Using Enriched Ontology Structure for Improving Statistical Models of Gene Annotation Sets

Frank Rügheimer

Institut Pasteur, Laboratoire Biologie Systémique
Département Génomes et Génétique
F-75015 Paris, France
CNRS, URA2171, F-75015 Paris, France
`frueghei@pasteur.fr`

Abstract. The statistical analysis of annotations provided for genes and gene products supports biologists in their interpretation of data from large-scale experiments. Comparing, for instance, distributions of annotations associated with differentially expressed genes to a reference, highlights interesting observations and permits to formulate hypotheses about changes to the activity pathways and their interaction under the chosen experimental conditions. The ability to reliably and efficiently detect relevant changes depends on properties of the chosen distribution models. This paper compares four methods to represent statistical information about gene annotations and compares their performance on a public dataset with respect to a number of evaluation measures. The evaluation results demonstrate that the inclusion of structure information from the Gene Ontology enhances overall approximation quality by providing suitable decompositions of probability distributions.

1 Introduction

The Gene Ontology (GO) [1] establishes standardized sets of annotation terms for genes and gene products. Terms are grouped in three separate sub-ontologies that are concerned with intracellular location, molecular functions and associated biological processes of gene products respectively. In addition, the ontology provides a network of relations that formalize the relationships between annotation terms. This is used, for example, to associate a general category with more specific terms subsumed under that category, so annotations on different levels of detail may be applied concurrently. The resulting formal description of domain knowledge has been a key contribution to expanding the use of computational methods in the analysis and interpretation of large scale biological data. For instance, the GO enables the definition of semantic similarity measures, which in turn can be used to compare or cluster gene products and groups thereof [7] or to implement semantic search strategies for retrieval of information from domain-specific text collections [5].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part II, CCIS 81, pp. 55–64, 2010.

The utility of the term relations is further increased when the ontology is combined with an annotation database. In the case of the GO term relations have been combined with databases of annotations for gene products to identify statistically significant term enrichment [2] within subset of genes undergoing expression changes in large scale experiments (microarrays, ChIP-chip etc.). For this reason plug-ins for GO annotations have been integrated into standard data visualization and analysis software for systems biology [3]. In a similar way annotation sources and term relations can be combined with data from experiments investigating effects of interventions, e.g. from knockout or RNAi studies. The relational information provided by GO contributes to the integration of observations on different levels of detail so a subsequent statistical analysis of the results becomes possible. Beyond this role in data fusion, the ontology structure can guide the construction of statistical models by providing decompositions of probability distributions over annotations. As an additional benefit this approach establishes consistency between distributions regardless of the level of detail under which data is viewed for the purpose of the analysis.

In annotation databases following the GO standard each entry assigns an annotation term to a particular gene. Both genes and annotation terms may occur in several entries. Therefore several terms may be annotated to the same gene, and such combinations of annotations terms are often used to indicate roles in the interaction of pathways associated with different biological functions. While an analysis of the annotation sets as a whole is desirable, a direct representation via empirical distributions is usually impractical as the theoretical size of the sample space for annotations with n possible terms is on the order of 2^n. To construct probabilistic models of annotation frequencies a number of representations are employed, which differ in their inherent modeling assumptions and the simplifications applied. In this paper I compare such strategies using publicly available data sets and discuss their differences in the light of the resulting properties.

Section 2 provides a brief exposition of the data set used in the comparison, its connection to the Gene Ontology and the preprocessing applied to it. In section 3 the three different types of distribution models are presented and pointers to the relevant literature given. This is followed by details of the evaluation method and the evaluation measures employed (section 4). Results are summarized and discussed in section 5.

2 Data Sets and Preprocessing

The data set used throughout the experiment was constructed from a collection of annotations on the function of the genes and gene products of the baker's and brewer's yeast *Saccharomyces cerevisiae* – one of the most well-studied eukariotic model organisms. The collection is maintained by the Saccharomyces Genome Database project [11] and will be referred to as the SGD. The annotations provided by that source are compliant with the GO annotation standard.

