Condensed Random Sets for Efficient Quantitative Modelling of Gene Annotation Data

Frank Rügheimer
Biologie Systémique, Institut Pasteur
75015 Paris, France
frueghei@pasteur.fr

Ernesto William De Luca
Otto-von-Guericke University of Magdeburg
39106 Magdeburg
ernesto.deluca@ovgu.de

Abstract

With the widespread use of annotations in biological databases efficient models for statistical properties of set-valued attributes become increasingly relevant. In this work we introduce condensed random sets (CRS) as compact representations of distributions over annotation sets. The approach is discussed for both unorganized term vocabularies and term hierarchies, applied to an annotated yeast genome dataset and evaluated in comparison to an alternative representation. Encouraged by the results of the evaluation we explore further applications by pointing out how the representation can be used to support the construction of new semantic similarity measures for information retrieval.

1 Introduction

Genome sequencing has become a widely used tool in modern biology. Yet, in higher organisms, genomic information alone does not suffice to predict and characterize the manner in which particular gene products affect metabolic and signaling pathways, as it does not reflect the large number of interactions in the cell. To obtain a better understanding of biological processes the investigation of other layers of the cell machinery that are closer to the biological function has recently drawn much attention, e.g. the regulatory mechanisms involving RNAs [Toledo-Arana *et al.*, 2009] (transcriptome) and eventually the resulting proteins and their post-translational modifications themselves (proteome).

Due to a combination of recent advances of experimental techniques and extensive efforts to systematically survey literature, biologists have succeeded in establishing curated collections of information concerning gene products and their role for a number of model organisms. One of the results of those efforts was the realization that the same genes are frequently involved in several, sometimes seemingly unrelated biological processes. That information has been released to the public in the form of standardized public databases in which gene identifiers are associated with annotations terms describing their function. Using associated relational knowledge representations such as the Gene Ontology, annotation terms can be organized and linked with each other. In particular the Gene Ontology defines a hierarchy of annotation terms thus allowing to specify properties on different levels of detail.

In the present work we concern ourselves with the enrichment of relational knowledge representations with quantitative information extracted from annotated reference datasets. In particular we have investigated sets of annotation terms on the biological processes linked to the products of known genes. Observed statistical relationships between those annotations, and between terms and their generalizations were overlaid with term hierarchies extracted from the Gene Ontology. The resulting data representation can be used, e.g.:

1. To summarize and compare properties of datasets;
2. To compute likely expansions of coarse annotations to a higher level of detail, e.g. when making predictions from incomplete information or integrating data from different measurements;
3. To improve semantic similarity measures by taking into account empirically derived statistical relationships between annotation terms.

In the following section we establish how datasets featuring annotations with multiple terms can be modelled using condensed random sets. Following that we extend that approach to integrate it with relational knowledge representations in the form of term hierarchies. Both variants of the model are then applied to and evaluated on an annotated dataset for the bakers and brewers yeast *Saccharomyces cerevisiae*, which has been extensively studied as a eukaryotic model organism (Section 4). The results are compared to a representation based on an independent modelling of annotation terms. Finally Section 5 explores the prospect of applying the modeled term-set distributions for the construction of new context-specific semantic similarity measures.

2 Condensed Random Sets

Due to their relative flexibility and extensibility annotations have become a popular way to enrich existing data. Unlike conventional attributes that may only take one value out of a fixed domain, the same data-object may be simultaneously annotated with several terms that together describe a property. Denoting the set of potentially admissible annotation terms as Ω, annotations instantiate a set-valued at-

tribute A^* that takes values from the set 2^Ω formed by the subsets of Ω. Apart from the simple list of associated terms, that very information may yield other interesting findings. For the annotated yeast genome, for example, it allows to investigate whether the activity of a gene is specific to a biological process or not. Conversely, when analysing expression data, the interpretation of the process annotations yields lists of candidates for pathways or functions that are affected by targeted interventions. Applied to whole genomes, the focus shifts from individual sets to frequency distributions over sets. From these distributions, one can obtain a quantitative characterization, for instance, of the level of complexity (fraction of genome involved) and relative importance (fraction of specialized genes/proteins) of biological processes in the organisms of interest.

In the probabilistic framework such a distribution over the possible annotation sets gives rise to a *Random Set* [Nguyen, 1978]. The two characteristics suggested as complexity and specificity assessments respectively correspond to the one-point coverage and the single-element probabilities of the random set, with a distribution p^* specifying the probability of each individual annotation-term combination. Since the number of combinations grows exponentially with the number of admissible annotation terms, however, a direct representation strategy allows for a very limited choice of annotation terms only. Even if all values can be represented in memory, providing estimates for a large number of – in most cases very small – probabilities with acceptable precision would require unrealistically large samples [Wasserman, 2006].

Fortunately, many applications do not require representations with detailed probability values for all set-valued outcomes. Due to their role in interpretations probabilities of singletons and the probability of term coverage by set-valued annotations provide useful information summaries. By focusing on these pieces of information the condensed random sets achieve a compact representation of statistical information regarding set attributes.

