Intelligent Decision Technologies 3 (2009) 233–238
DOI 10.3233/IDT-2009-0058
IOS Press

Ranking genes based on kernels

Nantia Iakovidou[a,*], Alexandros Nanopoulos[b] and Yannis Manolopoulos[a]
[a]*Department of Informatics, Aristotle University, GR-54124 Thessaloniki, Greece*
[b]*Institute of Computer Science, University of Hildesheim, D-31141 Hildesheim, Germany*

Abstract. Retrieval queries in microarray databases can rank genes according either to their similarity by detecting functionally related genes, or to their importance by detecting genes with significant regulation role. Although both rankings are useful, they can be contradicting. For instance, similar highly ranked genes may have low importance and vice versa. Thus, we propose a Web-inspired kernel method for fusing the two rankings according to the user needs.

Keywords: Gene similarity, gene importance, ranking

1. Introduction

Several microarray databases have been developed in recent years, such as the Stanford Microarray Database and the GenBank Database, serving as storage sites for microarray data and facilitating their public dissemination.[1] These data can be expression data or gene ontology terms. Users query such databases to retrieve information and associated annotations about genes, by specifying certain criteria, e.g. retrieve genes whose expression level varies by a certain amount.

Similarity queries are a versatile primitive for querying microarray databases. They seem to be very useful in important applications, such as gene clustering and ranking. Chabalier et al. [1] cluster genes according to biological, medical, genomic and expression annotations, which results in several gene functional and biologically relevant networks. In general, a similarity query aims at retrieving similar genes, which means detecting those genes that are functionally related to the query gene. Using techniques from Information Retrieval (e.g., the vector space model and the cosine correlation similarity), similarity between two genes is measured in terms of expression levels or gene ontology (GO) annotations [1]. The output of a similarity query contains genes ranked in descending order of their similarity.

The problem that arises here is that a similarity query usually returns a vast amount of results. The order of the returned results plays a very important role to the user needs; for instance, every user would wish the first result also to be the desirable one. Inspired from algorithms used in web search engines, such as the well-known Google's PageRank [2], many methods have been developed that rank results according to their importance. Especially in relational databases the notion of using web inspired algorithms for ranking is very common in recent years. Authors in [3] have developed a method for progressively identifying the top-k answers from a relational database, while authors in [4] propose a rank-join algorithm that makes use of the individual orders of its inputs to produce join results ordered on a user-specified scoring function.

The need for ranking exists in biological databases, as well. The idea of ranking genes by importance has appeared in [5], where the GeneRank algorithm was introduced showing great resemblance to PageRank. Google's PageRank was originally devised for assessing the importance of web pages in search engine results and was based on the premise that a web page should be highly ranked if other highly ranked pages contain hyperlinks to it. By analogy, GeneRank classifies a gene higher, if it is correlated, either in terms of GO annotations or in terms of expression levels, to other highly ranked genes. Ranking by importance can

*Corresponding author. Tel.: +30 2310 991924; Fax: +30 2310 991913; E-mail: niakovid@csd.auth.gr.

[1]See genome-www5.stanford.edu and www.ncbi.nlm.nih.gov.

ISSN 1872-4981/09/$17.00

draw attention to genes with low expression level but with an important regulation role among other genes. In [6] gene ranking is performed with the Kleinberg's HITS algorithm [7].

Therefore, given a query gene, we can rank search results either by their similarity to the query or by their importance, but not by both at the same time. This means that ranking by importance may show up several genes, while the similarity measure may show up some other different kind of genes. For example, according to Schilde et al. [8] some genes that are related to each other (which denote high similarity) are under-expressed in gskA null cells (which denote that they are not important). As will be shown later, the two rankings may differ significantly, i.e., results with high similarity may have low importance and vice versa. However, both rankings detect useful properties, i.e., functional relation vs. overall regulation role. Thus, it is necessary to devise a method allowing users to fuse the two rankings according to their needs.

Here, we investigate the problem of ranking genes by respecting at the same time the trade-off between similarity and importance and by taking into account user needs each time. We study two sets of data separately: one with expression values and one with GO terms. We propose a kernel-based fusion method that uses an intuitive parameter to express favor towards the one or the other ranking scheme and combine their advantages. The role of this regulatory mechanism plays the Kernel Similarity measures that have been proposed for the World Wide Web [9]. We also provide interesting results that prove the need for the existence of such a fusion method.