Within the GO, terms are organized into three non-overlapping term sets. Each of the three sets covers one annotation aspect, namely biological processes,

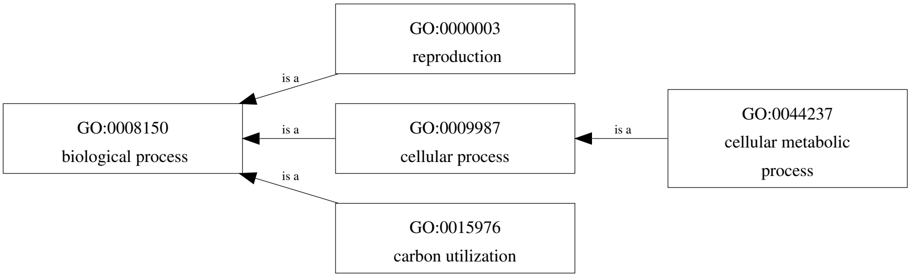

Fig. 1. Extract from the quasi-hierarchical term structure as specified by the Gene Ontology relations

molecular functions or cellular components. For the "cellular components" the term set is structured by a quasi-hierarchical partial ordering defined via the "`part_of`" relation whereas the "`is_a`" relation fulfills the same role for the two other aspects (Figure 1). The "molecular function" annotation refers to biochemical, e.g. enzymatic, activity and is closely linked to the protein (and therefore gene) sequence. However, many proteins that posses very similar functions or even a shared evolutionary history are found in largely unrelated pathways. The "cellular component" annotation provides information on cell compartments in which the gene products are known to occur. On the biological side this information allows to restrict a search for potential additional players in a partially known mechanism. This enables experimenters to look specifically at, for instance, candidates for a postulated membrane bound receptor. Finally, the "biological process" annotations provide an idea of the overall functionality to which a gene product contributes. The terms occurring in the ontology fragment of Figure 1 are examples of this annotation type. Because the targeted interactions in large scale expression studies are focused on overall biological processes only annotations for this aspect were considered in the experiments.

It should also be noted that the full term set of the GO (at the time of this writing >30,000 terms) provides a very high level of detail. Only a small subset of the available annotation terms are actually used in the SGD. Even among the terms that are used, many have a low coverage in the database. To obtain a standardized, broader perspective on the data that lends itself to a statistical analysis, less specific versions of the ontology can be employed. These so-called "slim ontologies" define subsets of comparatively general Gene Ontology terms. In the case of *S. cerevisiae* a species-specific slim version of the ontology has been released together with the full annotation data [12]. For the study described in this paper any annotation terms from the SGD that were not already included in the slim version of the ontology were mapped to their most specific ancestors in the reduced term set. Note that for the selected sub-ontology the corresponding term subset of the Yeast Slim GO has tree structure.

The resulting term sets consists mainly of leaf nodes of the slim ontology, but still contains elements representing coarser descriptions. For the evaluation these

```
tY(GUA)M1 {GO_0006412}
YAL004W   {GO_0008150}
YAL005C   {GO_0006412 GO_0006457 GO_0006810 GO_0006950 GO_0006996}
YAL010C   {GO_0006810 GO_0006996 GO_0016044}
YAL014C   {GO_0006810}
YAL017W   {GO_0005975 GO_0006464 GO_0007047}
YAL018C   {GO_0008150}
YAL019W   {GO_0006996}
YAL026C   {GO_0006810 GO_0006996 GO_0016044 GO_0016192 GO_0042254}
```

Fig. 2. Fragment of constructed gene list with associated GO term identifiers

terms were considered as competing with their more specific hierarchical children, reflecting the GO annotation policy of assigning the most specific suitable term supported by the observations. For the analysis the example database was constructed by aggregating the mapped annotations for each of the known genes into gene-specific annotation sets. The resulting file summarizes the known biological processes for each of 6849 genes using a total of 909 distinct annotation sets (Figure 2).

