Semantic similarity measures as tools for exploring the Gene Ontology

P.W.Lord[a], R.D. Stevens, A. Brass and C.A.Goble
Department of Computer Science
University of Manchester
Oxford Road
Manchester
M13 9PL
UK
p.lord@russet.org.uk
robert.stevens@cs.man.ac.uk
abrass@man.ac.uk
carole@cs.man.ac.uk

Abstract

Many bioinformatics resources hold data in the form of sequences. Often this sequence data is associated with a large amount of annotation. In many cases this data has been hard to model, and has been represented as scientific natural language, which is not readily computationally amenable. The development of the Gene Ontology provides us with a more accessible representation of some of this data. However it is not clear how this data can best be searched, or queried. Recently we have adapted information content based measures for use with the Gene Ontology (GO). In this paper we present detailed investigation of the properties of these measures, and examine various properties of GO, which may have implications for its future design.

1 Introduction

Historically bioinformatics has largely grown out of efforts to deal with the increasingly large amount of data produced by molecular biology in the form of protein or DNA sequences. This sort of data can be modelled straightforwardly as a list of characters, and then searched, stored, and manipulated computationally.

During the development of the many repositories that store this data, a large amount of "annotation" has been associated with these sequences. This ranges from semi-structured data, such as species information, to unstructured free text descriptions. Often there is a large amount of annotation. Although, for example, SWISS-PROT is often described as a protein sequence database, it could also be considered to be a protein annotation database.

[a] To whom correspondence should be addressed

This has served the community well in the past, when the annotation was meant for humans to read. However it causes difficulties when trying to analyse the annotation computationally for the purpose, for example, of summarising many different SWISS-PROT entries comprising a protein family. [1] While the text is accessible by computer applications, it is not easy to interpret computationally.

It is partly because of these difficulties that there has been growing interest in ontologies within bioinformatics.[2] They provide a mechanism for capturing a community's view of a domain in a shareable form, that is accessible by humans and also computationally amenable. An ontology provides a set of vocabulary terms that label domain concepts. These terms should have definitions and be placed within a structure of relationships, the most important being the "is-a" relationship between *parent* and *child* and the "part-of" relationship between *part* and *whole.* [3,4]

The Gene Ontology (GO) is one of the most important ontologies within the bioinformatics community. [5] It is specifically intended for the purpose of extending free text annotation commonly found, with ontological annotation. As the name suggests it is limited to annotation of gene products. It comprises three orthogonal taxonomies or "aspects", that hold terms describing the *molecular function*, *biological process*, and *cellular component* for a gene product. GO is rapidly growing having over 11 000 terms (as of April 2002). Additionally new ontologies covering other regions of biology are being developed.[b]

GO represents terms within a Directed Acyclic Graph (DAG) consisting of a number of terms, represented as nodes within the graph, connected by relationships, represented as edges. Terms can have multiple parents, as well as multiple children along the "is-a' relationships ("photoreceptor" and "transmembrane receptor" are children of "receptor"), together with part-of relations that describe, for instance, that "mitochondrial membrane" is part of "mitochondrion".

The terms held within this structure are used to annotate database entries.[c]. For example, the SWISS-PROT protein, OPSR_HUMAN, has the molecular function annotation of "red-sensitive opsin", (GO:0015061). By providing a standard vocabulary across many biological resources such as SWISS-PROT and InterPro, this shared understanding should enable querying across these databases. One obvious way to query these databases would be to ask for proteins which are *semantically similar* to a query protein.

In a previous paper[6] we adapted an existing measure for semantic similar-

[b] http://www.geneontology.org/doc/gobo.html

[c] http://www.geneontology.org/goa

ity for use with GO. This measure was based on the *information content*, which uses the notion that the less frequently used terms are more informative. We tested this measure by analysing semantic similarity, and correlating it with sequence similarity, showing that, as would be expected, the more closely similar two sequences are, the more similar their ontological annotation is. We also demonstrated how this measure could be used as the basis for a simple search tool, operating over the ontological annotation.

