Quantifying the biological similarity between gene products using GO: an application of the vector space model

Spiridon C. Denaxas[1] *Student Member* IEEE, Christos Tjortjis[1] *Member* IEEE

[1] School of Informatics, University of Manchester, PO Box 88, Manchester, M16 1QD, UK

S.Denaxas@postgrad.manchester.ac.uk, christos.tjortjis@manchester.ac.uk

***Abstract*— Recent advances in biological experiments, such as DNA microarrays, have produced large multidimensional data sets for examination and analysis. Scientists however, heavily rely on existing biomedical knowledge in order to fully analyze and comprehend such datasets. The approach we propose combines statistical natural language processing techniques with the GO annotation ontology, for assessing the biological relatedness of gene products clusters. We explore the application of the vector space model as a means of quantifying this relatedness between gene products, based on their underlying biological properties, as indicated by the GO terms associated with them. We report on experimental results on a small subset of *saccharomyces* gene products. We also propose and validate a biological similarity figure of merit which can assess gene expression cluster analysis results. Finally, we deploy our approach combined with hierarchical clustering in order to illustrate its application to gene expression clustering experiments.**

Index Terms— **Clustering, Gene Ontology, Text Mining, Vector Space Representation.**

I. Introduction

Recent advances in biological experiments such as DNA microarray technology have made it possible to simultaneously monitor the expression levels of thousands of genes in parallel during important biological processes and across large collections of samples, providing insight into gene functionality and their regulatory mechanisms. Microarrays enable researchers to identify and comprehend genes and their respective functions that would have otherwise remain unknown.

Large scale experiments like this however induce and heavily rely on massive amounts of generated information. The measured patterns during such experiments are very often explained retrospectively by examining and analyzing the underlying biological properties of the respective gene products composing the data set. Thus, the amount of scientific discoveries, hypotheses and cross-references, stored mainly in raw text format across a number of specialized systems, is growing rapidly.

Existing biological knowledge is critical in order to comprehend such data sets. Researchers have argued towards the effectiveness of deploying computational methods that incorporate external information sources in order to assist the interpretation and organization of such experiments [1]. External information sources include ontology-based knowledge, primary and secondary sequence databases and medical literature. Published scientific text contains a distilled version of the most biologically significant discoveries and is a potent source of information for integrating in experiments [2].

A number of solutions yielding high accuracy results exist but they often rely on the integration of information from a number of external information sources such as MEDLINE, making them less flexible and perhaps in many cases organism oriented. A more detailed overview of these approaches is provided in section 2.

Our approach demonstrates that statistical text processing techniques can be deployed solely on the Gene Ontology and the information therein and yield fruitful results. We feel that solutions which are based mainly on medical literature, such as MEDLINE abstracts and raw text, offer a broader notion of similarity between gene products since biomedical literature contains knowledge regarding gene relations discussed in a variety of contexts.

On the other hand, Gene Ontology (GO) annotation terms are specific and explicitly denote a gene product's molecular function, the biological process in which it takes part in or the molecular component in which it resides [3]. Thus, making extensive usage of the GO annotation terms will provide more specific biomedical information and a more accurate measure on the correlation between gene products.

Our main goal is to develop an approach that exploits and incorporates the vast amounts of biological information GO offers in the analysis of groups of genes; for example gene

clusters resulting from high-throughput gene expression analysis experiments. Unlike the majority of existing approaches, the framework developed solely operates on the information within the GO ontology and does not rely on external information sources.

The remaining of the paper is organized as follows: Section 2 reviews related work. Section 3 details the methods used in the proposed approach. Section 4 presents experimental results. Section 5 discusses and evaluates the results and concludes the paper with directions for further work.

II. Related work

Information retrieval, text mining and statistical natural processing methods have been recently deployed in order to discover and assess the biological similarity between individual pairs and clusters of genes based on existing literature. The majority of methods use existing biomedical databases containing textual information on gene products such as MEDLINE [4] and SWISS_PROT [5]. Additionally, several methods use the GO annotation as source of existing knowledge for both analyzing and evaluating results from large scale biological experiments and yield encouraging results.

