Appears in Proceedings of the Data Mining and Text Mining for Bioin- formatics Workshop at PKDD2003

Improving Information Extraction through Biological Correlation

Francisco M. Couto, Mário J. Silva
LaSIGE, Departamento de Informática
Faculdade de Ciências
Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
{fjmc,mjs}@di.fc.ul.pt

Pedro Coutinho
UMR 6098, Architecture et Fonction des
Macromolécules Biologiques
Centre National de la Recherche Scientifique
13402 Marseille CEDEX 20, France
pedro@afmb.cnrs-mrs.fr

ABSTRACT

We present a new method for improving the efficiency of information extraction systems applied to biological literature, using the correlation between structural and functional classifications of gene products. The method evaluates extracted information by checking if gene products from a common family match a common set of biological properties. To evaluate the method, we implemented it in a case-study, where the method annotated carbohydrate-active enzymes with functional properties extracted from literature. Each carbohydrate-active enzyme is assigned to one or more families of catalytic and carbohydrate-binding modules according to its modular structure. To compute the relatedness between functional properties, we implemented a semantic similarity measure in GO, a biological ontology. The results present our quantitative measure of the correlation between the modular structures and functional properties, showing that our method is a viable approach for automatic validation of extracted biological information.

1. INTRODUCTION

Relevant facts discovered in molecular biology research, like in other fields, have been mainly published in scientific journals throughout the last century [9]. Extracting knowledge from this large amount of unstructured information is a painful and hard task, even to an expert. The solution was to create and maintain structured databases, such as GenBank and SwissProt that collect and distribute biological information, in particular biological sequences. These databases describe properties of common biological entities, such as genes and proteins. In the past few decades, the explosion of data has caused the exponential growth of these databases [3], where efforts to compensate the lack of annotation of many entries (mostly genomic) are at the origin of the significant misannotations, underprediction and overprediction of properties found today in biologic databases [5]. The integration of literature-derived annotation to different sources of data corrects and completes our knowledge about these biological entities [16]. However, a substantial amount of knowledge important to the integration is still only recorded in literature [18], which motivates the development of automatic tools that could extract part of this knowledge.

Information extraction methods find relevant information in unstructured texts and encode it in a structured form, like a database [11]. The application of these methods to biological literature is a recent research topic with a high activity despite its youth [2, 8, 17]. However, the use of different nomenclatures, different data classifications, and misannotations are hard barriers to overtake.

Our work aims to enhance information extraction systems for automatic annotation of biological databases through a new method, which we named CAC (Correlate the Annotations' Components). It evaluates whether an annotation is valid or not, based on the biological correlation between structure and function of gene products [14]. To check the effectiveness of CAC, we implemented it in case-study, where CAC annotated carbohydrate-active enzymes with functional properties extracted from literature. From these annotations, we identified a correlation between biological structure and function by using a semantic similarity measure [10].

The rest of this paper is structured as follows. Section 2 describes CAC method in detail. In Section 3 we present our case-study, describing its sources of biological information, the validation process, and the results. Section 4 discusses related work. Finally in Section 5 we express our main con-

clusions and directions for future work.

2. CAC

In CAC, we restrict an annotation to a pair composed by a gene product and a biological property. The gene products have to be classified in families according to their structural information, and the biological properties have to be organized in an ontology structured as a graph. CAC aims to validate an annotation only when its components have a biological relationship between them.

We define that two annotations converge if they relate different gene products from a common family with similar biological properties. CAC assumes that an annotation is valid if it has a significant number of convergent annotations in any of its gene product's families. This assumption is supported by the dogma of molecular biology, which postulates that sequences should be correlated with their biological activity, i.e. gene products from a common family usually share a common set of biological properties. To validate an annotation in a family, it is not necessary to have all gene products from the family sharing the biological property, but only a significant subset of them.