In this paper we extend our analysis to two different methods for measuring semantic similarity, again validating the measures against sequence similarity. We also use these measures to investigate the annotation from different aspects of GO. We discuss the implications that these results have for future development of a search tool, and speculate on the implications this may have for future development of the Gene Ontology.

2 Semantic Similarity Measures

All of the measurements used here are based on the information content of each term. This is defined as the number of times each term, or any child term, occurs in the corpus. This is expressed as a probability. Although there are several available corpora we have limited our analysis to SWISS-PROT-Human (see Section 3). It is possible to interpret "child term" in a number of ways, by considering links of all semantic type, or just is-a inks. In this paper all links are used, as the distribution of link types across the different aspects, differ widely (molecular function, 6207 is-a's to 35 part-of's, cellular component, 542 to 619, biological process, 5697 to 989).

In Figure 1 these probabilities are shown diagrammatically for a small section of GO. From the definition, we can guarantee that the information content of each node increases monotonically toward the root node, which will have an information content of 1. As the three aspects of GO are disconnected subgraphs, this also holds true if we ignore the top level node ("Gene Ontology", (GO:0003683)), and take, for example, "molecular function", (GO:0003674) as our root node instead.

Given these probabilities, there are several measures of semantic similarity.[7,8,9] All three of these measures use the information content of the shared parents of the two terms, as defined in Equation (1), where $S(c1, c2)$ is the set of parental concepts shared by both $c1$ and $c2$. As GO allows multiple parents for each concept, two terms can share parents by multiple paths. We take the minimum $p(c)$, where there is more than one shared parent. We call this p_{ms} for *probability of the mimimum subsumer.*

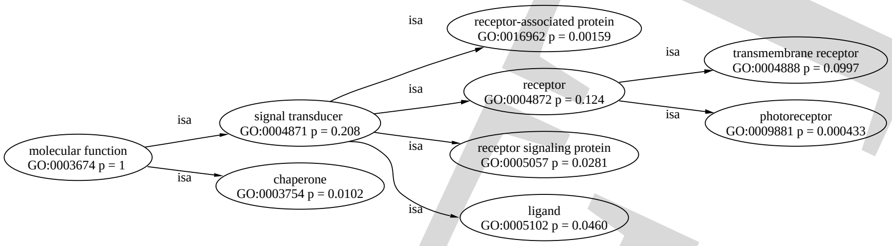


Figure 1: Probabilities in the Gene Ontology. Each node is annotated with its GO accession and the probability of this term occurring in the SWISS-PROT-Human database. This figure was produced from GO, using the graphviz tools (`http://www.graphviz.org`).

$$p_{ms}(c1, c2) = \min_{c \in S(c1,c2)} \{p(c)\} \tag{1}$$

The first of the three measures shown in Equation (2), is after Resnik, [7] and uses only the information content of the shared parents. As p_{ms} can, in general, vary between 0 and 1, this measure varies between infinity (for very similar concepts) to 0. In practise, for terms actually present in the corpus, the maximum value of this measure is defined by $-\ln(1/t) = \ln(t)$ where t is the number of occurrences of any term in the corpus.

$$\mathrm{sim}(c1, c2) = -\ln p_{ms}(c1, c2) \tag{2}$$

The next measure, after Lin, [9] uses both the information content of the shared parents, and that of the query terms. In this case, as $p_{ms} \geq p(c1)$ and $p_{ms} \geq p(c2)$, this value varies between 1 (for similar concepts) and 0.

$$\mathrm{sim}(c1, c2) = \frac{2 \times [\ln p_{ms}(c1, c2)]}{\ln p(c1) + \ln p(c2)} \tag{3}$$

The final measure, after Jiang, [8] shown in Equation (4), is of semantic distance, which is the inverse of similarity. It uses all the same terms as Equation (3), but not in the same order. [8] As with Equation (2) this can give arbitrarily large values although in practice has a maximum value of $2\ln(t)$.

$$\mathrm{dist}(c1, c2) = -2\ln p_{ms}(c1, c2) - (\ln p(c1) + \ln p(c2)) \tag{4}$$

