BMC Bioinformatics

BioMed Central

Software

Open Access

ProbCD: enrichment analysis accounting for categorization uncertainty

Ricardo ZN Vêncio and Ilya Shmulevich*

Address: Institute for Systems Biology, 1441 North 34th street, Seattle, WA 98103-8904, USA

Email: Ricardo ZN Vêncio - rvencio@gmail.com; Ilya Shmulevich* - ishmulevich@systemsbiology.org

* Corresponding author

Published: 12 October 2007

BMC Bioinformatics 2007, **8**:383 doi:10.1186/1471-2105-8-383

Received: 11 July 2007
Accepted: 12 October 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/383

Abstract

Background: As in many other areas of science, systems biology makes extensive use of statistical association and significance estimates in contingency tables, a type of categorical data analysis known in this field as enrichment (also over-representation or enhancement) analysis. In spite of efforts to create probabilistic annotations, especially in the Gene Ontology context, or to deal with uncertainty in high throughput-based datasets, current enrichment methods largely ignore this probabilistic information since they are mainly based on variants of the Fisher Exact Test.

Results: We developed an open-source R-based software to deal with probabilistic categorical data analysis, ProbCD, that does not require a static contingency table. The contingency table for the enrichment problem is built using the expectation of a Bernoulli Scheme stochastic process given the categorization probabilities. An on-line interface was created to allow usage by non-programmers and is available at: http://xerad.systemsbiology.net/ProbCD/.

Conclusion: We present an analysis framework and software tools to address the issue of uncertainty in categorical data analysis. In particular, concerning the enrichment analysis, ProbCD can accommodate: (i) the stochastic nature of the high-throughput experimental techniques and (ii) probabilistic gene annotation.

Background

The system-level approach to data analysis known as enrichment analysis (also known as over-representation or enhancement analysis) is now commonplace. Moreover, the number of available software tools to perform such analysis is large (see [1,2] for comprehensive reviews). The preferred way to formalize the enrichment problem is by means of a contingency table, often 2 × 2.

The mathematical problem is conceptually generic, being applied to diverse types of data, such as genomics, transcriptomis or proteomics datasets; diverse types of analysis, including multiple and/or ordered outcomes; and diverse types of gene classification schemes, such as Gene Ontology (GO), KEGG or organism-specific ones. For a given ontology term t defining the set of genes G_t and its complementary set G_t^c, the general enrichment analysis contingency table is:

	G_t	G_t^c
$outcome_1$	$X_{1,1}$	$X_{1,2}$
$outcome_2$	$X_{2,1}$	$X_{2,2}$
$\cdots$	$\cdots$	$\cdots$
$outcome_k$	$X_{k,1}$	$X_{k,2}$

Besides measuring the statistical significance of the null hypothesis that the rows and columns are independent, as yielded by Fisher's Exact Test [3] and Fisher-like methods [1,2], it is also possible to measure statistical association between a table's rows and columns [4] (a detailed discussion on significance vs. association in the enrichment problem context can be found in [5]).

Most of the attention in the enrichment analysis problem has focused on issues such as the search for the best multiple-test correction or the implementation of better user-friendly software interfaces to facilitate biologist's exploratory work [1]. However, one of the limitations that the available approaches still share is that they assume, explicitly or implicitly, that one is able to construct the contingency table exactly, without uncertainty in populating its cells. Some efforts to consider ranked lists of genes, ranked by their reliability, were proposed to ameliorate the aforementioned limitations [6], however they do not work on the categorical data framework and incorpore the probabilitic information in a heuristic fashion [7].

Recently, the computational biology community has been witnessing an increasing interest in probabilistic approaches to gene annotation, particularly in the Gene Ontology (GO) context, as a realization of the limitations imposed by the traditional deterministic and context-independent gene annotation schemes [8-15]. These efforts are motivated by: the necessity to assess the error propagation in automatic gene annotation [9,15]; desire to include different types of evidence sources such as protein-protein interaction [8,13] or phylogenomics [10,12] and annotation extrapolation from model organisms to others [11,14]. Meanwhile, the probabilistic nature of data obtained by high-throughput measurement techniques is well recognized and a number of attempts to model it were proposed over the past decade in various experimental contexts [16,17]. However, these efforts are not integrally taken into account when usual enrichment analysis is performed.