The condensed random set approach [Rügheimer, 2007], builds on a partitioning of the set of the subsets of a sample space Ω and a mapping of set-distributions to a probability/possibility distribution over the condensed domains. In the formalization of that approach a special attribute value is introduced to label outcomes that are multi-valued w.r.t. a frame of discernment Ω or correspond to the empty set. For simplicity, the representation is initially discussed for the case of an unstructured repository of annotation terms.

Definition 1 *Let Ω be a set of distinct labels. Furthermore let $\omega^\diamond$ be a special symbol uniquely associated with and not already contained in Ω. Consider a mapping σ from the set of subsets 2^Ω to the* **extended set universe** $\Omega \cup \{\omega^\diamond\}$

$$\begin{aligned} \sigma :\ 2^\Omega &\rightarrow \Omega \cup \{\omega^\diamond\} \\ \forall S \subseteq \Omega :\ \sigma(S) &= \begin{cases} \omega & \text{if } S = \{\omega\}, \omega \in \Omega, \\ \omega^\diamond & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

We call σ the **set reduction mapping** *w.r.t. Ω.*

It is easily seen that σ preserves the distinction between singleton elements of 2^Ω, but collects the multi-valued outcomes in a separate class. Consider now a set of objects or cases O and their description via a set-valued attribute A^* taking values from 2^Ω. Using the definition of the set reduction mapping, it is possible to define a condensed set-valued attribute $A^\diamond$ that is linked to the values of A^*:

Definition 2 *Let A^* be a set-valued attribute $A^* : O \rightarrow 2^\Omega$. Additionally let $\sigma : 2^\Omega \rightarrow \Omega \cup \{\omega^\diamond\}$ denote the set reduction mapping w.r.t. Ω. The* **condensed set-valued attribute** $A^\diamond$ **induced by** A^* *is a mapping:*

$$\begin{aligned} A^\diamond :\ O &\rightarrow \Omega \cup \{\omega^\diamond\} \\ \forall o \in O :\ o &\mapsto \sigma(A^*(o)). \end{aligned} \tag{2}$$

The relation between the attribute domain conveyed by the set reduction mapping is illustrated in Figure 1. The underlying term set Ω is referred to as the *basic domain* of the condensed set-valued attribute $A^\diamond$ (written $\Omega = \mathrm{bdom}(A^\diamond)$).

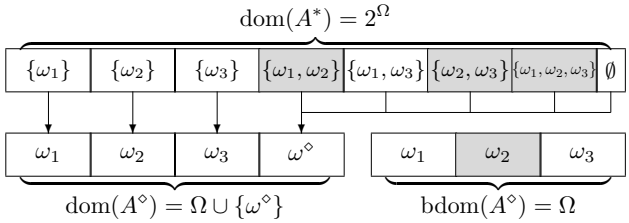

Figure 1: Domains of a set-valued attribute A^*, the induced condensed set-valued attribute $A^\diamond$ and underlying basic domain Ω. Arrows indicate the set reduction mapping w.r.t Ω. Shaded elements of $\mathrm{dom}(A^*)$ mark multi-valued outcomes covering ω_2.

Per Definition 2 the values of $A^\diamond$ depend directly on the values of A^*. Consequently a probability distribution p^* over $\mathrm{dom}(A^*)$ induces a probability distribution $p^\diamond$ over $\mathrm{dom}(A^\diamond)$, which summarizes p^*.

$$\begin{aligned} p^\diamond(\omega) &= P^*(\{S\colon \sigma(S) = \omega\}) \\ &= P^*(\sigma^{-1}(\omega)) \\ &= \begin{cases} p^*(\{\omega\}) & \text{if } \omega \in \mathrm{bdom}(A^\diamond), \\ \sum\limits_{\substack{S \in \mathrm{dom}(A^*) \\ |S| \neq 1}} p^*(S) & \text{if } \omega = \omega^\diamond. \end{cases} \end{aligned} \tag{3}$$

It is important to realize that for any element $\omega \in \mathrm{bdom}(A^\diamond)$, the value $p^\diamond(\omega)$ refers to the probability of ω being the *only* element in an annotation list, rather that just being one of them. The probability mass originally associated with multi-valued outcomes $S\colon S \in \mathrm{dom}(A^*), |S| > 1$ or with the empty-set outcome is assigned to a surrogate attribute value $\omega^\diamond$ in the condensed probability distribution. This approach has two immediate benefits: Since $p^\diamond$ is still a probability distribution, well established operations of the probabilistic framework like conditioning and marginalization can be employed with this representation. In addition to that, Definition 2 can be applied to estimate the condensed probability distributions directly from data, that is without prior computation of the distribution p^*.