Raychaudhuri et al. recently developed the Neighbour Divergence per Gene (NDPG) concept in order to assess the functional coherency of a group of genes by utilizing existing knowledge from public repositories such as MEDLINE [6]. Based on the GO gene annotation ontology, Gibbons & Roth developed a method to judge the quality of gene expression clustering methods [7] and combined it with the Saccharomyces Genome Database (SGD) database [8] and existing datasets [9,10] Glenisson et al. [11] evaluated the vector space representation [12] in text-based clustering of genes selected from the MIPS [13] functional catalogue.

Lord et al. [14] used a similar approach when they explored the semantic similarity between GO terms by making use of Resnik's [15] notion of *shared information content.* Similarity between annotation and literature has also been shown to augment sequence similarity searches. In their work, Chang et al. [16] augmented PSI-BLAST [17] with similarity scores calculated over the annotations and MEDLINE references cited by entries retrieved by the individual sequence similarity searches. Finally, Karypis et al. describe a method of textual analysis of documents associated with pairs of genes and describe how their approach can be utilized for discovering, identifying and annotating functional relationships among genes [18].

III. Methods

A. Constructing gene profiles

Ontologies are the most common form for the representation of knowledge in the bioinformatics community. An ontology is the specification of the key concepts in a given field of operations combined with the description of the relationships that exist amongst these concepts. In the majority of cases, an ontology is composed by a strictly controlled vocabulary. Additionally, the relationships between the concepts are established as axioms that capture the network structure of the knowledge that they model.

A number of different ontologies have been developed in the past years and have been widely used in the bioinformatics field such as the Unified Medical Language System (UMLS) [19] and the Gene Ontology. The GO ontology consists of a widely accepted and standardized gene annotation vocabulary used by scientists in order to express and define in a clear and concise manner certain biological attributes about a specific gene. GO consists of three separately structured ontologies called molecular function, biological process and cellular component. Biological process refers to the biological objective in which the gene or gene product contributes to. The molecular function ontology denotes the biochemical activity of a gene and finally, the cellular component refers to the place in the cell where the gene product resides.

Every GO term follows the *true path rule*: *the pathway from a child term all the way up to its top-level parent(s) must always be true*. If a specific child term describes a gene product, then all its parents also apply to that gene product. By exploiting this rule we are able to construct more accurate and concise gene profiles since additional GO terms are assigned to each gene product.

We used the SGD database to construct a smaller gene subset, consisting of 88 genes from three biologically distinct groups. The first group contains genes related to the DNA metabolism biological process, the second group is related to the process of transport and finally genes composing the third group are involved in the yeast sporulation process. A detailed summary of the gene product subset that was constructed can be seen in Table I.

For every gene product, the path from its assigned GO term up to the root node of the ontology is extracted. This is easily achieved by querying a local version of the latest GO relational database port and parsing the results. This effectively assigns a set of GO terms to the gene product. For every GO term assigned to the gene, the *definition* field is extracted from the GO ontology and appended to the genes textual profile. For example, as seen in Fig. 1, the textual profile constructed for the APN1 will include the textual information extracted from the *definition* fields from the GO:0006281, GO:0006259, GO:0006139, GO:0044237, GO:0050875 and GO:0007582 annotation terms which were previously associated with the gene product by exploiting the *true path* rule. Standard stemming algorithms where applied [20] on each profile.

Table I: A summary of the respective GO terms which compose the yeast subset used

Biological group	Term name	Number of genes
sporulation	sporulation	*13*
	sporulation (sensu funghi)	*19*
transport	amino acid transport	*15*
	aromatic amino acid transport	*1*
	basic amino acid transport	*7*
	neutral amino acid transport	*4*
DNA metabolism	DNA repair	*5*
	mismatch repair	*11*
	bypass DNA synthesis	*1*
	error-free DNA repair	*4*
	postreplication repair	*8*

B. Vector space model representation

We encoded the individually constructed gene text profiles using a bag-of-words following the vector space model paradigm. The vector space model effectively encodes an entire document into a k-dimensional vector which represents the terms found within the document and their occurrence. The grammatical structure of the document is generally ignored and terms are individually extracted, therefore making this approach also known as *bag-of-words.*

In the vector space model representation, a document is represented by a weighted vector (also known as a profile) of which each individual component corresponds to a single term from the entire set of terms within the constructed vocabulary [21]. For every term found in the document, a value denotes its presence and is represented by a weight within the documents profile as shown in equation (1).

$$d_i = (w_{i,1}, w_{i,2}, \ldots w_{i,N}) \quad (1)$$

Each weight w_{ij} within the document vector d of document *i*, represents the weight of term j from the vocabulary of size *N*.