For our purposes we are most interested in the semantic similarity between proteins, rather than GO terms *per se*, so we needed to combine these measures when a protein was annotated with several terms. In previous work, based on

WordNet,[10] a similar problem was found, as individual words have more than a single sense.[11] In this case the maximum similarity between the word senses was taken, as generally only a single word sense is used at a time. With GO annotated gene products this is not the case. A gene product will generally have all of the roles attributed to it by the annotators at the same time. We have therefore taken the average similarity between all terms. For this paper only those terms with evidence codes of “Traceable Author Statement” [d] have been used. With this dataset, most proteins have been annotated with a single term from each aspect, so values are rarely combined in this way.

3 Implementation

All results shown are from analysis performed on the April 2002 release of GO database available from `http://www.godatabase.org/dev`. The perl API available from the same source was used as an interface to this database, running over a MySQL RDBMS. The work was limited to those associations between GO terms, and SWISS-PROT proteins. In this paper SWISS-PROT-Human refers to those proteins in SWISS-PROT for which GO annotations were available, which, at the time of writing was limited to approximately 7 000 human proteins. Only those associations with “Traceable Author Statement” tags were used. The semantic similarity measures were implemented using a perl library developed for this work. All software is available on request.

BLAST searches were performed using local copies of the NCBI BLAST program over the complete SWISS-PROT protein database, available from the NCBI FTP site (`http://www.ncbi.nlm.nih.gov/BLAST/`). An “expect” value of 100 was used for all searches. Self matches, usually the best match for any protein, were excluded from the analysis. Searches were launched using the “bioperl” API (`http://www.bioperl.org`).

Results shown in Figure 2(a) and similar figures was analysed from the raw data, by taking “slices” down the X axis (ln[bit score]), and calculating the average values at each point. Scripts were written in perl, and results displayed using gnuplot (`http://www.gnuplot.info`).

Correlation is calculated as shown in Equation (5), where x_i and y_i are the semantic similarity between two proteins, over different aspects of GO, for all possible pairs of proteins in the SWISS-PROT-Human dataset.

$$\mathrm{corr}(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{5}$$

[d] see `http://www.geneontology.org/doc/GO.Evidence.html`

4 Similarity Measures Over Different Aspects

In previous work,[6] we investigated the semantic similarity measure described in Equation (2). In order to validate that this measure was producing appropriate results we compared them to sequence similarity. Results of this comparison are shown in Figure 2(a). Combined with a correlation coefficient measure it appears that sequence similarity is strongly correlated with semantic similarity based on the "molecular function" aspect of GO. This fits with the biological expectations. The sequence of a protein determines its molecular function, but does not necessarily relate to the biological process that it is involved in, or its cellular localisation.

We therefore extended this analysis to the other sequence similarity measures given in Section 2. The results of this analysis are shown in Figure 2, and Table 1.

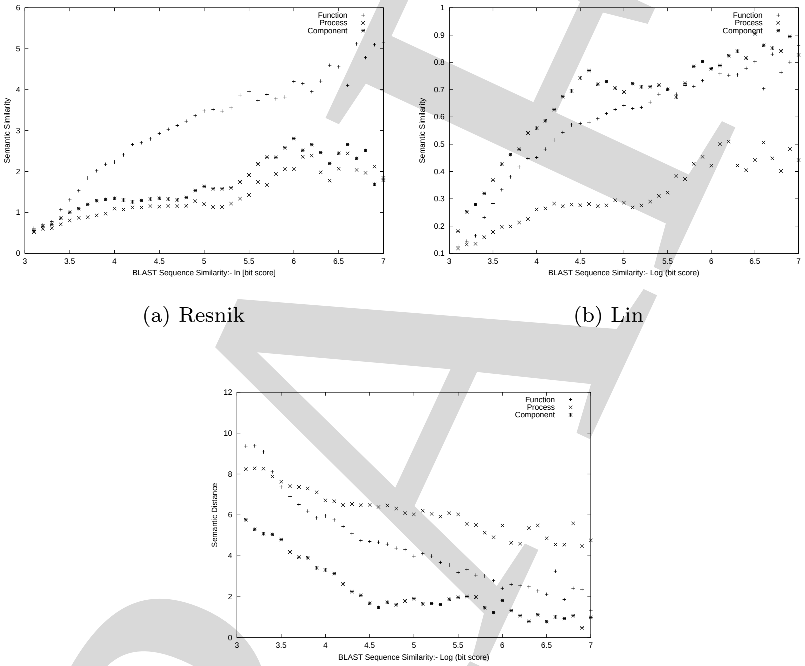
Aspect	Resnik	Lin	Jiang
Molecular Function	0.577	0.541	-0.483
Biological Process	0.280	0.303	-0.312
Cellular Component	0.368	0.452	-0.414

