction To Modern Information Retrieval*. McGraw-Hill, 1983.

[18] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. In *Pacific Symposium on Biocomputing*, pages 391–402, 2003.

[19] Gene Golub and Charles F. van Loan. *Matrix computations*. The Johns Hopkins University Press, 1983.

[20] J. Herrero, R. Diaz-Uriarte, and J. Dopazo. Gene expression data processing. *Bioinformatics*, 19:655–656, 2003.

[21] R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry. Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics*, 21(1):104–115, 2005.

[22] Douglas A. Hosack, Glynn Dennis Jr., Brad T. Sherman, H. Clifford Lane, and Richard A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(6):P4, 2003.

[23] T. R. Hvidsten, A. K. Sandvik, A. Laegreid, and J. Komorowski. Predictive gene function from gene expressions and ontologies. *In Proceeding of Pacific Symposium on Biocomputing*, 2001.

[24] Peter D. Karp. What we do not know about sequence analysis and sequence databases. *Bioinformatics*, 14(9):753–754, 1998.

[25] Purvesh Khatri, Pratik Bhavsar, Gagandeep Bawa, and Sorin Drăghici. Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Research*, 32:W449–56, Jul 2004.

[26] Purvesh Khatri, Bogdan Done, Archana Rao, Arina Done, and Sorin Draghici. A semantic analysis of the annotations of the human genome. *Bioinformatics*, 21(16):3416–3421, 2005.

[27] Purvesh Khatri, Sorin Drăghici, G. Charles Ostermeier, and Stephen A. Krawetz. Profiling gene expression using Onto-Express. *Genomics*, 79(2):266–270, February 2002.

[28] Oliver D. King, Rebecca E. Foulger, Selian S. Dwight, James V. White, and Frederick P. Roth. Predicting gene function from patterns of annotation. *Genome Research*, 13:896–904, 2003.

[29] David Kozono, Masato Yasui, Landon S. King, and Peter Agre. Aquaporin water channels: atomic structure molecular dynamics meet clinical medicine. *J. Clin. Invest.*, 109(11):1395–1399, 2002.

[30] W. Pan, J. Lin, and C. Le. Model-based cluster analysis of microarray gene expression data. *Genome Biology*, 3(2):research0009.1–0009.8, 2002.

[31] John Quackenbush. Microarrays–Guilt by Association. *Science*, 302(5643):240–241, 2003.

[32] Soumya Raychaudhuri, Jeffrey T. Chang, Patrick D. Sutphin, and Russ B. Altman. Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research*, 12:203–214, 2002.

[33] Joel Richardson. Vlad: A new GO tool for visual annotation display. http://www.informatics.jax.org/∼jer/vlad/.

[34] Karine G. Le Roch, Yingyao Zhou, Peter L. Blair, Muni Grainger, J. Kathleen Moch, J. David Haynes, Patricia De la Vega, Anthony A. Holder, Serge Batalov, Daniel J. Carucci, and Elizabeth A. Winzeler. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639):1503–1508, 2003.

[35] Jonathan Schug, Sharon Diskin, Joan Mazzarelli, Brian P. Brunk, and Christian J. Stoeckert Jr. Predicting Gene Ontology functions from PromDom and CDD protein domains. *Genome Research*, 12:648–655, 2002.

[36] Jeffrey Skolnick and Jacquelyn S. Fetrow. From genes to protein structure and function: Novel applications of computational approaches in the genomic era. *Trends in Biotechnology*, 18:283–287, 2000.

[37] Michael G. Walker, Wayne Volkmuth, Einat Sprinzak, David Hodgson, and Tod Klingler. Prediction of Gene Function by Genome-Scale Expression Analysis: Prostate Cancer-Associated Genes. *Genome Research*, 9(12):1198–1203, 1999.

[38] Haiying Wang, Francisco Azuaje, Olivier Bodenreider, and Joaqun Dopazo. Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 25–31, 2004.

[39] LF Wu, TR Hughes, AP Davierwala, MD Robinson, R Stoughton, and SJ. Altschuler. Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, 31(3):255–265, July 2002.

[40] Barry R. Zeeberg, Weimin Feng, Geoffrey Wang, May D. Wang, Anthony T. Fojo, Margot Sunshine, Sudarshan Narasimhan, David W. Kane, William C. Reinhold, Samir Lababidi, Kimberly J. Bussey, Joseph Riss, J. Carl Barrett, and John N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, March 2003.

[41] B Zhang, D Schmoyer, S Kirvo, and J. Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics.*, 5(16), February 2004.

[42] GH Zhou, XY Wen, H Liu, MJ Schlicht, MJ Hessner, PJ Tonellato, and MW. Datta. BEAR GeneInfo: A tool for identifying gene-related biomedical publications through user modifiable queries. *BMC Bioinformatics.*, 5(46), April 2004.