The individual weights representing terms found within the document are calculated during the indexing operation. A number of popular indexing schemes exist and were taken into consideration [22].

Eventually, we used the IDF indexing scheme in order to minimize the noise within our data set and additionally minimize the impact of very common biological terms. Automatic indexing of the profiles as well as stop word elimination was performed by using the *doc2mat* script; a part of the CLUTO toolkit [23].

C. Quantifying biological similarity

Similarity between a pair of documents d_i and d_j is calculated by measuring the cosine of the angle between the normalized weighted vectors representing the two documents [24], as shown in equation (2):

$$\mathrm{sim}(d_i,d_j) = \cos(d_i, d_j) \quad (2)$$

The same concept applies when calculating the similarity between a document d_i and a query document d_j. The underlying hypothesis behind this statistical approach for assessing document similarity states that a high degree of similarity between the documents also denotes a high degree of relevance and semantic similarity between them.

Based on this concept, we can define a similarity metric which can be used to quantify the functional relationship between individual GO terms assigned to genes. Subsequently, the metric can act as a measurement of biological relatedness between pairs of genes that the respective terms have been assigned to. Since the text profiles constructed for the gene products essentially describe their biological properties, should two genes share common biological properties, they will also share a very high degree of similarity between their associated text profiles.

Based on this notion, given two genes i and j, represented by their previously constructed textual profiles di, dj we define *BIOsim* as the cosine of the angle between the normalized weighted vectors representing their individual textual profiles (3).

$$\mathrm{BIOsim}(i,j) = \cos(di,dj) \quad (3)$$

Gene products which share a high degree of biological correlation will have BIOsim values closer to 1 whereas lower values towards zero will illustrate a very low degree of similarity. Similarly, we can also assess and quantify the biological relatedness and coherency of a group of genes based on the same metric notion. Given a group of genes, we can define the clusters functional coherence, *BIOco*, based on the arithmetic mean of their normalized weighted vector representations (4).

$$\frac{1}{n}\sum(BIOsim(i,j) = \cos(di,dj)) \quad (4)$$

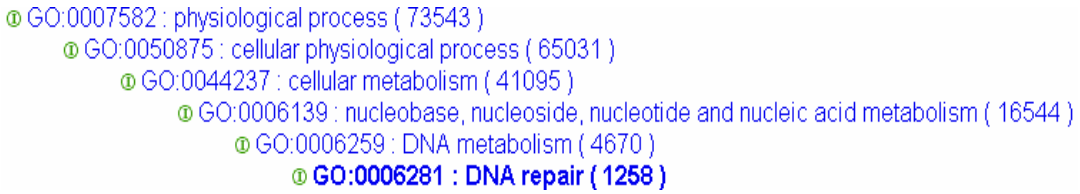

Fig. 1: A complete path from the DNA repair annotation term up to the top level parent

Based on (8), clusters which are biologically coherent will have a BIOCo value close to 1 whereas lower values will denote smaller degrees of biological relatedness shared between the gene products composing the cluster. Both measures are able to quantify the biological similarity between individual pair of genes or a cluster of genes based on the medical and biological knowledge extracted from their associated GO annotation terms.

IV. Results

A. *Validating functional similarity*

In order to validate our approach, we initially clustered all individual gene products based on the respective document profiles we constructed. Initially, we attempted to cluster the profiles in three clusters, each of them representing the three major biological processes in which the gene products belong too as seen in Table 1.

In order to cluster the profiles, we used CLUTO, a software package for clustering high dimensional data [23]. Using the cosine similarity as a metric, in the first run all 88 genes were correctly clustered in the three major clusters representing the biological processes. Fig. 2 illustrates the matrix representation of the first cluster produced which contains the 32 gene products composing the *sporulation* category. A term represented with a bright grey colour illustrates greater presence of that term within the cluster, whereas lighter shades of red and white depict lower values and zero respectively. The 12 most descriptive stemmed features, with their respective percentage, for the cluster containing the gene products composing the sporulation biological process, which best illustrate it can be seen in Table II.