Since they represent the non-ambiguous cases, singleton annotations are enriched in many real-world datasets. In the biological application considered here 56.9% of all genes are annotated with just one term. Should all annotations consist of a single term (no ambiguity) the representation is equivalent to a probability distribution over $\mathrm{dom}(A) = \mathrm{bdom}(A^\diamond) = \Omega$ as $p^\diamond(\omega) = p(\omega)$ and $p^\diamond(\omega^\diamond) = 0$ hold for that case. To support the reconstruction of one-point coverages, however, a richer representation is required. Given a probability distribution p^* for

a set-valued attribute A^* taking values from 2^Ω, the one-point coverage of individual elements $\omega \in \Omega$ is computed as follows:

$$\begin{aligned} \forall \omega \in \Omega : \mathrm{opc}(\omega) &= P^*(S : S \subseteq \Omega \wedge \omega \in S) \\ &= \textstyle\sum_{\substack{S \subseteq \Omega \\ \omega \in S}} p^*(S). \end{aligned} \tag{4}$$

For each $\omega \in \Omega$ one element of the sum in the right-hand expression of Equation 4 is obtained directly from the distribution $p^\diamond$ of the induced condensed attribute $A^\diamond$. For $S = \{\omega\}$ the summand is recovered due to the equality $p^*(S) = p^*(\{\omega\}) = p^\diamond(\omega)$. To represent the contribution from all other subsets of Ω, the latter are encoded as proportions relative to $p^\diamond(\omega^\diamond)$ (called coverage factors):

Definition 3 *Let p^* denote a distribution linked to a set-valued attribute (A^*) over 2^Ω and $p^\diamond$ the distribution over the domain $\mathrm{dom}(A^\diamond)$ of an induced condensed set-valued attribute $A^\diamond$ obtained by applying equation 3. Then the* **coverage function** *$c^\diamond$ relative to multi-valued outcomes of A^* is defined as a function*

$$\begin{aligned} c^\diamond : \ \Omega &\rightarrow [0,1] \\ \omega &\mapsto \begin{cases} \dfrac{\sum_{\substack{S \subseteq \Omega, \omega \in S, \\ |S|>1}} p^*(S)}{p^\diamond(\omega^\diamond)} & \text{if } p^\diamond(\omega^\diamond) > 0, \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

For $p^\diamond(\omega^\diamond)$ the value $c^\diamond(\omega)$ denotes the conditional probability for ω being *contained* in a non-singleton outcome. Although the contributions to the one-point coverage could have been stored directly, the representation via relative coverage factors was chosen to better support probabilistic conditioning and marginalization operations. In the case $p^\diamond(\omega^\diamond) = 0$, the conditional coverage factors are undefined, but can be set to a constant. Alternatively the problem can be avoided altogether by using a Laplace correction.

Like the distribution $p^\diamond$, the *relative coverage factors* assigned by $c^\diamond$ can be computed directly from data. Replacing the sum in Equation 4 the one-point coverage may now be rewritten as

$$\begin{aligned} \forall \omega \in \Omega : \mathrm{opc}(\omega) &= \textstyle\sum_{\substack{S \subseteq \Omega \\ \omega \in S}} p^*(S) \\ &= p^\diamond(\omega) + p^\diamond(\omega^\diamond) \cdot c^\diamond(\omega) \end{aligned} \tag{6}$$

In the following, the term *condensed distribution* is understood to refer to a tuple $(p^\diamond, c^\diamond)$ that is formed by a condensed probability distribution and the corresponding coverage function.

The advantage of the condensed set-valued attribute $A^\diamond$ and the function $p^\diamond$ and $c^\diamond$ as compared to the full random set representation is the reduction of the number of parameters. For each term of the attribute domain only the probability for the singleton outcome and the coverage factor need to be stored. For practical reasons, it is also advantageous to explicitly represent the combined probability mass of all multi-valued outcomes, which is required for the calculation of every one-point coverage. This raises the total number of model parameters to $2|\Omega| + 1$. With the condensed random sets the number of distribution parameters grows linearly in the size of the underlying base domain Ω. In contrast, a full distribution over sets would have to encode the probabilities of $2^{|\Omega|}$ possible instantiations.

3 Application to Term Hierarchies

So far the individual annotation terms were considered as largely unrelated. In practise, however, terms are frequently organized in a hierarchy. For several application fields such term hierarchies are specified as part of an ontology. We refer to such term relations using the functions parent_H/$\mathrm{children}_H$ to denote a terms direct predecessors/descendants in the hierarchy and more generally anc_H/desc_H for compatible terms of different specificity. The hierarchical term structure acknowledges that annotations may originate from different sources and provide information on distinct levels of detail. This means that it is no longer sufficient to trace which terms or labels have been used in an annotations itself, but also to consider other applicable terms that are implied. For example, in the hierarchy depicted in Figure 2 the term a_1 is a generalization of a_{12}. Therefore, whenever a_{12} applies to a situation, so does a_1. In contrast, if a case is labeled only as a_1 we do not know which of the more specific labels a_{11}, a_{12}, a_{13} apply (Figure 2). However, the probability of different refinement alternatives may be estimated by looking at the conditional distributions for term usage in fine grained annotations in reference data or simply the remainder of the dataset.