In parallel, the preprocessing assembled information about the annotation scheme employed. To that end the term hierarchy was extracted from the ontology and converted into a domain specification. This specification serves to describe how the annotation on different levels of details relate to each other and was later used to by one of the models to integrate the ontology information during the learning phase.

3 Distribution Models

In order to cover a broad spectrum of different strategies four representations for distributions on annotation sets were implemented:

a) A model using binary random variables to encode presence or absence of elements in a set. The variables are treated as independent, so the distribution of set-instantiations is obtained as a product of the proabilities for the presence or absence of the individual elements of the underlying carrier set.
b) A condensed distribution [8] model using an unstructured attribute domain
c) An enriched term hierarchy using condensed random sets for the representation of branch distribution [10,9]
d) A random set representation [6]

The representation task is formalized as follows: Let Ω denote the set of available annotation terms. The preprocessed annotation database is rendered as a list $D = \{S_1, \ldots, S_m\}$ $m \in \mathbb{N}$, $S_i \in 2^{\Omega}$, where the S_i represent annotation term sets associated with individual genes. The representation task is to model relevant properties of the generating probability distribution characterized via its probability mass function $p_{\text{Annot}} : 2^{\Omega} \rightarrow [0, 1]$. To this end distribution models are trained from a non-empty training set $D_{\text{trn}} \subset D$ and subsequently tested using the corresponding test set $D_{\text{tst}} = D \setminus D_{\text{trn}}$. To increase the robustness of evaluation results several training and test runs are embedded into a cross-validation framework (cf. section 4).

The independence assumptions in (a) allow a compact representation of probability distributions by decomposing them into a small set of binary factor distributions $\hat{p}_{\omega_i} : \{+, -\} \rightarrow [0, 1]$, where the outcomes $+$ and $-$ denote presence or absence of the term ω_i in the annotation set. This results in the decomposition

$$\forall S \subseteq \Omega : \quad \hat{p}_{\text{Annot}}(S) = \left(\prod_{\omega \in S} \hat{p}_\omega(+) \right) \left(\prod_{\omega \in \Omega \setminus S} \hat{p}_\omega(-) \right) \tag{1}$$

$$= \left(\prod_{\omega \in S} \hat{p}_\omega(+) \right) \left(\prod_{\omega \in \Omega \setminus S} 1 - \hat{p}_\omega(+) \right) . \tag{2}$$

Because only one value per term needs to be stored this results in a very compact model. Moreover, the approach allows to rapidly compute probability estimates and is thus popular in text mining and other tasks involving large term sets. The strong independence assumptions, however, are also a potential source of errors in of the representation of probability distribution over the set domain 2^Ω.

Approach (d) represents the opposite extreme: Each possible combination of terms is represented in the sample space, which for this model is the power set 2^Ω of the term set. Therefore the the target distribution is estimated directly from observations of the samples. Due to the size of the distribution model, and the sparse coverage of the sample space no explicit representation of the model was provided. Instead all computations were conducted at evaluation time based on counts of annotation term sets shared by the training and test database applying a subsequent modification for the Laplace correction (see page 60). Nevertheless, computation time for evaluating the random set model exceeded that of the other models by several orders of magnitude and, giving its scaling behavior, is not considered an option for application in practice.

Finally approach (b) and (c) represent two variants of the condensed random set model introduced in [8] and [10] respectively. The central idea of these approaches is to use a simplified sample space that represents annotations consisting of single terms separately, but groups those for non-singleton instantiations. The probability mass assigned to the non-singleton instantiations is then further distributed according to a re-normalized conditional probability distribution, which is encoded using the method proposed in (a). This two-step approach allows to better reflect the singletons (which are overrepresented in GO-Annotations), while retaining the performance advantages of (a). Approach (c) additionally integrates structure information from the ontology by associating a condensed random set with each branch of the ontology structure. Because condensed random sets use the probabilistic aggregation operation observations on coarsened versions of the enriched ontology remain consistent with aggregated