During the second run, we set the total number of clusters to 11 in order to explore the possibility of identifying all the relevant biological sub-groups existing within the data set and grouping the genes together accordingly. Using the cosine similarity as a metric of similarity between textual profiles, all gene products were clustered accurately and assigned to the relevant cluster which denoted their respective biological property.

Using these two scenarios we were able to validate the correctness of our approach since all gene products were correctly clustered with biologically similar genes based on the similarity of their textural profiles. Using a similar line of attack, we attempted to calculate the coherency of the entire dataset. By forcing all gene products in one cluster we were able to quantify the biological coherency score of the cluster (*0.281)*.

Table II: The features for the cluster which contains the gene products composing sporulation

Feature	Percentage	Feature	Percentage
reproduct	25.1%	format	19.6%
spore	19.6%	sporul	19.6%
fungi	6.3%	taxonomi	1.6%
funghi	1.6%	ncbi	1.6%
psysiolog	1.6%	4751	1.6%
sensu	1.6%	organ	0.3%

The low value we obtained denotes a very small degree of biological relatedness between the gene products composing the cluster – something predictable since all three biological groups were clustered together. Similarly one could quantify the biological coherency of other gene clusters and use the obtained values in order to prioritize clusters for further analysis.

B. *Further experiments*

Further experiments were carried on actual budding yeast *S. cerevisae* data, as collected from microarray experiments [10]. For this purpose, we made use of the dataset utilized by Eisen et al during their data clustering experiments. Similar to the original experiments, we applied *pairwise average-linkage* cluster analysis to the gene expression dataset using a form of correlation coefficient similar to Pearson's correlation coefficient.

A striking result of the process is the tendency of large groups of genes which are clustered together to share common biological properties; more specifically a strong display of similarity in the biological process area. This validates one of the basic assumptions under which microarray scientists operate on, the fact that genes which share common expression patterns are most likely to share common biological properties as well.

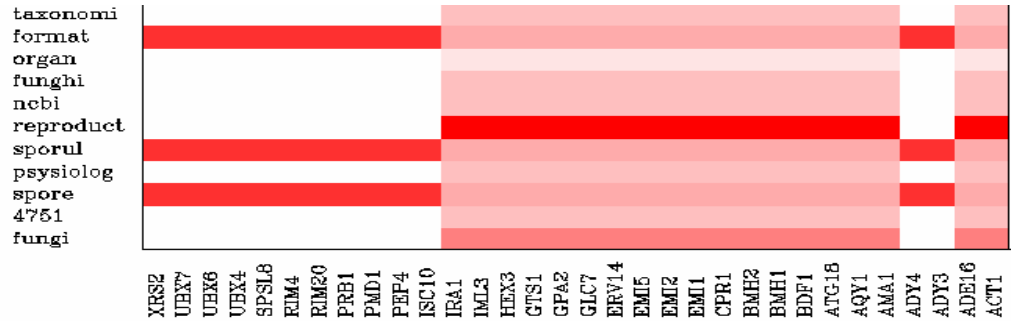

Fig. 2: A matrix representing the cluster which contains the gene products composing sporulation.

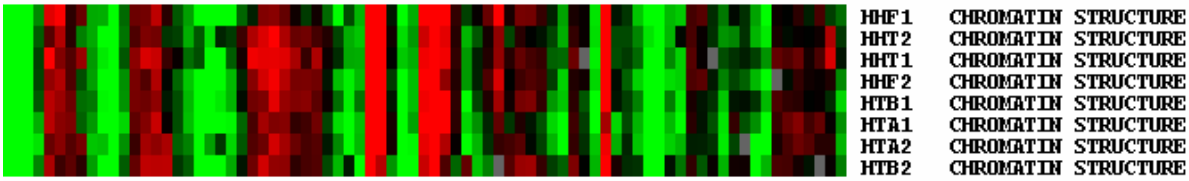

Fig. 3: Clustered display of the eight histone genes which are clustered together. These genes essentially duplicates of the histones and it has been shown elsewhere that they are coregulated at a particular point of the cell cycle.