We describe a computational solution that is able to deal with the uncertainty introduced in enrichment analysis due to: (i) the stochastic nature of the results obtained with such high-throughput experimental techniques or (ii) probabilistic gene annotation.

Implementation

ProbCD is an open-source software designed to perform probabilistic categorical data analysis. ProbCD is written in R [18] with a level of modularity that makes it suitable to be incorporated by existing development efforts of integrative tools [19]. To facilitate the usage by researchers with no knowledge of R, we implemented a user-friendly web-based interface for the software, which is not limited to any particular organism. The on-line interface and the source-code are available on the project's website [20].

The idea behind ProbCD's implementation is to formally represent the intuitive process of building a contingency table in a probabilistic manner. Informally speaking, each element to be placed in the contingency table is not considered to be indivisible, but instead is "shared", according to probabilistic rules, among the contingency table's cells in a manner that is conceptually similar to fuzzy membership. The theoretical and computational implementation aspects are described in detail below.

Without loss of generality, the following descriptions are applied considering one particular ontology term t that is associated with a set of genes, named simply as G_t. It should be noted that G_t is not restricted to the Gene Ontology categorization and can be any kind of classification or annotation.

The vector $\boldsymbol{q}$ contains a probabilistic annotation for all g of the organism's genes: $q_j = \mathbb{P}\ (gene_j \in G_t)$ for $j \in \{1, \cup, g\}$. This probabilistic annotation is assumed to be given, typically obtained from some analysis process. The deterministic scenario corresponds simply to $\mathbb{P}\ (gene_j \in G_t) \in \{0, 1\}$, and hence is a special case.

The matrix $\mathbf{P}$ contains a probabilistic description for all k possible outcomes of the property being studied. Therefore, $\mathbf{P}$ is a $k \times g$ matrix with elements $P_{i,j} = \mathbb{P}\ (gene_j \in outcome_i)$ for $j \in \{1, \cup, g\}$ and $i \in \{1, \cup, k\}$. This probabilistic description of the data uncertainty is assumed to be given.

To motivate the general probabilistic model, it is useful to examine an arbitrary 2 × 2 example in the deterministic scenario:

	G	G^c
H	$x_{1,1}$	$x_{1,2}$
H^c	$x_{2,1}$	$x_{2,2}$

where all x's are the counts of a regular contingency table over the gene sets G and H. In its matrix representation:

$$\begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \end{pmatrix} = \begin{pmatrix} \sum_j \mathbf{1}_{\{gene_j \in H\}} \mathbf{1}_{\{gene_j \in G\}} & \sum_j \mathbf{1}_{\{gene_j \in H\}} \mathbf{1}_{\{gene_j \in G^c\}} \\ \sum_j \mathbf{1}_{\{gene_j \in H^c\}} \mathbf{1}_{\{gene_j \in G\}} & \sum_j \mathbf{1}_{\{gene_j \in H^c\}} \mathbf{1}_{\{gene_j \in G^c\}} \end{pmatrix}$$

where $\mathbf{1}_{\{\}}$ is the indicator function.

Inspired by this representation, it is easy to see that the "hard" indicator functions may be substituted by Bernoulli random variables in order to account for the categorization uncertainty. Since all sets are finite, the indicator functions can be represented as vectors in $\{0, 1\}^g$ and the sums over all genes as dot products. In a generic scenario, with given non-deterministic $\mathbf{P}$ and $\boldsymbol{q}$, the contingency table represented by $\mathbf{X}|\mathbf{P}, \boldsymbol{q}$ is a random matrix that is difficult to describe in closed form. It is also not compatible with the statistical formalism supporting Fisher's Exact Test or other well-known Fisher-like approaches, as these are not applicable to random tables.