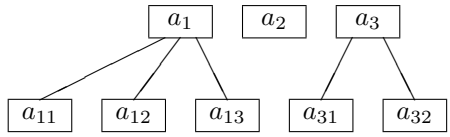

Figure 2: Attribute Value Hierarchy Example

3.1 Model Construction

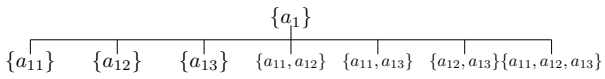

Figure 3: Possible Refinements of Label a_1 in the Hierarchy from Figure 2

A general approach to use the condensed random set framework to deal with to such hierarchies has been described in [Rügheimer and Kruse, 2008]. Each branch in the term hierarchy H is associated with a condensed random set that models the empirical distributions of the possible expanded annotations in reference data. The combined set of labels in the term hierarchy is denoted by $\mathcal{L}$.

The above representation strategy presumes that the expanded annotations w.r.t. different parent labels are (statistically) independent of one another given those parents in the hierarchy. Applied to all expandable labels of the hierarchy, this leads to the data structure depicted in Figure 4. For each non-leaf label λ_r an additional label $\lambda_r^\diamond$ is introduced. In the condensed representation the conditional probability assigned to that label refers to the event that the label λ_r is split into more than one applicable child labels during the next refinement step. This is complemented with conditional coverage factors, which are stored for each element in the direct refinement of λ_r.

To estimate the model parameters from empirical data the Equations 3 and 5 are applied to the branch distributions of non-leaf labels λ_r. The respective reference set is formed by those observations, for which λ_r is both applicable and has been expanded on the observed frame. In

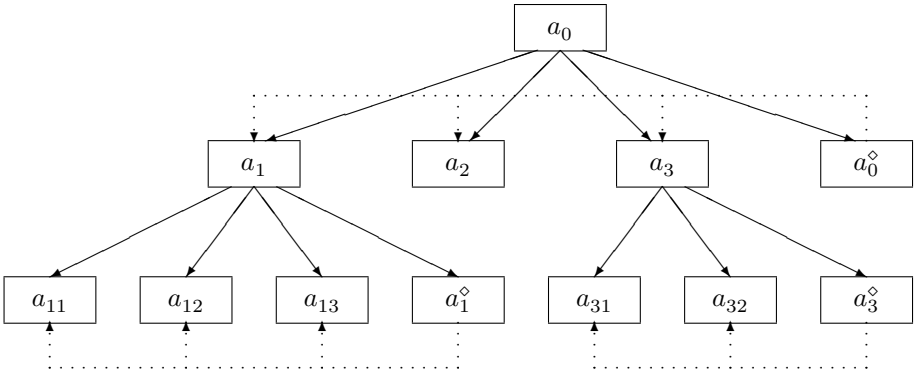

Figure 4: Extended attribute value hierarchy as data structure for the condensed representation of distributions over multi-valued instantiations (conditional probabilities and coverage factors indicated by solid and dotted arrows respectively)

that case information on the applicability of the individual child labels of λ_r is available too. An algorithm to calculate the branch distributions for a given label hierarchy H is given below (Figure 5). For each instantiation from the training data, all compatible nodes in the term hierarchy are marked. Following that, affected branch distributions are traversed to update counters for element coverage (in the case of a multi-label instantiation) or for the occurrence of the respective singleton. Counter updates for the first label in each instantiation are delayed until a distinction of single- and multi- label instantiations becomes possible. After all instantiations have been processed the condensed distribution function and coverage functions are calculated.

The branch distribution on the originally set-valued selections of applicable labels from the elementary refinement of λ_r is represented using the condensed distribution $p^\diamond_{H,\lambda_r}$, with the new element $\lambda^\diamond_r$ representing the non-singleton annotation sets, and the associated coverage function $c^\diamond_{H,\lambda_r}$. The lcorr parameter denotes a user-defined constant for an optional Laplace correction, which is applied for both the induction of branch probabilities and conditional coverage factors (the latter being instances of a two class problem). The bounding of the normalization factors ensures that all marginal probabilities will be defined, even if the Laplace correction is not applied. This guarantee does not extend to conditional branch probabilities though. By altering the normalization factors the algorithm is easily adapted to alternative interpretations of the non-expanded values in the training data set.

3.2 Recalling Information

To facilitate the use of the above representation to model distributions, let us now address how stored information is accessed. To recover a set-distribution from an existing model, the conditional branch distributions on the hierarchy are recombined into respective distributions on the frames. For singleton outcomes this amounts to multiplying branch probabilities along a path of label refinement, i.e. $\forall \lambda \in \mathcal{L}$:

$$p^{*\prime}_H(\{\lambda\}) = \prod_{\lambda' \in (\{\lambda\} \cup \mathrm{anc}_H(\lambda)) \setminus \lambda_0} p^\diamond_{H,\mathrm{parent}_H(\lambda')}(\lambda'). \quad (7)$$