Similarity of gene products and biological properties are fuzzy concepts, but we can still define metrics to estimate them. In our case, we need to define two type of metrics:

- Given two gene products g_1 and g_2 from a common family, we express their *structural distance* as $\Delta(g_1, g_2)$.
- Given two biological properties p_1 and p_2, we express their *functional distance* as $\Delta(p_1, p_2)$.

Since it is only necessary to calculate the structural distance between gene products from a common family, we should calculate the structural distance according to factors that characterize the family. For instance, we can measure the sequence similarity of common modules using BLAST (Basic Local Alignment Search Tool) [1]. Functional distance between biological properties can be measured through semantic similarity measures, which details are described in section 2.1

We define a measure of the convergence between two annotations as being proportional to its structural distance and inversely proportional to its functional distance. Without loss of generality, we consider a set of annotations $\mathcal{A}_f$ whose gene products belong to a common family f.

Definition 1. Given two annotations $(g_1, p_1) \in \mathcal{A}_f$ and $(g_2, p_2) \in \mathcal{A}_f$, their *annotation convergence* is defined as:

$$\Gamma((g_1, p_1), (g_2, p_2)) = \frac{\Delta(g_1, g_2)}{\Delta(p_1, p_2)}$$

We can visualize the notion of annotation convergence by representing an annotation as an arrow from a gene product to a biological property. If two arrows start from distant locations and finish in close locations, then they are convergent. In other words, two annotations are convergent if their structural distance is larger than their functional distance.

EXAMPLE 1. *Figure 1 presents three different cases for the annotations a_1=(g_1, p_1), a_2=(g_2, p_2) and a_3=(g_3, p_3), whose components we placed on the shaded regions according to their structural or functional distance. In case (a) we have $\Delta(g_1, g_2) = \Delta(g_2, g_3) = \Delta(g_1, g_3)/2$ and $\Delta(p_1, p_2) = \Delta(p_2, p_3) = \Delta(p_1, p_3)$, which makes $\Gamma(a_1, a_3) > \Gamma(a_1, a_2)$. In case (b) we have $\Delta(g_1, g_2) = \Delta(g_2, g_3) = \Delta(g_1, g_3)$ and $\Delta(p_1, p_2) = \Delta(p_2, p_3) = \Delta(p_1, p_3)/2$, thus $\Gamma(a_1, a_2) > \Gamma(a_1, a_3)$. Finally, in case (c) we have $\Gamma(a_1, a_2)$=0 and $\Gamma(a_2, a_3)$ =∞, since a_1 and a_2 annotate the same gene product (origin) whereas a_2 and a_3 annotate the same biological property (destiny).*

Since we defined the concept of annotation convergence by two measures from different universes, it is not reasonable to establish a coefficient from which we can consider the annotations convergent. However, it is possible to define a more flexible relation that considers two annotations convergent if their annotation convergence is greater than a certain value.

Definition 2. Given two annotations $a_1 \in \mathcal{A}_f$ and $a_2 \in \mathcal{A}_f$, and a threshold h, they are *h-convergent* if $\Gamma(a_1, a_2) \geq h$.

Finally, given a threshold h, we define the correlation degree of an annotation as the number of h-convergent annotations in a common family.

Definition 3. Given an annotation $a_0 \in \mathcal{A}_f$ and a threshold h, the *correlation degree* of a_0 for h is defined as $\mathcal{D}_h(a_0) = \#\{a_x : a_x \in \mathcal{A}_f \wedge \Gamma(a_0, a_x) \geq h\}$.

The method validates annotations that have a correlation degree larger than a certain value in at least one family. This value and the convergence threshold are parameters that statistical classification methods can adjust [26].