The contingency table is defined in terms of Bernoulli Schemes [21] which is the generalization of the Bernoulli Process to more than two possible outcomes. The notation $\boldsymbol{Z} \sim \mathrm{Be}(p_1, \cup, p_n)$ represents the distribution:

$$z = \begin{cases} (1,0,0,\cdots,0) & \text{with probability } p_1; \\ (0,1,0,\cdots,0) & \text{with probability } p_2; \\ (0,0,1,\cdots,0) & \text{with probability } p_3; \\ \cdots & \\ (0,0,0,\cdots,1) & \text{with probability } p_n. \\ p_1 + \cdots + p_n = 1 & \end{cases}$$

The random variable $\mathbf{X}$ is a matrix representation of a $k \times 2$ contingency table:

$$\begin{pmatrix} X_{1,1} & X_{1,2} \\ \cdots & \cdots \\ X_{k,1} & X_{k,2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{d}_1 \cdot \boldsymbol{a}_1 & \boldsymbol{d}_1 \cdot \boldsymbol{a}_2 \\ \cdots & \cdots \\ \boldsymbol{d}_k \cdot \boldsymbol{a}_1 & \boldsymbol{d}_k \cdot \boldsymbol{a}_2 \end{pmatrix}$$

where $\cdot$ is the usual dot-product, $\boldsymbol{a}_i = (A_{i,\,1}, \cup, A_{i,\,g})$ is a row-vector of a $2 \times g$ binary matrix $\mathbf{A}$ such that $(A_{1,\,j}, A_{2,\,j})|q_j \sim \mathrm{Be}(q_j, 1 - q_j)$ and $\boldsymbol{d}_i = (D_{i,\,1}, \cup, D_{i,\,g})$ is a row-vector of a $k \times g$ binary matrix $\mathbf{D}$ such that $(D_{1,\,j}, \cup, D_{k,\,j})|(P_{1,\,j}, \cup, P_{k,\,j}) \sim \mathrm{Be}(P_{1,\,j}, \cup, P_{k,\,j})$.

It is very easy to extend this framework for completely generic $k \times m$ tables ($m > 2$), but this would be outside the scope of the ontology enrichment problem.

To measure statistical association between rows and columns in contingency tables, analogously to correlations for non-categorical data, we recall the pivotal works by L.A. Goodman and W.H. Kruskal [4]. Depending on the problem under consideration, an appropriate association measure function ρ can be chosen. ProbCD calculates the statistical association accounting for the stochastic nature of the table's categorization, reporting $\rho = \rho\,(\mathbb{E}\ [\mathbf{X}|\mathbf{P}, \boldsymbol{q}])$, where $\mathbb{E}$ is the expectation operator. If the categorical data is represented by a regular 2 × 2 matrix, then Yule's Q can be used as the statistical association function $\rho \equiv Q : \mathbb{R}^4 \to [-1, 1]$. If one is dealing with ordered contingency tables, then Goodman-Kruskal's gamma, $\rho \equiv \gamma : \mathbb{R}^{2k} \to [-1, 1]$, can be used since it is the generalization of Yule's Q. Considering non-ordered categories, there is no analogy with the usual correlations in [-1, 1] and in this case, as suggested by [4], Cramer's T is used with $\rho \equiv T : \mathbb{R}^{2k} \to [0, 1]$.

All the association measures implemented can be calculated for $\mathbb{E}\ [\mathbf{X}|\mathbf{P}, \boldsymbol{q}] \in \mathbb{R}^{2k}$, while 2 × 2 Fisher's Exact Test p-value cannot, since it is a function in $^4 \to [0, 1]$. Moreover, a p-value is related to the significance only, containing no information about the actual association level.