The remainder of the paper is organized as follows. Section 2 mentions previous related work on ranking genes. Section 3 describes our proposed fusion method as well as the terms similarity and importance. The following section provides an analytical and detailed description of the data sets used in our study, the derived experimental results and the whole process of ranking genes in general. Lastly, we draw some conclusions derived from this work.

2. Related work

Ranking of search results in microarray databases is a relatively new research problem. Chabalier et al. [1] suggest a clustering method that computes semantic similarities between genes. Clustering is performed according to different kinds of knowledge:

1. biological knowledge provided by the GO terms that are organized according to three hierarchies: Biological Process (P), Molecular Function (F) and Cellular Component (C),
2. medical knowledge supplied by the Unified Medical Language System (UMLS) [10],
3. genomic knowledge corresponding to the sequence features of the studied genes, and
4. expression pattern knowledge provided by experimental results. According to this method, the more information two genes share in common, the more functionally related they are and, thus, the different descriptions of each gene are taken into account.

Gudivada et al. [6] propose another gene prioritization method that detects functional relationship between genes and diseases. This method ranks genes according to model-driven semantic relationships and enables to utilize the combination of mouse phenotypes and human disease clinical features apart from GO and pathways in their prioritization approach.

GeneRank algorithm proposed by Morrison et al. [5] provides a method to assess the importance of a gene. Just like the ranking score of a web page will be high if it is linked to other highly ranked pages, the relative ranking of a gene will be increased if it is linked to other highly differentially expressed genes. In [11] authors suggest a multi-criterion approach to perform ranking of genes that are both biologically and statistically significant using signal processing methods.

Bie et al. [12] present an approach to provide answers to a recently identified problem in bioinformatics, which is to discover disease genes to diagnose and understand the biology of disease processes. Their method fuses gene datasets using kernels to perform disease gene hunting. A method for combining multiple kernel representations is also proposed by Lanckriet et al. [13]. The method is applied to the problem of predicting yeast protein functional classifications using a support vector machine (SVM) trained on several types of data.

3. Proposed fusion method

Here, we focus on two data types that are common in microarray databases:

(i) arrays including expression levels and
(ii) arrays including GO annotations.

In case (i), we have an $n \times k$ array A, where element $A(i, j)$ is the expression level (real number) of gene i to condition j. In case (ii), we have an $n \times k$ array A, where element $A(i, j)$ equals 1, if gene i is annotated with term j and 0 otherwise. To treat both cases equivalently, we use the Biclustering Analysis Toolbox[2] for case (i) to discretize the data to binary values. Henceforth, we use the A notation without ambiguity to refer to an inclusion matrix either for expression levels or GO annotations. In the sequel, we first describe separately ranking by similarity and by importance, and next the proposed method to fuse them.

3.1. Gene similarity

As mentioned above, two genes can be functionally related in several ways. For instance, they can be involved in a same biological process (for example, iron ion transport), where they can carry a specific molecular function (for example, ferric iron binding) or they can be involved in the same disease (for example, liver disease). Therefore, the more information two genes share in common, the more similar they are.

Co-citation coupling [14] is a classical means in Information Retrieval for defining relatedness between documents as the number of other documents that cite them both. Bibliographic coupling [15] also defines relatedness between two documents as the number of common references cited by the two. The two aforementioned measures can be formally defined as follows. Let A be an adjacency matrix of a citation graph. Then, the number of co-citations between nodes i and j is given by the (i, j)-element of the co-citation matrix $M = A^T A$. Similarly, the bibliographic coupling matrix $M = AA^T$ gives the values of bibliographic coupling. Since these matrices are symmetric, their graph counterparts, the co-citation graph and bibliographic coupling graph, are undirected.

In an analogous manner, when we have to do with gene similarity, relatedness between genes is measured by the co-occurrence of genes in an inclusion matrix. Thus, given an inclusion matrix A, apparently the matrix $M = A^T A$ is the co-occurrence matrix. As the similarity between two vectors is represented by the angle between these two vectors, then given a co-occurrence matrix M and by normalizing with the lengths, we can get the well-known cosine similarity matrix $\frac{M(i,j)}{(M(i,i) \cdot M(j,j))^{\frac{1}{2}}}$, which computes similarity between genes.

[2] Downloaded from http://www.tik.ee.ethz.ch/sop/bicat/.

3.2. HITS and gene importance

As already mentioned, co-citation coupling defines relatedness between documents, but computing importance of documents from their contents is a very difficult task. Citation counts have long been used as the index of document importance. Even though citations are made for various reasons, a positive correlation was observed between the number of citations and the significance or impact of the cited work.