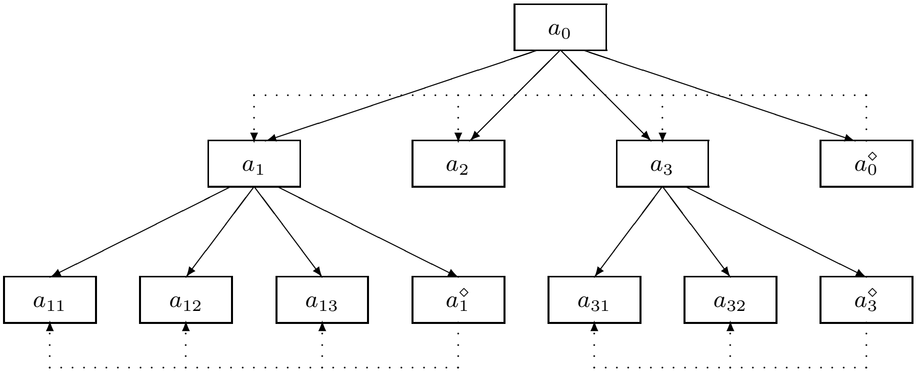

Fig. 3. Decomposition principle for the hierarchical version of the Condensed Random Set model. Conditional probabilities and coverage factors are indicated by solid and dotted arrows respectively (image from [10]).

results from observations on higher levels of detail. For an in-depth discussion of parameters and the model induction algorithm see [9].

In all cases, the parameters were estimated from the observed frequencies in the training data with a Laplace correction applied. The value of the Laplace correction was set to of 0.5 for models (a), (b), (c) and to $2.5 \cdot 10^{-9}$ for model (d), contributing similar portions of the total probability mass for all models.

4 Evaluation

Preprocessing resulted in a database of annotation sets for 6849 genes. To limit sampling effects, the evaluation measures were computed in a 5-fold cross-validation process [4]. To this end the data set was split into five partitions with genes randomly assigned (4 partition with 1370 genes each and one partition with 1369 genes). In each of the five runs a different partition was designated as a test data set whereas the remaining partitions used in the role of training data. Evaluation measures were chosen to provide complementary information on how well different aspects of the set-distribution are captured by each model type. All measures are described with respect to and evaluation against a test data set $D_{\text{tst}} \subset D$.

Log-Likelihood. A common way to evaluate the fit of a probability-based model M is to consider the likelihood of the observed test data D_{tst} under the model, that is, the conditional probability estimate $\hat{p}(D_{\text{tst}} \mid M)$. The closer the agreement between test data and model, the higher that likelihood will be. The likelihood is also useful to test model generalization, as models that overfit the training data tend to predict low likelihoods for test datasets drawn from the same background distribution as the training data. To circumvent technical limitations concerning the representation of and operations with small numbers in the computer, the actual measure used in practice is based on the logarithm of the likelihood:

$$\log L(D_{\text{tst}}) = \log \prod_{S \in D_{\text{tst}}} \hat{p}(S \mid M) \tag{3}$$

$$= \sum_{S \in D_{\text{tst}}} \log \hat{p}(S \mid M). \tag{4}$$

In that formula the particular term used to estimate the probabilities $\hat{p}(S \mid M)$ of the records in D are model-dependent. Since the likelihood takes values from $[0, 1]$ the values for the log-transformed measure are from $(-\infty, 0]$ with larger values (closer to 0) indicating better fit. The idea of the measure is that the individual cases (genes) in both the training and the test sets are considered as independently sampled instantiations of a multi-valued random variable drawn from the same distribution. The likelihood of a particular test database D_{tst} is computed as the product of the likelihoods of its $|D_{\text{tst}}|$ elements. Due to the low likelihood of individual sample realizations even for good model approximation, the Log-Likelihood is almost always implemented using the formula given in Equation 4, which yields intermediate results within the bounds of standard floating point format number representations.