In general the approximation will be imperfect. In addition to the unavoidable sampling error, the branch distributions do not distinguish between real singletons and cases where a label is merely the only applicable element in the local branch. Provided sufficient training data is available,

Algorithm 1 Inducing Branch Distributions

```
procedure GETFREQUENCIES(H,Observations,lcorr)
    RESETCOUNTERS
    for currentObs ∈ Observations do
        for λ ∈ currentObs do
            MARKLABEL(H, λ)
            MARKANCESTORS(H, λ)
        end for
        UPDATECOUNTERS(H)
        obsCnt ← obsCnt + 1
    end for
    λ_r ← ROOTLABEL(H)
    while VALIDLABEL(λ_r) do
        nrm_p ← max {1, GETNUMSGLTEXP(λ_r) + GETNUMMLTVEXP(λ_r)
                    +(1 + |children_H(H, λ_r)|) · lcorr}
        nrm_c ← max {1, GETNUMMLTVEXP(λ_r) + 2 · lcorr}
        for λ ∈ CHILDRENH(λ_r) do
            p^◇_{H,λ_r}(λ) ← (GETNUMASSGLT(λ) + lcorr) / nrm_p
            c^◇_{H,λ_r}(λ) ← (GETNUMASCVRD(λ) + lcorr) / nrm_c
        end for
        p^◇_{H,λ_r}(λ^◇_r) ← (GETNUMMLTVEXP(λ_r) + lcorr) / nrm_p
        λ_r ← NEXTLABELINDEPTHFIRSTSEARCHORDER(H)
    end while
end procedure
```

Algorithm 2 Inducing Condensed Branch Distributions (Counting)

```
procedure UPDATECOUNTERS(H)
    p ← ROOTLABEL(H)
    while VALIDLABEL(p) do
        nMarkedChildren ← 0
        for c ∈ CHILDRENH(p) do
            if ISMARKED(c) then
                if nMarkedChildren = 0 then
                    firstChild ← c
                    nMarkedchildren ← 1
                else
                    COUNTASCOVERED(c)
                    nMarkedChildren ← nMarkedChildren + 1
                end if
            end if
        end for
        if nMarkedChildren = 1 then
            COUNTASSINGLETON(firstChild)
            COUNTSGLTEXPANSION(p)
        else if nMarkedChildren > 1 then
            COUNTASCOVERED(firstChild)
            COUNTMLTVEXPANSION(p)
        end if
        CLEARMARK(H, p)
        p ← NEXTMARKEDLABELINDEPTHFIRSTSEARCHORDER(H)
    end while
end procedure
```

Figure 5: Calculation of Condensed Branch Distributions from Data

a higher precision can be obtained by adding a separate set of branch distributions though.

The one point coverages of individual labels are retrieved by recursively accumulating conditional probabilities and coverage factors for each elementary refinement leading to the label in question. For a single recursion step the reconstructed one-point coverage of a given label is obtained by application of Equation 6. Because each branch distribution refers only to those cases where the respective ancestor labels are applicable, the result is than multiplied with the respective one-point coverage for the ancestors: $\forall \lambda \in \mathcal{L} \neq \lambda_0,\ \lambda_r \stackrel{\text{def}}{=} \text{parent}_H(\lambda):$

$$\begin{aligned} \text{opc}'_H(\lambda) &= \text{opc}'_H(\lambda_r) \cdot \Big(p^{\diamond}_{H,\lambda_r}(\lambda) \\ &+ p^{\diamond}_{H,\lambda_r}(\lambda^{\diamond}_r) \cdot c_{H,\lambda_r}(\lambda) \Big), \end{aligned} \tag{8}$$

where λ_r is used as a shorthand notation for the parent label of λ in the hierarchy and $\lambda^{\diamond}_r$ the corresponding surrogate label that indicates multiple applicable elements in the extension of λ_r. For each level in the hierarchy an additional factor is supplied until the root label λ_0 is reached. If the empty annotation sets are excluded the one-point coverage of that label is always one[1]. To efficiently compute one-point coverages for several elements of a frame an implementation would reuse partial results whenever the recursion runs over shared ancestors in the hierarchy. Under the assumption that applicability of the individual labels within an elementary refinement is independent for non-singleton instantiations, the one-point coverages can also be used to approximate probability values for annotations sets with more than one term, though the approximation quality is lower than for the singletons.

Finally case-specific information on one-point coverages and probabilities can be integrated to allow reasoning. This is achieved by temporarily fixing conditional branch distributions to externally supplied inputs. In the next step the distributions on the target frames are recomputed with the provided values taking precedence over those supplied by the model. Recursions are broken early whenever one of the externally provided values is encountered and only the missing conditional branch probabilities are supplemented by the model.

4 Experimental Evaluation

The evaluation has been conducted on an annotated genome dataset released to the public via the Saccharomyces Genome Database project [SGD Curators, a]. The SGD-project maintains a curated database that summarizes published results about the function of the genes and gene products of the baker's and brewer's yeast Saccharomyces cerevisiae, as well as their respective roles in biological processes and their intracellular activity sites. Annotation follows a domain-wide standard defined in the gene-ontology [The Gene Ontology Consortium, 2000]. The latter also defines term relations that allow to link annotations on different levels of specificity to each other. The terms are organized into three non-overlapping term hierarchies for the tree aspects of annotation (processes, functions, cellular component). Each of these term hierarchies forms a separate branch of the ontology and is connected to the other two only via the common root node.