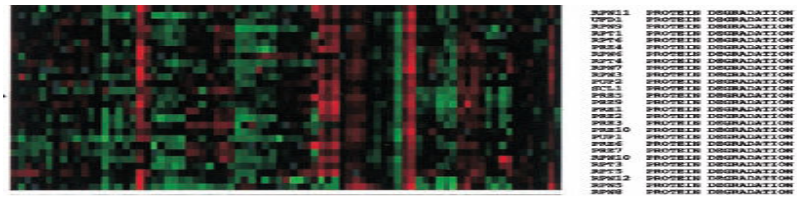

Fig. 4: Clustered display of the 27 genes which are clustered together and are involved in the proteasome. The assigned *BIOsim* value of 1 denotes a perfectly functional coherent cluster since all of the genes composing it share an identical biological_process term from the GO. Partial image segment extracted from [10].

Initially, two very tight clusters immediately stand out from the results and are displayed in Fig. 3 and Fig. 4. The first cluster displayed in Fig. 3 is composed by eight histone genes which are essentially duplicates of the histones H2A, H2B, H3 and H4. Hereford et al. showed that these genes display similar regulation patterns at a particular point of the cell cycle [25].

Similarly, the cluster displayed in Fig. 4, contains 27 genes which encode the bulk components of the protease. Both clusters immediately stand out from the hierarchical tree constructed during the process since they both have a *BIOsim* value of **1**.

The glycolysis cluster contained 15 gene products involved in the biological process of glycolysis within the cell. Additionally, it also contains the TKL1 gene product which takes part in the pentose phosphate cycle process and ACS2 which takes part in the acetyl-conenzyme biosynthesis process. The calculated *BIO*sim score for the cluster was **0.723**.

V. Conclusion

In this paper we described a statistical natural language processing approach based on the vector space model in order to assess and quantify the biological similarity between pairs and clusters of gene products. Our main aim was to explore the potential of utilizing the vector space model solely on biological information extracted from the GO terms associated with individual gene products.

By exploiting the *true path* rule, we associated a number of GO terms with each gene product, the terms which compose the path from its assigned term up to the parent term of the taxonomy. We then constructed a textual profile of an average of 150 terms based on the *definition* field of the respective terms. Since the textual profiles constructed essentially describe the underlying biological properties of the gene products, a high degree of semantic similarity between the profiles translates to a high degree of biological similarity between the gene products. We are able to measure and quantify the biological relatedness between gene products and clusters composing them by calculating the dot product

between pairs and the average dot product between genes composing a cluster respectively. Values close to 1 denote a high degree of biological similarity and coherency respectively whereas values closer to 0 denote a very low degree of similarity. In order to validate our approach and obtain some initial experiments we constructed a small subset of 88 *saccharomyces* genes from 3 distinct biological groups. We then constructed their individual text profiles and clustered the associated gene products based on the degree of semantic similarity between them.

One of the main aims in our research is the application and integration of the above mentioned approach within the context of gene expression clustering. We have previously explored this approach by developing a graph oriented approach to assessing a clusters biological coherency based on GO [26].

We are also currently working on constructing *query documents* in order to accurately identify the dominant biological properties of a potential cluster of genes. A query document will essentially be a text profile which will contain a large number of features, all associated with a major biological concept. A high degree of similarity between the textual profiles of the gene products composing the cluster and the actual query document is a very good indication that the biological property it describes is dominant within the cluster. Finally, we are considering the implementation of a *weighting* scheme for the respective annotation terms assigned to each gene product as previously explored in [27].

REFERENCES

[1] Altman, R.B., Raychaudhuri, S., "Whole-genome expression analysis: challenges beyond clustering", *Current Opinion on Structured Biology*, vol. 11, 2001, pp. 340-347.

[2] Raychaudhuri, S., Chang, J., Imam, F., Altman, B., "The computational analysis of scientific literature to define and recognize gene expression clusters", *Nucleic Acids Research*, vol. 31, no. 15. 2003, pp. 4553-4560.

[3] Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology", *Nature Genetics*, vol. 25, 2000, pp. 25-29.