Kleinberg's HITS algorithm, along the lines of PageRank, is a more recent and sophisticated method for evaluating document as well as web document importance. HITS discusses the problem of finding the "most relevant" web pages in response to a given broad query. The algorithm assigns the so-called authority and hub scores to each web document. An underlying assumption behind HITS is that mutual reinforcement relation exists between authorities and hubs: authoritative documents are cited by many hub documents and hub documents are those that cite many authoritative documents.

By analogy, we can use Kleinberg's HITS algorithm to assign authority and hub scores to each gene. In this case, a mutual reinforcement relation between authorities and hubs exists as well: authoritative genes are related to many hub genes, and hub genes relate to many authoritative genes. Let A be an inclusion matrix of genes. We apply HITS and compute the authority scores of genes, as the dominant eigenvector of the co-occurrence matrix $M = A^T A$.

3.3. Neumann kernels

Some formulations of link analysis measures that are intermediate between importance and relatedness have been introduced lately. These formulations are based on the family of symmetric positive semi-definite kernels [16], which define an inner product of graph nodes. Here, it is important to note that an intermediate concept between importance and relatedness might seem strange, since importance is a measure defined on individual nodes, whereas relatedness is defined between them. However, given an importance score vector v such as the HITS authority vector, vv^T defines a matrix where every row (and column) i gives a ranking of nodes identical to the one given by v except for an i such that $v(i) \neq 0$. Importance can thus be treated as a function over a pair of nodes, or a matrix, as well. The Neumann kernel was proposed for computing the semantic similarity between documents represented as set

of terms [17]. It defines document similarity and term similarity by using their complementary relation [18]. This bears reminiscence to the complementary relation of authorities and hubs in HITS. Given an inclusion matrix A and a parameter p, where $0 \leqslant p < 1$, we propose to first compute a C matrix:

$$C(i,j) = \frac{M(i,j)}{(M(i,i) \cdot M(j,j))^{\frac{1-p}{2}}} \quad (1)$$

The kernel matrix K_p, which defines the gene ranking, is equal to $C(I - p||C||^{-1}C)^{-1}$. Thus, the kernel ranking is user-controlled by tuning the parameter p in the range [0,1). The interpretation of Neumann kernels is as follows: When $p = 0$, clearly the kernel matrix K_p is actually the cosine similarity matrix. When p approaches 1, as described in [9], the ranking induced by the Neumann kernel is identical to the authority score of HITS. Thus, the kernel ranking yields more similar genes when p approaches 0 and it yields more important genes when p approaches 1.

4. Experimental results

We examined the properties of our proposed ranking approach (denoted as Kern) compared to ranking by similarity with the cosine measure (denoted as Sim) and by importance determined by the HITS algorithm (denoted as HITS). We used two microarray data sets. We ranked separately the two datasets first by using Sim and HITS and then by using the proposed fusion method Kern.

The first dataset contains expression levels of size 6,152 rows $\times$ 173 columns.[3] Rows represent genes, whereas columns represent conditions. Every value corresponding to a particular gene and a particular condition represents the expression level of the gene when this specific condition is applied. This data set is a combination of many microarrays and the data contained in it represent the normalized, background-corrected log2 values of the Red/Green ratios measured on the DNA microarrays.

The second dataset contains GO annotations with terms related to three aspects, i.e., C, F, and P (32,724 lines in total).[4] C stands for cellular component, F for molecular function and P for biological process.

[3]Downloaded from www.genome.stanford.edu/yeast_stress/data.shtml.

[4]Downloaded from dictybase.org/Downloads.

Table 1
KMin distances.

p	Expression Data		GO Data	
	D(K,H)	D(K,S)	D(K,H)	D(K,S)
0	0.34	0	0.45	0
0.5	0.15	0.18	0.23	0.26
0.9999	0.02	0.33	0.02	0.46

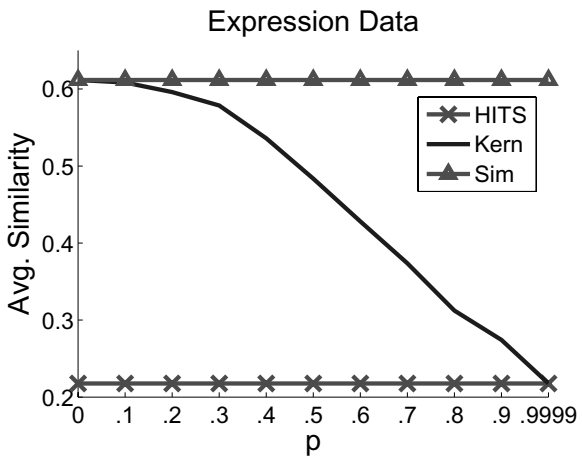

Fig. 1. Similarity measure for expression data.