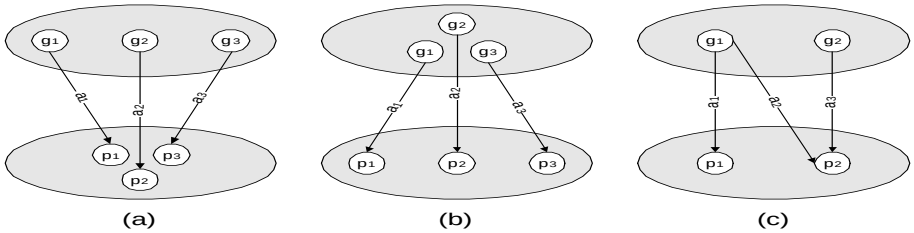

Figure 1: Annotations examples

2.1 Semantic Similarity Measures

Semantic similarity measures compute distances between terms structured in a hierarchical taxonomy. Two kinds of approaches are prevalent: information content (node based) and conceptual distance (edge based). Information content considers the similarity between two terms the amount of information they share, where a term contains less information when it occurs very often. Conceptual distance is a more intuitive approach. It identifies the shortest topologic distance between two terms in the scheme taxonomy. Budanitsky et al. experimentally compared five different proposed semantic similarity measures in WordNet [10]. The comparison shows that Jiang and Conrath's semantic similarity measure provides the best results overall [19]. This semantic similarity measure is a hybrid approach, i.e. it combines information content and conceptual distance with some parameters that control the degree of each factor's contribution.

To compute the functional distance between functional properties we propose to use Jiang and Conrath's measure. However, to measure the distance between two functional properties, we must be able to compute the following factors: their closest common ancestor; the shortest path between each term and their common ancestor; and for each term in these paths its information content, its depth and the number of its direct descendents (i.e. local density).

The information content computation depends on the terms' frequency. We have to compute the number of occurrences of each term in the corpora. However, if a term occurs then all its ancestor terms also occur. Thus, we have to propagate the term occurrences throughout the hierarchy, reaching a frequency for the root node equal to the sum of all the occurrences, as it does not represent any relevant information.

The conceptual distance is based on the node depth and density factors. The node depth factor relies on the argument that similarity increases as we descend the hierarchy, since the relations are based on increasingly finer details. The density factor relies on the argument that when the parent node has several child nodes (high density) they tend to be more similar.

3. CASE-STUDY

This section presents a case-study to evaluate if CAC method is a viable approach. The case-study uses the following biological information sources:

- CAZy (Carbohydrate Active enZYmes) is a database of carbohydrate-active enzymes identified and classified in various families by careful sequence and structural comparisons [13]. It describes the families of structurally-related catalytic and carbohydrate binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. It also links the sequences to GenBank(GenPept) [6], SwissProt [4] and PDB [7] entries. These databases are repositories of gene and protein sequence and structural data used to characterize CAZy's enzymes.

- GO (Gene Ontology) provides a structured controlled vocabulary of gene and protein biological roles [12]. The three organizing principles of GO are molecular function, biological process and cellular component. Rison et al. discuss the reasons for choosing GO as the functional scheme in a survey about functional classification schemes [23]. They describe GO as "representative of the 'next generation' of functional schemes". Unlike other schemes, GO is not a tree-like hierarchy, but a directed acyclic graph (DAG), which permits a more complete and realistic annotation.

CAZy and GO provide the structural and functional classification schemes, respectively, for our case-study. Thus, CAZy enzymes and GO terms will assume the role of gene products and biological properties, respectively, in our concept of annotation.

PubMed is an online interface for the MEDLINE database [20]. MEDLINE provides a vast collection of abstracts and bibliographic information, which have been published in biomedical journals. In this paper, we consider a document as a bibliographic item whose citation is present in MEDLINE.

3.1 Validation

To assure that CAC method produces valuable results, we need to identify a correlation between CAZy and GO classification schemes. Our strategy to identify this correlation is to compare the probability of extracting similar terms in a family with the probability of extracting similar terms in general. We structured the validation process in three steps:

1. Retrieve a set of documents related to each enzyme from available literature.
2. Extract annotations that associates each enzyme with GO terms extracted from its related documents.
3. Compute the probability of similar terms inside a family and in general.