Table 1: Correlation co-efficients between BLAST bit scores, and semantic similarity. Data was calculated as for Figure 2. Correlation co-efficients were calculated for each data set.

For all three measures the correlation coefficients show that sequence similarity is most tightly correlated (or in the case of the distance measure inversely correlated) with the Molecular Function aspect of GO, followed by the Cellular Component aspect, and finally the Biological Process aspect. Of the three, the measure after Resnik, shows the strongest correlation with sequence similarity. Interestingly this measure also provides the weakest correlation against the biological process aspect. This suggests that the Resnik measure may be the most discriminatory. However there is no *a priori* reason to suspect that relationships will be linear, which may affect the correlation co-efficients. By inspection of Figure 2, it appears that the Resnik measure is the more linear.

5 Correlating Aspects

As discussed previously, GO is split into three different aspects. In terms of the GO DAG, although these three aspects are collected under a single top level term, "Gene Ontology", (GO:0003673), they are entirely orthogonal, being disconnected subgraphs. This part of the design of GO is justified, because these aspects "are all attributes of genes[...]. Each of these may be assigned



(a) Resnik

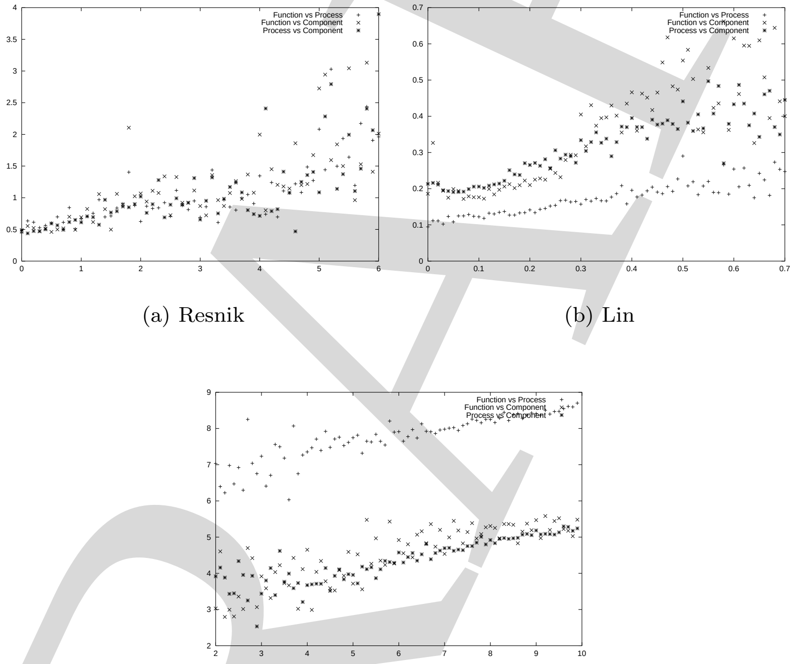
(b) Lin

(c) Jiang

Figure 2: Comparing sequence and semantic similarity. BLAST searches were performed for each SWISS-PROT-Human protein, and all matches analysed for semantic similarity with the search protein. Intervals were taken along the x-axis, $ln[bitscore]$, and values averaged, see Section 3 for details.

independently".[5] Furthermore "simply recognizing that [the aspects] represent independent attributes is by itself clarifying". Although in terms of the GO DAG these aspects are independant, we were interested in whether this was also true of the usage of the terms within SWISS-PROT.