Since the full annotation is very detailed, a considerable fraction of the annotation terms is only applied to a very small subset of the database. Due to their extremely low term coverage it is not well-justified to include them into a statistical analysis. To provide a standardized broader view of the represented knowledge, less specific versions of the ontology have been released by the consortium. These so-called "slim ontologies" define species-specific subsets of comparatively general Gene Ontology terms and are usually released together with the full annotation data collected in coordinated efforts to analyze the genome and proteome of selected model organisms. The dataset used in the experiments was based on a projection of the full SGD annotations to a subset of relatively broad gene-ontology terms – the GO-Slim terms for yeast [SGD Curators, b]. Term that were not included in the in the slim version of the ontology were mapped to their most specific hierarchical ancestor in the reduced term set. Both that mapping and the GO-Slim itself are maintained at the SGD website.

To evaluate the proposed framework, test its underlying assumptions and compare its predictions with those of alternative frameworks, we implemented three different approaches:

- A model in which presence or absence of elements in a set are encoded using binary variables. The latter variables are treated as independent, so the distribution of set-instantiations is obtained as a product of binary distributions for the state of the elements of the underlying carrier set. The set-distribution is described via its one-point-coverage.
- A condensed distribution model using an unstructured attribute domain
- An enriched term hierarchy using condensed random sets for the representation of branch distribution (described in Section 3).

For the experiment all models were trained using the distribution of annotation sets from a randomly sampled subsets of the yeast genome. The resulting distribution models were then compared with the distribution of the annotation term combinations on the remaining genes. To that end approximation quality and generalization were evaluated using several measures that emphasize either overall quality of fit, the representation of singleton outcomes or the prediction of element coverage. To increase robustness of that evaluation against sampling effects a cross-validation strategy was employed for all experiments.

4.1 Data Preparation and Experimental Setup

Due to the structure of the SGD projects internal database, each assignment of an annotation term to a gene is represented as separate database record. Apart form the gene name and annotation term these records contain supplementary information, such as alternative gene names, the annotation aspect class, types of information sources used to assign the annotation, references to the location of the gene within the genome or connected publications.

Historically several genes have been described and named by two or more research group independently. Often these groups investigated seemingly unrelated biological functions in different organisms. Only later, when refined sequencing and sequence comparison techniques allowed to locate genes within a genome and identify homologue genes in different species, these discoveries have

[1] Otherwise, empty instantiations can easily be represented by inserting a "virtual" root label with an unnormalized branch distribution at the top of the hierarchy. In that case, the one point coverage of the original root label is computed using Equation 8, whereas the one-point coverage of the new root label is set to one.

been found to refer to identical or analogous objects. As a result several genes are known by more than just one name. In order ensure that annotations can be attributed correctly, the first step of preprocessing consisted in mapping all alternative gene names to unique standard identifiers which are used throughout all subsequent processes.

Following that, the records where filtered w.r.t. the annotation aspect given. For the purpose of this evaluation the annotation w.r.t. the "biological process" aspect was chosen. In comparison to the other annotation classes, the annotations on the biological processes provide a comparatively reliable and extensive higher-level description of the role of the gene product in the organism. In the remaining part of the database, annotations for individual genes are still spread over several database records. To better support a gene-based view on the data annotations where grouped by the genes they refer to. The resulting file summarizes the known biological function for each of 6849 genes using 909 distinct annotation sets.

In parallel, the preprocessing routines assembled information about the annotation scheme employed. To that end the term hierarchy structure was extracted from the ontology and converted them into a domain specification for the hierarchical version of the condensed distribution models. Similar domain specifications were prepared for the non-hierarchical version and for the model based on independent binary variables. In those cases however the domain specifications were limited to a list of annotation terms, that is the information on term organization was disregarded. The generated domain specifications were later used to preconfigure distribution models in the training phase.

The above preprocessing method resulted in a database of annotation sets for 6849 genes. To study the properties of the model types this set was split into five partitions with genes randomly assigned (4 partition with 1370 genes each and one partition with 1369 genes). To limit sampling effects, the evaluation measures were computed in a 5-fold cross-validation process [Kohavi, 1995] with a different partition serving as a test data set and the remaining partitions providing training data in each run.