Instead of extracting information from the entire available corpora, we retrieve only documents somehow related to each enzyme. CAZy links its enzymes to external databases (GenBank, SwissProt and PDB) that contain bibliographic references. We retrieve for each enzyme the documents cited in its linked external database entries.

We extract the annotations based on the occurrences in text [18, 15, 25]. We assume that if a document mentions a GO term then there is an underlying biological relation between the enzymes related to the document and the GO term, i.e. we annotated the enzymes with the GO term. This is a very strong assumption and a source of misannotations. However, it satisfies our goal of evaluating CAC by using it to filter misannotations.

The three organizing principles of GO represent three orthogonal ontologies, thus we did not mix annotations from different organizing principles. We choose to start by extracting only molecular functional terms, given its greater importance to CAZy.

We consider that two GO terms are similar if their functional distance is smaller than a given threshold. We implemented Jiang and Conrath's semantic similarity measure in GO to compute the functional distance. Given a specific term, we can define its probability of similar terms in a set of terms as the number of its similar terms over the total number of terms.

Definition 4. Given a term t, a set of terms T, and a similarity threshold k, we define the term's *probability of similar terms* in the set as:

$$\mathcal{P}_{sim}(t,T) = \frac{\#\{t_x : t_x \in T \ \wedge \ 0 \leq \Delta(t,t_x) \leq k\}}{\#T}$$

We assign to each family the set of terms annotated with its enzymes.

Definition 5. Considering the set of all extracted annotations $\mathcal{A}$, we define *the set of all extracted terms* as $\mathcal{T} = \{t : (e,t) \in \mathcal{A}\}$, and given a family f we define its *set of terms* as $\mathcal{T}_f = \{t : (e,t) \in \mathcal{A} \wedge e \in f\}$.

We define the probability of extracting similar terms in a particular family as the average of its term's probability of similar terms in the family.

Definition 6. Given a family f, we define the *family's probability of extracting similar terms in it* as $\mathcal{P}_{in}(f) = \overline{\{\mathcal{P}_{sim}(t, \mathcal{T}_f \setminus \{t\}) : t \in \mathcal{T}_f\}}$.

We define the probability of extracting similar terms in a family as the average of $\mathcal{P}_{in}(f)$ for all the families.

Definition 7. Given a set of families F, we define the *probability of extracting similar terms in a family* as $\mathcal{P}_{in} = \overline{\{\mathcal{P}_{in}(f) : f \in F\}}$.

To provide a good source of comparison, we define the probability of extracting similar terms in general analogously to $\mathcal{P}_{in}$. The only difference is that for each family's term we identify its similar terms in all the extracted annotations.

Definition 8. Given a family f, we define the *family's probability of extracting similar terms in general* as $\mathcal{P}_{all}(f) = \overline{\{\mathcal{P}_{sim}(t, \mathcal{T} \setminus \{t\}) : t \in \mathcal{T}_f\}}$.

Definition 9. Given a set of families F, we define the *probability of extracting similar terms in general* as $\mathcal{P}_{all} = \overline{\{\mathcal{P}_{all}(f) : f \in F\}}$.

If $\mathcal{P}_{in}$ is significantly larger than $\mathcal{P}_{all}$ for a given similarity threshold then there is a correlation between CAZy and GO classification schemes, which is a strong argument to conclude that the annotations validated by CAC method have a larger precision than all the extracted annotations.

EXAMPLE 2. *Consider* $\mathcal{T}=\{t_1,t_2,t_3\}$ *with* $\Delta(t_i,t_j) = i+j$, $\mathcal{T}_f=\{t_1,t_2\}$ *for the family* f, *and* $k=4$. *Then we have* $\mathcal{P}_{in}(f) = \overline{\{\mathcal{P}_{sim}(t_1,\{t_2\}), \mathcal{P}_{sim}(t_2,\{t_1\})\}} = \overline{\{1,1\}}=1$, $\mathcal{P}_{all}(f)=\overline{\{\mathcal{P}_{sim}(t_1,\{t_2,t_3\}), \mathcal{P}_{sim}(t_2,\{t_1,t_3\})\}} = \overline{\{1,1/2\}}=3/4$, *and therefore* $\mathcal{P}_{in}(f) > \mathcal{P}_{all}(f)$ *because* $\Delta(t_2,t_3) > k$.