One particular difficulty connected with the Log-Likelihood, resides in the treatment of previously unobserved cases in the test data set. If such values are assigned a likelihood of zero by the model then this assignment entails that the whole database is considered as impossible and the Log-Likelihood becomes undefined. In the experiment this undesired behavior was countered by applying a Laplace correction during the training phase. The Laplace correction ensures that all conceivable events that have not been covered in the training data are modeled with a small non-zero probability estimate and allow the resulting measures to discriminate between databases containing such records.

Average Record Log-Likelihood. The main idea of the log-likelihood measure is to separately evaluate the likelihood of each record in the test database with respect to the model and consider the database construction process a sequence of a finite number of independent trials. As a result log-likelihoods obtained on test databases of different sizes are difficult to compare. By correcting for the size of the test database one obtains an average record log-likelihood measure that is better suited to a comparative study:

$$\text{arLL}(D_{\text{tst}}) = \frac{\log L(D_{\text{tst}})}{|D_{\text{tst}}|}. \tag{5}$$

Note that in the untransformed domain the mean of the log-likelihoods corresponds to the geometric mean of the likelihoods, and is thus consistent with the construction of the measure from a product of evaluations of independently generated instantiations.

Singleton and Coverage Rate Errors. In addition to the overall fit between model and data, it is desirable to characterize how well other properties of a set-distribution are represented. In particular it has been pointed out that the condensed distribution emphasizes the approximation of both singleton probabilities and the values of the element coverage. To assess how well these properties are preserved by the investigated representation methods, two additional measures – $\mathrm{d}_{\mathrm{sglt}}$ and $\mathrm{d}_{\mathrm{cov}}$ – have been employed. These measures are based on the sum of squared errors for the respective values over all elements of the base domain:

$$\mathrm{d}_{\mathrm{sglt}} = \sum_{\omega \in \Omega} \left(p'(\omega) - p(\omega)\right)^2 , \tag{6}$$

$$\mathrm{d}_{\mathrm{cov}} = \sum_{\omega \in \Omega} \left(\mathrm{opc}'(\omega) - \mathrm{opc}(\omega)\right)^2 . \tag{7}$$

In this equation the function opc computes the one-point coverage of an element by a random set, defined as the cumulative probability of every instantiation containing the argument.

5 Results

For the assessment and comparison of the different methods, a 5-fold cross-validation was conducted. All approaches were applied with the same partitioning of the data. Evaluation results of the individual cross-validation runs were collected and – with the exception of the logL measure – averaged. These results are summarized in Table 1.

The two condensed random set-based models (b) and (c) achieve a better overall fit to the test data (higher value of arLL-measure) than the model assuming independence of term coverages (a) indicating that those assumptions are not well suited for annotation data. The highest accuracy for all models and variants is achieved using the hierarchical version of the CRS model. This is interpreted as a clear indication of the benefits provided due to additional structure information. Despite its large number of parameters the full random set representation (d) does not achieve acceptable approximation results. Due to the large sample space that model is prone to overfitting.

For the prediction of singleton annotations model (a) exhibits large prediction errors. This again is explained by the independence assumption in that representation being too strong. In contrast, with their separate representation of singleton annotation sets, the CRS-based models show only small prediction errors for the singleton frequencies, though the incomplete separation between real singletons and single elements in local branch distributions appear to leads to a slightly increased error for the hierarchical version.

Table 1. Evaluation results for individual runs and result summaries; best and second best results highlighted in dark and light gray respectively (from top left to bottom right: Model using independent binary variables (a) with Laplace correction of 0.5, Condensed Random Sets on unstructured domain (b) with Laplace correction of 0.5, Condensed Random Sets on hierarchically structured domain (c) with Laplace correction of 0.5, Random Set representation (d) with Laplace correction of $2.5 \cdot 10^{-9}$)

a)

$\log L$	arLL	d_{sglt}	d_{cov}
-9039.60	-6.60	0.067856	0.001324
-8957.19	-6.54	0.064273	0.001524
-9132.09	-6.67	0.060619	0.001851
-8935.82	-6.52	0.074337	0.001906
-9193.44	-6.72	0.059949	0.001321
	-6.61	0.065406	0.001585

b)