To test this, we performed pairwise comparisons of all proteins in SWISS-PROT-Human. For each pair a semantic similarity score was calculated using each of the three aspects. The individual pairs of scores were then extracted, and compared. Results are shown in Figure 3 and Table 1.



(a) Resnik

(b) Lin

(c) Jiang

Figure 3: Comparing semantic similarity over aspects of GO. Pairwise comparisons of semantic similarity over all three aspects of GO, and for all proteins in SWISS-PROT-Human were performed. Results were split into pairs, and averaged as in Figure 2.

It is clear that for all the measures there is a significant, but weak correlation between all three of aspects. For the two similarity measurements the ordering of the correlation is conserved (molecular function to cellular com-

Aspects	Resnik	Lin	Jiang
Molecular Function - Cellular Component	0.290	0.318	0.0877
Molecular Function - Biological Process	0.219	0.244	0.269
Biological Process - Cellular Component	0.202	0.175	0.166

Table 2: Correlation coefficients for semantic similarity scores over different aspects of GO. Data was collected as for Figure 3, and correlation coefficients calculated for each data set.

ponent, molecular function to biological process, biological process to cellular component). This order is different for the distance measure after Jiang (Table 2). We are unclear whether the unexpectedly low correlation observed with this measure for the molecular function versus cellular component aspects is either because the Jiang measure is one of distance rather than similarity, or because of its behaviour over small ontologies (the cellular component ontology is about 1/5 the size of the other two aspects).

It therefore appears that while the use of the three aspects of GO is not, in fact, independent, the correlation between their use is quite weak.

6 Discussion

One of the obvious uses for these semantic similarity measures is in the development of a "semantic search" tool. A previous study compared these measures, in a different context, and found that the measure after Jiang, gave the best results. [12]

The results presented in Section 4, suggest that all three of the measures show a strong correlation between sequence similarity and molecular function semantic similiarity. The Resnik measure shows the highest correlation, as well as having the lowest correlation for the other two aspects, so it may be the most discriminatory. Further, it provides us with more information. Results are bounded between 0 and $\ln(t)$, where t is the number of terms in the corpus, while the Lin measure is bounded between 0 and 1. A large numerical value therefore indicates a large corpus. The numerical value also reveals information about the usage within corpus of the part of the ontology queried. The score from comparing a term with itself depends on where in the ontology the term is, with less frequently occurring terms having higher scores.

The Lin measure hides this information, as term compared to itself will always score 1. However, it has a significant advantage. One difficulty with using these measures in a search is that many protein pairs share identical scores (see Table 3) which hinders ranking. As the Resnik measure depends solely on the information content of the shared parents, there are only as many discrete scores as there are ontology terms. By using the information content of the query terms the Lin measure increases the number of discrete scores at least quadratically with the ontology size. The example search shown in Table 3 demonstrates this point. The Resnik measure ranking places a protein annotated as “Androgen Receptor”, (GO:0004882), and “RNA polymerase II Transcription Factor”, (GO:0003702) equally, because the former term is a child of the latter. The Lin measure, however, ranks all proteins annotated with “RNA polymerase II Transcription Factor”, (GO:0003702) first, as it is capable of differentiating between a term and one of its children.

Swissprot ID	Description	Similarity	GO Term
Resnick			
ANDR_HUMAN	Androgen receptor (Dihydrotestosterone receptor)	3.412	“Androgen receptor”, (GO:0004882)
AP1_HUMAN	Transcription factor AP-1	3.412	“RNA polymerase II transcription factor”, (GO:0003702)
ATF4_HUMAN	Cyclic-AMP-dependent transcription factor	3.412	“RNA polymerase II transcription factor”, (GO:0003702)
Lin			
AP1_HUMAN	Transcription factor AP-1	1	“RNA polymerase II transcription factor”, (GO:0003702)
ATF4_HUMAN	Cyclic-AMP-dependent transcription factor	1	“RNA polymerase II transcription factor”, (GO:0003702)
BTF3_HUMAN	Transcription factor	1	“RNA polymerase II transcription factor”, (GO:0003702)
Jiang			
ENL_HUMAN	ENL protein.	0.634	“RNA polymerase II transcription factor”, (GO:0003702)
AF4_HUMAN	AF-4 protein (Proto-oncogene AF4)	0.934	“Transcription factor”, (GO:0003700)
AIRE_HUMAN	Autoimmune regulator (APECED protein)	0.934	“Transcription Factor”, (GO:0003700)