	bibliographic references	distinct documents
GenBank	22849	4575
SwissProt	8998	4006
PDB	3561	785
Total		6377

Table 1: Number of items retrieved

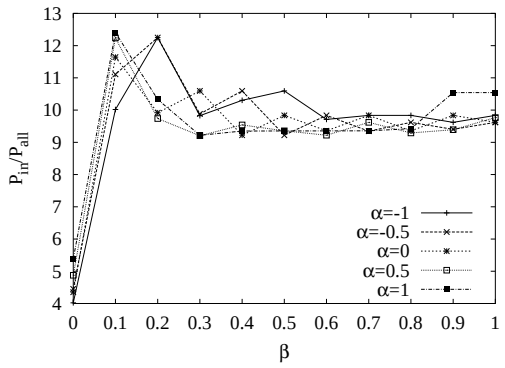

Figure 2: $\mathcal{P}_{in}$ over $\mathcal{P}_{all}$

3.2 Results

This section describes the results of our last analysis performed on the January 2003 release of GO and CAZy databases. Table 1 presents the number of bibliographic references retrieved and the number of documents cited by them. From these documents, we extracted 13869 annotations. We computed the probability of extracting similar terms for 90 families of glycoside hydrolases (GHs), which are the best curated enzymes in CAZy. These families were associated with 3748 documents, from which were extracted 343 distinct GO terms.

Figure 2 shows the ratio of $\mathcal{P}_{in}/\mathcal{P}_{all}$. We computed these values for the similarity threshold that maximized the difference between $\mathcal{P}_{in}$ and $\mathcal{P}_{all}$. The parameters α and β control the degree of how much the node depth and density factors contribute to semantic similarity computation. These contributions become less significant when α approaches 0 and β approaches 1. The values achieved show that the probability of extracting similar terms is significantly larger inside a family, as anticipated. The graph has a peak when $\beta \in [0.1, 0.2]$, where $\mathcal{P}_{in}/\mathcal{P}_{all}$ is larger than 12 for all α, except for α=0. This means that the density of the DAG and the depth of each node are important conceptual distance factors to amplify the correlation.

The maximum value of $\mathcal{P}_{in}/\mathcal{P}_{all}$ is obtained with α=1.0 and β=0.1. Figure 3 uses this configuration to show $\mathcal{P}_{in}$ against $\mathcal{P}_{all}$ for different similarity thresholds. As expected, both probabilities are proportional to the similarity threshold, since a larger similarity threshold implies also a larger number of similar terms. The relevant fact in the graphic is that $\mathcal{P}_{in}$ is always significantly larger than $\mathcal{P}_{all}$, which shows that enzymes with similar modular structure tend to be annotated with similar functional terms.

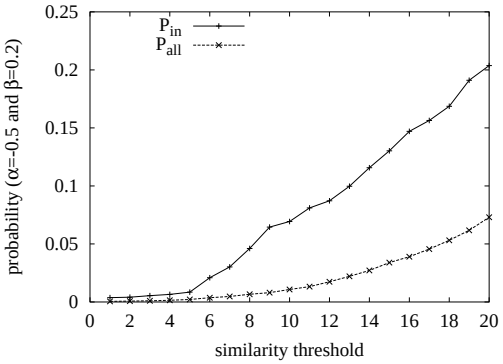

Figure 3: $\mathcal{P}_{in}$ against $\mathcal{P}_{all}$

4. RELATED WORK

Different techniques for computing similarity measures between terms have been developed to address a variety of problems. Early approaches were based only on counting edge distances between terms [22]. These were later improved by using the information content of each term, a classic Information Retrieval technique [24].