The structure of a GO term is quite simple. In general, each GO entry consists of a term name (e.g. cell) and a zero-padded seven-digit identifier (or accession number) prefixed by GO: (e.g. GO: 0005623), which is used as a unique identifier and database cross-reference. Terms may have one or more secondary IDs, alternate IDs that refer to them. GO terms should be equipped with a text definition, which includes an indication of the definition source. Terms may also have a comment, which gives more information about the term and its usage.

From the two datasets we formed two inclusion matrices. In each case, we considered each gene as the query gene, and ranked the rest genes according to the three approaches (Kern, Sim and HITS). We kept the top k results (default $k = 20$) and computed the normalized Kendall distance D that counts the number of pair wise disagreements between two lists [9]. Number 0 corresponds to the minimum distance, whereas number 1 corresponds to the maximum distance, represented by value p. As mentioned before in this paper parameter p is set by the user in the range [0,1). Table 1 presents the average distances between Kern (K) and HITS (H) and between Kern (K) and Sim (S), for several p values. When $p = 0$, as expected, Kern and Sim produce identical results. When p approaches 1, then its distance from HITS approaches 0. For intermediate

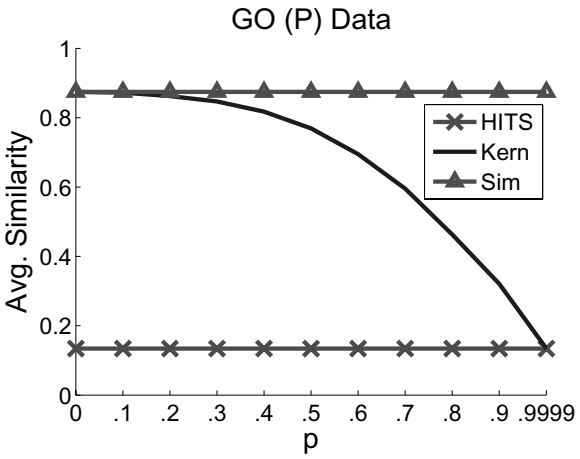

Fig. 2. Similarity measure for GO data.

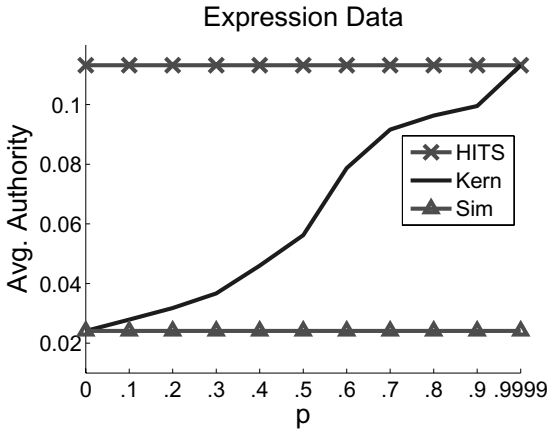

Fig. 3. Authority measure for expression data.

p values, Kern is an amalgamation of the two others, as it is about equally distant from them.

Next, we measured the average similarity and authority (namely the HITS score that indicates importance) of the k results by varying p. Figures 1 and 2 illustrate the similarity measure for the two inclusion matrices, whereas Figs 3 and 4 depict the authority measure for the two inclusion matrices as well. Clearly, HITS and Sim differ significantly in all four cases. The former produces results that are important (high authority) but not similar and the latter produces results that are quite similar but not important. An example from real life that confirms our statements comes from the authors of [8]. Gene 2C that encodes polypeptides of similar size is related to other 5 genes ($DDB_G0280871$, $DDB_G0281003$, $DDB_G0280953$, $DDB_G0282317$ and $DDB_G0281019$). Relation between all these genes denotes similarity. However, on