4.2 Parameter Estimation

Using the model configuration files prepared in the preprocessing step and the training data for each validation run, the different model types were trained for the distribution of gene annotation sets. In the case of the reference model with independent binary variables the parameter set consists of one value per element in the carrier set, which describes the probability of an instantiation containing that very element. The modeled probability $\hat{P}(S)$ of any set-instantiation $S \subseteq \Omega$ is obtained by computing the products

$$\hat{P}(S) = \left(\prod_{\omega \in S} \mathrm{opc}(\omega) \right) \cdot \left(\prod_{\omega \in \Omega \setminus S} (1 - \mathrm{opc}(\omega)) \right), \quad (9)$$

with the model parameters $\mathrm{opc}(\omega)$ denoting the (estimated) probability of the element ω being an element of the realization. Coverage rates for elements in the carrier set are estimated from the frequencies of the two possible outcomes "element is present in the instantiation" and "element is absent in the instantiation".

For the condensed distribution and hierarchy-based condensed distribution model the parameters are singleton probabilities and conditional coverage factors either for the distribution as a whole, or – in the hierarchical version – for subtrees of the label hierarchy. For a detailed description of parameters and the model induction procedures see Sections 2 and 3 respectively. In all cases, the parameters were estimated from the observed frequencies in the training data with a Laplace correction of 0.5 applied.

4.3 Evaluation Measures

Having discussed the different model classes, their training and the general evaluation method, we shall now investigate the evaluation measures employed for this task. The measures where chosen to provide complementary information on how well different aspects of the set-distribution are captured by each model type.

Log-Likelihood To describe those measures it is assumed that all models are evaluated against a test data set $D_{\mathrm{tst}} = (d_1, d_2, \ldots, d_m)$ with each d_i formed by the set of annotations applicable to one particular gene. 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 $\hat{P}(D_{\mathrm{tst}} \mid M)$. 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:

$$\log L(D_{\mathrm{tst}}) = \log \prod_{d \in D_{\mathrm{tst}}} P(d \mid M) \quad (10)$$

$$= \sum_{d \in D_{\mathrm{tst}}} \log P(d \mid M). \quad (11)$$

In that formula the particular term used to estimate the probabilities $P(d \mid M)$ of the records in D are model-dependent. Since the likelihood takes values form $[0, 1]$ 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 set are considered as independently sampled instantiations of a multi-valued random variable drawn from the same distribution. The Likelihood of a particular test database of size m is computed as the product of the likelihoods of its m records. 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 11, 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 of $\mathrm{lcorr} = 0.5$ during the training phase. This modification 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-likelihoods 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 as a more suitable measure:

$$\mathrm{arLL}(D_{\mathrm{tst}}) = \frac{\log L(D_{\mathrm{tst}})}{|D_{\mathrm{tst}}|} \tag{12}$$

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 particular 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 the quality of the approximations from an application-oriented viewpoint and compare it to results achieved using other methods, two additional measures – $\mathrm{d_{sglt}}$ and $\mathrm{d_{cov}}$ – have been employed. These measures are based on the sum of squared errors for the respective values over all elements of the base domain:

$$\mathrm{d_{sglt}} = \sum_{\omega \in \Omega} \left(p'(\omega) - p(\omega)\right)^2, \tag{13}$$

$$\mathrm{d_{cov}} = \sum_{\omega \in \Omega} \left(\mathrm{opc}'(\omega) - \mathrm{opc}(\omega)\right)^2. \tag{14}$$

4.4 Experimental Results

For increased robustness of the results the evaluation was conducted using 5-fold cross-validation. In each of the five runs the models were trained using a Laplace correction of 0.5. To obtain a basis for the assessment and comparison of the different methods, the evaluation results of the individual runs were collected and – with the exception of the logL measure[2] – averaged. These results are summarized in the Tables 1–3.

$\log L$	arLL	$\mathrm{d_{sglt}}$	$\mathrm{d_{cov}}$
-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

Table 1: Evaluation Results for Model Using Independent Binary Variables (One-Point-Coverage) with Laplace Correction of 0.5

As anticipated the two condensed random set-based models achieve a considerably better fit to the test data (higher value of arLL-measure) than the model assuming independence of term coverages. Among the two CRS-based models the variant that uses the term hierarchy structure clearly benefits from this additional information and consistently yields better results than its competitor. The large error obtained for the prediction of singleton annotations in the model based on independent binary variables, points out the inadequacy of the independence assumption in the latter representation. 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 separations between real singletons and single elements in local branch distributions appears to leads to a slightly increased error for the hierarchical version. This is consistent with the higher error d_{cov} of that model in the prediction of coverage factors. The two non-hierarchical models represent one-point coverages directly and therefore achieve identical prediction error[3].