More recently, Lord et al. investigated an information content semantic similarity measure, and its application to annotations found in SwissProt [21] that also associate gene products with GO terms. They present results showing that semantic similarity is correlated with sequence similarity, i.e. function is correlated with structure. Since we propose an effective information extraction tool for biological literature, we evaluated the measure of this correlation with annotations automatically extracted from free text, instead of using human curated annotations. In our work, we replaced the sequence similarity by a modular structure classification, which is a more precise structural classification. We also tested the application of a hybrid semantic similarity measure, which integrates the information content with other valuable factors.

5. CONCLUSIONS & FUTURE WORK

We presented CAC method, which improves the efficiency of information extraction systems applied to biological lit-

erature. The method uses the correlation between structure and function to increase the precision of automatically extracted annotations.

We automatically annotated carbohydrate-active enzymes with functional terms extracted from literature. From the annotations, we computed the probability of extracting similar terms, which was significantly larger for enzymes from a common family. This result shows a correlation between modular structure and molecular function, which assures that CAC method increases the precision of extracted annotations, thus making it an effective tool for automatic information extraction of biological literature.

We implemented a hybrid semantic similarity measure to compute the similarity between GO terms, which shows that this kind of measures is feasible in a biological setting. Moreover, our results show that the information content measure improves its effectiveness when integrated with a conceptual distance measure.

To present results where the annotations validated by CAC have a larger precision than all the extracted annotations we first need to curate the extracted annotations. Besides human curation, we also intend to incorporate human curated annotations recorded in different biological sources to accelerate this procedure.

6. REFERENCES

[1] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, pages 403–410, 1990.

[2] M. Andrade and P. Bork. Automated extraction of information in molecular biology. *FEBS Letters*, 476:12–17, 2000.

[3] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Longman Higher Education, 1999.

[4] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28:45–48, 2000.

[5] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach, Second Edition*. MIT Press, 2001.

[6] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, B. Rapp, and D. Wheeler. GenBank. *Nucleic Acids Research*, 30:17–20, 2002.

[7] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[8] C. Blaschke, L. Hirschman, and A. Valencia. Information extraction in molecular biology. *Briefings in BioInformatics*, 3:1–12, 2002.

[9] C. Blaschke, R. Hoffmann, J. Oliveros, and A. Valencia. Extracting information automatically from biological literature. *Comparative and Functional Genomics*, 2:310–313, 2001.

[10] A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000), Pittsburgh, PA, June 2001.

[11] C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18:65–79, 1997.

[12] T. G. O. Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11:1425–1433, 2001.

[13] P. Coutinho and B. Henrissat. Carbohydrate-active enzymes: an integrated database approach. *Recent Advances in Carbohydrate Bioengineering*, pages 3–12, 1999.

[14] F. Couto, M. Silva, and P. Coutinho. Curating extracted information through the correlation between structure and function. In *third meeting of the special interest group on Text Data Mining*. Intelligent Systems for Molecular Biology (ISMB), 2003.

[15] J. D. et al. Mining MEDLINE: Abstracts, sentences, or phrases? In *PSB*, pages 326–337, 2002.

[16] M. Gerstein. Integrative database analysis in structural genomics. *Nature Structural Biology*, Structural genomics supplement:960–963, November 2000.

[17] L. Hirschman, J. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[18] T. Jenssen, A. L. greid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28, may 2001.

[19] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research on Computational Linguistics (ROCLING X)*, Taiwan, 1997.

[20] MEDLINE. PubMed database at the National Library of Medicine. http://www.ncbi.nlm.nih.gov.

[21] P.W.Lord, R. Stevens, A. Brass, and C.A.Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, pages 601–612, 2003.

[22] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems*, 19(1):17–30, 1989.

[23] S. Rison, T. Hodgman, and J. Thornton. Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, 1:56–69, 2000.

[24] G. Salton and M. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[25] B. Stapley and G. Benoit. Biobiblimetrics: Information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. In *PSB*, pages 326–337, 2002.

[26] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.