$\log L$	arLL	d_{sglt}	d_{cov}
-7992.76	-5.83	0.000241	0.001342
-7885.19	-5.76	0.000222	0.001531
-8045.31	-5.87	0.000411	0.001838
-7839.16	-5.72	0.000612	0.001895
-8195.49	-5.99	0.000268	0.001316
	-5.83	0.00035	0.001584

c)

$\log L$	arLL	d_{sglt}	d_{cov}
-7629.66	-5.57	0.000539	0.008293
-7559.38	-5.52	0.000457	0.011652
-7752.21	-5.66	0.000857	0.006998
-7529.83	-5.50	0.001014	0.004767
-7828.44	-5.72	0.000567	0.009961
	-5.59	0.000686	0.008334

d)

$\log L$	arLL	d_{sglt}	d_{cov}
-8259.66	-6.03	0.001823	0.098462
-8346.13	-6.09	0.001311	0.100860
-8651.12	-6.31	0.000964	0.095850
-8288.68	-6.05	0.003105	0.103200
-8534.30	-6.23	0.000671	0.094305
	-6.14	0.001574	0.098536

This is consistent with the higher error d_{cov} of that model in the prediction of coverage factors. The non-hierarchical models (a) and (b) represent one-point coverages directly and therefore achieve identical prediction error[1]. Large deviations for coverage rate predicted by the Random Set model (d) are explained by the cumulative effect of the Laplace correction after aggregating over the a large number of combinations.

6 Conclusions

The presented contribution analyzed the effect of different modeling assumptions for representing distributions over annotation sets. Although parsimonious models should be preferred whenever justified from the data, the often applied independence assumption for term occurrence do not seem to hold for annotation data in biology. It could be shown that the inclusion of background information on relations between annotation terms contributes to improving the overall accuracy of the representation at some cost for the accuracy of coverage rates and singleton frequencies. In combination with the additional benefit of consistent aggregation operations the results indicate that the probabilistic

[1] The minor differences between the tables are merely artifacts of the two-factor decomposition of coverage factors in the condensed distribution.

enrichment of ontologies provides an both effective approach to the statistical modeling of distributions over annotation sets and integrates well with already available resources for data analysis in biology.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
2. Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G.: GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with lists of genes. Bioinformatics 20(18), 3710–3715 (2004)
3. Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B., Aittokallio, T.: GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. Bioinformatics 23(3), 394–396 (2006)
4. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. of the 14th Int. Joint Conference on Artificial Intellligence (IJCAI 1995), pp. 1137–1145 (1995)
5. Müller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLoS Biology 2(11) (2004)
6. Nguyen, H.T.: On random sets and belief functions. Journal Math. Anal. Appl. 65, 531–542 (1978)
7. Ovaska, K., Laakso, M., Hautaniemi, S.: Fast Gene Ontology based clustering for microarray experiments. BioData Mining 1(11) (2008)
8. Rügheimer, F.: A condensed representation for distributions over set-valued attributes. In: Proc. 17th Workshop on Computational Intelligence. Universitätsverlag Karlsruhe, Karlsruhe (2007)
9. Rügheimer, F., De Luca, E.W.: Condensed random sets for efficient quantitative modelling of gene annotation data. In: Proc. of the Workshop "Knowledge Discovery, Data Mining and Machine Learning 2009" at the LWA 2009, pp. 92–99. Gesellschaft für Informatik (2009) (published online)
10. Rügheimer, F., Kruse, R.: An uncertainty representation for set-valued attributes with hierarchical domains. In: Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2008), Málaga, Spain (2008)
11. SGD Curators: Saccharomyces genome database, `http://www.yeastgenome.org`, (accessed 2008/11/16)
12. SGD Curators: SGD yeast gene annotation dataset (slim ontology version). via Saccharomyces Genome Database Project [11], `ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/go_slim_mapping.tab` (accessed November 16, 2008)