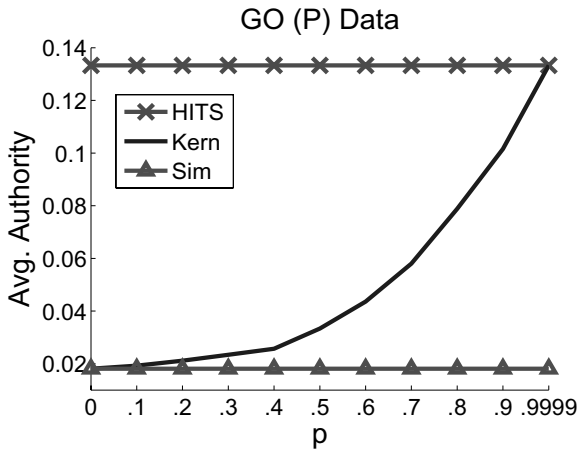

Fig. 4. Authority measure for GO data.

the other hand, all six genes are under-expressed in gskA (Glycogen Synthase Kinase) null cells, which mean that they are not important under these particular circumstances. Our proposed ranking approach Kern adapts smoothly between the two aforementioned cases (similarity and importance), either we have expression or GO data and can be tuned according to the user requirements.

5. Conclusions

Existing methods that rate results according to similarity or importance are unable to provide the user with information about both the aforementioned measures. We proposed a kernel-based method for ranking retrieval results for microarray data. Our results indicate that the proposed scheme can fuse the two measures and detect genes that are both similar and important. It can also adapt to the users' requirements, by varying the parameter p in the range [0,1).

References

[1] J. Chabalier, J. Mosser and A. Burgun, A transversal approach to compute semantic similarity between genes, *BMC Bioinformatics* **8** (2007), 1–12.

[2] S. Brin and L. Page, The anatomy of a large-scale hypertextual (web) search engine, *Computer Network and ISDN Systems* **30**(1–7) (1998), 107–117.

[3] G. Li, J. Feng, F. Lin and L. Zhou, Progressive ranking for efficient keyword search over relational databases, in: *Proceedings of the 25th British National Conference on Databases* (*BNCOD*), Cardiff, UK, 2008, 193–197.

[4] I.F. Ilyas, W.G. Aref and A.K. Elmagarmid, Supporting top-k join queries in Relational Databases, *The VLDB Journal* **13**(3) (2004), 207–221.

[5] J.L. Morrison, R. Breitling, D.J. Higham and D.R. Gilbert, GeneRank: using search engine technology for the analysis of microarray experiments, *BMC Bioinformatics* **6** (2005), 1–12.
[6] R.C. Gudivada, X.A. Qu, A.G. Jegga, E.K. Neumann, B.J. Aronow, A genome Phenome integrated approach for mining disease-causal genes using semantic Web, in: *Proceedings of the Workshop on Healthcare and Life Sciences*, 2007.
[7] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM* **46** (1999), 604–632.
[8] C. Schilde, T. Araki, H. Williams, A. Harwood and J.G. Williams, GSK3 is a multifunctional regulator of dictyostelium development, *Development* (*published by The Company of Biologists*) **131** (2004), 4555–4565.
[9] D.J. Cook and L.B. Holder, *Mining Graph Data*, John Wiley & Sons, 2007.
[10] O. Bodenreider, The unified medical language system (UMLS): Integrating Biomedical Terminology, *Nucleic Acids Research* **32** (2004), Database issue: D267–270.
[11] A.O. Hero, Gene Selection and Ranking with Microarray Data, in: *Proceedings of the 7th IEEE International Symposium on Signal Processing and its Applications* (*ISSPA*), Paris, France, 2003, 457–464.
[12] T.D. Bie, L.C. Tranchevent, L.M.M. van Oeffelen and Y. Moreau, Kernel-based data fusion for gene prioritization, *ISMB/ECCB* **23** (2007), i125–i132.
[13] G.R.G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan and W.S. Noble, Kernel-based data fusion and its application to protein function prediction in yeast, in: *Proceedings of the Pacific Symposium on Biocomputing*, 2004, 300–311.
[14] H. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents, *Journal of the American Society for Information Science* **24** (1973), 265–269.
[15] M.M. Kessler, Bibliographic coupling between scientific papers, *American Documentation* **14**(1) (1963), 10–25.
[16] B. Scholkopf and A.J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge, MA, 2002.
[17] J. Kandola, J. Shawe-Taylor and N. Cristianini, Learning semantic similarity, in: *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems* (*NIPS*), Vancouver, Canada, 2003, 673–680.
[18] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas and R.A. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science* **41**(6) (1990), 391–407.