[4] Schuler, G.D., Epstein, J.A., Ohkawa, H., Kans, J., "Entrez: Molecular biology database and retrieval system", *Methods Enzymology*, 266, pp. 141-162.

[5] Bairoch, A., Boeckmann, B., "The SWISS-PROT protein sequence data bank", *Nucleic Acids Research*, vol. 20, 1992, pp. 2019-2022.

[6] Raychaudhuri, S., Schutze, H., Altman, B., "Inclusion of textual documentation in the analysis of multidimensional data sets: application to gene expression data", *Machine Learning*, vol. 52, 2003, pp. 119-145.

[7] Gibbons, F., Roth, F., "Judging the quality of gene expression-based clustering methods using gene annotation", *Genome Research*, vol. 12, 2002, pp. 1574-1581.

[8] Cherry, J et al., "SGD: Saccharomyces Genome Database", *Nucleic Acids Research*, vol. 26, no. 1, 1998, pp. 73.79.

[9] Cho, R. J., Cambell, M.J., Winzeler, E.A., et al., "A genome-wide transcriptional analysis of the mitotic cell cycle", *Mol. Cell*, vol. 2, 1998, pp. 65-73.

[10] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci*., vol. 95, 1998, pp. 14863-14868.

[11] Glenisson, P., Antal, P., Mathys, J., Moreau, Y., De Moor, B., "Evaluation of the vector-space representation in text-based gene clustering", *Pacific Symposium on Biocomputing*, 2003.

[12] Raghavan, V. V., Wong, K. M., "A critical analysis of vector space model for information retrieval", *Journal of the American Society for Information Science*, vol. 35, no. 5, 1986, pp. 279-287.

[13] Mewes, H., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., Frishman, D., "MIPS: a database for genomes and protein sequences", *Nucleic Acids Research*, vol. 27, no. 1, 2002, pp. 44-48.

[14] Lord, P., Stevens, R., Brass, A., Goble, A., "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation", *Bioinformatics*, vol. 23, no. 10, 2003, pp. 1275-1283.

[15] Resnik,P., "Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language". *J. Artif. Intelligence Res.*, vol. 11, 1999, pp. 95–130.

[16] Chang, J., Raychaudhuri, S., Altman, R., "Including biological literature improves homology search", Pac. Sym. *Biocomputing*, 2001, pp. 374-383.

[17] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Research*, vol. 25, 1997, pp. 3389-3402.

[18] Karypis, G., "CLUTO – a clustering toolkit", Technical Report, *TR 02-017*, Department of Computer Science, University of Minnesota, 2002.

[19] Barnett, G.O., Humphreys, B.L., Lindberg, D.A., Schoolman, H. M., "The unified medical language system: An information research collaboration", *Journal of the American Medical Information Association*, vol. 5, 1998, pp. 1-11.

[20] Porter, M.F., "An algorithm for suffix stripping", *Program*, vol. 14, no. 3, 1980, pp. 130-137.

[21] Baeza-Yates, R., Ribeiro-Neto, B., "Modern information retrieval", ACM Press / Addison-Wesley, 1999.

[22] Korfhage, R., "Information storage and retrieval", Wiley Computer Publishing, New York, 1999.

[23] Karypis, G., Nakken, S., Kauffman, C., "Finding functionally related genes by local and global analysis of MEDLINE abstracts", *SIGIR04 Bio Workshop: Search and Discovery in Bioinformatics*, 2004.

[24] Manning, D., Schutze, H., "Foundations of statistical natural language processing", The MIT Press, 2003.

[25] Hereford, L. M., Osley, M.A., Ludwig, T.R., McLaughlin, C.S., "Cell-Cycle regulation of yeast histone MRNA", *Cell*, vol. 24, no. 2, 1981, pp. 367-375.

[26] Denaxas, S.C., Tjortjis, C., "A hybrid knowledge-driven approach to clustering gene expression data", Proc. 10th *Pan'c Conf. on Informatics (PCI'2005)*, 2005, pp.205-216.

[27] Bodenreider, O., Aubry, M., Burgun, A., "Non-lexical approaches to identifying associative relations in the gene ontology", *Pacific Symposium on Biocomputing*, 2005, pp. 91-102.