Table 3: Tables shows results from a search against SWISS-PROT-Human, with the HXA1_HUMAN protein, against the molecular function aspect of GO, using the three different similarity measures for ranking. The top three results are shown for each measure. Following ranking by semantic similarity, proteins were sorted alphabetically.

The Jiang measure, which is the only distance measure, has the weakest correlation between the molecular function similarity and sequence similarity. However, as with the Lin measure, it combines information content from the shared parent, and the query terms.

Further investigation is required to determine which of the three measures is most appropriate for use within a search tool. It also seems likely that the relative advantages and disadvantages will change as GO increases in size and usage. The information hidden by the Lin measure will be less relevant if we know that both ontology and corpus are large, while the advantages of ranking for the Lin measure will increase with the size of the ontology. As well as further theoretical studies, we are developing a web delivered tool which should allow practical experimentation and user feedback.

We have also shown that the results from different aspects are only weakly correlated. *A priori* it is unclear whether users would prefer to perform semantic similarity searches over the GO as a whole, or over the different aspects independently. The data presented here suggests that as the aspects are largely independent combining results from the different aspects would be of little value, unless the user is looking for identically annotated proteins.

Although the aspects are largely independent, there is a correlation between them. This may have implications for the design of GO. Currently the aspects are completely disconnected subgraphs, which reflects the notion that these attributes are independent, when, in reality, they are not. Yet there is no formal linkage between, for example, the concept of “taste”, (GO:0007607), which is a biological process term, and the concept of a “taste receptor”, (GO:0008527) which is a molecular function term. As new ontologies emerge, covering further areas of biology, this problem may become more acute. Moving GO to a more expressive description logic based representation may be a good way to achieve this.[13]

In summary we have investigated several different measures of semantic similarity, all using information content as their basis. None of the three measures stand out as having a clear advantage over the others, although each has strengths and weaknesses. We have also investigated the behaviour of the different aspects of GO, and shown that they are largely independent, so that it will clearly be profitable to provide searches over different aspects of GO, rather than combining results. We believe that this work paves the way toward the development of a semantic similarity search tool, which will be a valuable additional tool in the armoury of the researcher.

Acknowledgements

This work was funded under the ESPRC/BBSRC Bioinformatics Programme (Grant number: BIF/10507)

1. P.W.Lord, J.R.Reich, A.Mitchell, R.D.Stevens, T.K.Attwood, and C.A.Goble. PRECIS: An Automated Pipeline for Producing Concise Reports About Proteins. In *IEEE International Symposium on Bioinformatics and Biomedical engineering*, pages 59–64. IEEE press, 2001.

2. R. Stevens, C.A. Goble, and S. Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, 1(4):398–416, 2000.
3. M. Winston, R. Chaffin, and D. Herrmann. A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11:417–444, 1987.
4. J.J. Odell. *Six Different Kinds of Aggregation*, pages 139–149. Cambridge University Press, 1998.
5. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33, August 2001.
6. P. W. Lord, R.D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 2002. Submitted.
7. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453, 1995.
8. J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan, 1998. ROCLING X.
9. D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.
10. C. Fellbaum, editor. *WordNet:- An electronic lexical database.* MIT Press, Cambridge, Massachusetts, 1998.
11. P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
12. A. Budanitsky and G.Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, 2001.
13. C. J. Wroe, R. D. Stevens, C. A. Goble, and M. Ashburner. An evolutionary methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL. In *Pacific Symposium of Biocomputing.*, 2003.