[2]See the discussion on the arLL measure to review the argument why averaging Log-Likelihoods is not meaningful here

$\log L$	arLL	$\mathrm{d_{sglt}}$	$\mathrm{d_{cov}}$
-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

Table 2: Evaluation Results for Condensed Distribution on Hierarchically Structured Domain with Laplace Correction of 0.5

$\log L$	arLL	$\mathrm{d_{sglt}}$	$\mathrm{d_{cov}}$
-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

Table 3: Evaluation Results for Condensed Distribution on Unstructured Domain with Laplace Correction of 0.5

5 Relevance for Semantic Similarity Measures

Measures of semantic similarity between concepts have been successfully applied in linguistics, where they are used in Word Sense Disambiguation [Patwardhan *et al.*, 2003], and in bioinformatics, where they are used to in connection with annotation databases to evaluate or enhance clustering or classification algorithms. The Gene Ontology [The Gene Ontology Consortium, 2000] has been developed with the aim of supporting users in using their biological background knowledge to find information in biological databases. But to actually retrieve the desired information it is necessary to relate the users query to stored pieces of information (e.g. documents). The majority of current search engines try to interpret the meaning of the query based on the keywords contained in it. The system uses these keywords to rank results by their degree of similarity to the applied query as defined by a similarity measure. If the keywords are well chosen, these methods fre-

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

quently provide an appropriate list of results. However, if the search terms are ambiguous or are used in different domains, then a rather inhomogeneous collection of results is returned. As this is the case for a considerable number of queries retrieval performance can be improved by applying automatic categorization / filtering techniques to separate those cases. In the biocomputing and the biomedical field semantic similarity measures have been employed to improve document retrieval [Lord *et al.*, 2003], [Pedersen *et al.*, 2007].

Whereas earlier semantic similarity measures were based on graph distance in a term hierarchy some of the more recent variants rely on measures of statistical interaction between pairs of terms [Lin, 1998] and context vectors, which essentially compare relative term coverage between the query and each semantic class [Patwardhan *et al.*, 2003]. In [De Luca, 2008], semantic prototype vectors were constructed from a combination of observed data and extensions acquired from ontological resources.

With condensed random sets it would be possible to obtain a more accurate representation of the distribution of annotation sets. Due to the additional parameters for modelling single valued annotations a typically large fraction of many real world datasets is represented with increased precision. Moreover term interaction are implicitly considered in the hierarchical version of the model. Currently the likelihood measures used in Section 4 are being developed into normalized similarity measures. Already the likelihood based assessment of similarity allow comparisons between groups of annotated objects as well as between groups and individual annotation sets. It can thus be applied both to compare clusters/groups (comparison: distribution–distribution) and to solve classification problems such as word sense disambiguation (comparison: instantiation–distribution).

6 Conclusions

Condensed Random Sets allow to efficiently model probability distributions over annotation sets. Because the number of model parameters is linear in the cardinality of the annotation term set, it can be applied to datasets that are inaccessible to a full random set representation.

It was demonstrated that the assumptions made to achieve this compact representation are in agreement with properties of a relevant real-world biological dataset leading to a high approximation quality – when compared to a reference approach with independent modeling of term annotations. At the same time the condensed representation allows to reduce the problem of overfitting, which constitutes another common problem with full random set representations.

A hierarchical version allows to condition distributions and to supplement information given on different levels of detail. Although the example discussed in this paper refers to a specific problem of biological data analysis, the internal representation employed is general enough to be applied to other random-set based knowledge models in a large field of applications.

References

[De Luca, 2008] Ernesto William De Luca. *Semantic Support in Multilingual Text Retrieval*. Shaker Verlag, Aachen, Germany, 2008.

[Kohavi, 1995] Ron Kohavi. 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 95)*, pages 1137–1145, 1995.

[Lin, 1998] D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning, Madison, WI*, pages 296–304, Madison, WI, USA, 1998.

[Lord *et al.*, 2003] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, pages 601–612, 2003.

[Nguyen, 1978] Hung T. Nguyen. On random sets and belief functions. *Journal Math. Anal. Appl.*, 65:531–542, 1978.

[Patwardhan *et al.*, 2003] S. Patwardhan, S. Banerjee, and T. Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257, Mexico City, Mexico, February 2003.

[Pedersen *et al.*, 2007] Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299, 2007.

[Rügheimer and Kruse, 2008] Frank Rügheimer and Rudolf Kruse. 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.

[Rügheimer, 2007] Frank Rügheimer. A condensed representation for distributions over set-valued attributes. In *Proc. 17. Workshop Computational Intelligence*, Karlsruhe, Germany, 2007. Universitätsverlag Karlsruhe.

[SGD Curators, a] SGD Curators. Saccharomyces genome database. (accessed 2008/11/16).

[SGD Curators, b] SGD Curators. SGD yeast gene annotation dataset (slim ontology version). via Saccharomyces Genome Database Project [SGD Curators, a]. (accessed 2008/11/16).

[The Gene Ontology Consortium, 2000] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[Toledo-Arana *et al.*, 2009] Alejandro Toledo-Arana, Olivier Dussurget, Georgios Nikitas, Nina Sesto, Hélène Guet-Revillet, Damien Balestrino, Edmund Loh, Jonas Gripenland, Teresa Tiensuu, Karolis Vaitkevicius, Mathieu Barthelemy, Massimo Vergassola, Marie-Anne Nahori, Guillaume Soubigou, Béatrice Régnault, Jean-Yves Coppée, Marc Lecuit, Jörgen Johansson, and Pascale Cossart. The listeria transcriptional landscape from saprophytism to virulence. *Nature*, May 2009.

[Wasserman, 2006] Larry Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, New York, 2006.