The dichotomous case, which is the simplest one, gives a more intuitive illustration on how the association is calculated in practice for the particular implementation: $\mathbb{E}\ [X_{1,1}|\mathbf{P}, \boldsymbol{q}] = E_{1,\,1} = P_{1,\,1}\,q_1 + \cup + P_{1,\,g}\,q_g$, $\mathbb{E}\ [X_{2,\,1}|\mathbf{P}, \boldsymbol{q}] = E_{2,\,1} = (1 - P_{1,\,1})\,q_1 + \cup + (1 - P_{1,\,g})\,q_g$, $\mathbb{E}\ [X_{1,\,2}|\mathbf{P}, \boldsymbol{q}] = E_{1,\,2} = P_{1,\,1}(1 - q_1) + \cup + P_{1,\,g}\,(1 - q_g)$, $\mathbb{E}\ [X_{2,\,2}|\mathbf{P}, \boldsymbol{q}] = E_{2,\,2} = (1 - P_{1,\,1})\,(1 - q_1) + \cup + (1 - P_{1,\,g})\,(1 - q_g)$ and $\rho = (E_{1,\,1}\,E_{2,\,2} - E_{1,\,2}\,E_{2,\,1})/(E_{1,\,1}\,E_{2,\,2} + E_{1,\,2}E_{2,\,1})$, which corresponds to Yule's Q.

To measure the statistical significance of the estimated association, ProbCD uses a randomization approach. The null distribution for the association measure, ρ^*, is proposed to be estimated from several permutation rounds. In each round a gene j receives randomly its probabilities $(P^*_{1,j}, \cdots, P^*_{k,j})$ from one of the g possible columns of $\mathbf{P}$ and an association value is calculated. The significance of the statistical association between rows and columns in the contingency table is calculated as $p = \mathbb{P}\ (\rho^* \geq \rho)$. A term t is significantly over-represented (or equivalently, the gene list is enriched for t) depending on user-defined thresholds for significance and/or association.

Results

The following examples illustrate the potential utility of considering probabilistic annotations and/or data uncertainty assessment in the enrichment analysis using

ProbCD on artificial datasets and a published yeast dataset.

The point of the following illustration is to show that even ontology terms annotated with modest probabilities can be considered to be over-represented if the list of genes obtained behave in a supportive pattern. Consider a hypothetical organism with 100 genes annotated in several GO terms, as described in the Additional Files. The genes $gene_1$ to $gene_{20}$ are deterministically annotated to the ontology term $t = a$. In other words, assume that it is well known that these 20 genes have some given functionality a. The experiment, for example from a hypothetical proteomics dataset, yielded a deterministic list of differentially expressed (DE) genes ranging from $gene_1$ to $gene_{10}$. The contingency table for this problem is, therefore:

	G_a	G_a^c
DE	10	0
DE^c	10	80

In this case, the DE gene list is clearly enriched for a within any meaningful significance cutoff. Consider now a second ontology term b obtained from a probability-based source with $\mathbb{P}$ $(gene_i \in G_b) = 40\%$, $i \in \{1, \cup, 20\}$. A probability of only 40% generally would not be sufficient evidence to warrant the inclusion of those 20 genes in G_b considering a usual deterministic framework and, therefore, would not be analyzed by deterministic-based methods, such as the Fisher's Exact Test. However, ProbCD is able to incorporate this information and yields: $\rho = 0.87$ and $p < 10^{-4}$ in 10000 permutation rounds, a significant enrichment for b. One can easily imagine, for example, genes that have a main function a but also have a different function b in, say, 40% of documented conditions.

The point of the following illustration is to show that the incorporation of probabilistic annotation information does not always translate to addition of terms into the enrichment result, as in the example above, but it can also mean the exclusion of non-relevant terms. Consider a hypothetical organism with 1100 genes. Let the genes $gene_1$ to $gene_{100}$ be grouped together in a cluster H after some genomic sequence analysis. Let the term a be annotated deterministically (Additional Files) yielding the contingency table:

	G_a	G_a^c
H	100	0
H^c	100	900

In this situation, H is clearly enriched for a within any meaningful significance cutoff. Now let the same annotation incorporate some evidence levels by defining: $\mathbb{P}$ $(gene_i \in G_a) = 99\%$ for $i \in \{1, \cup, 10\}$ and $\mathbb{P}$ $(gene_i \in G_a) = 1\%$ for $i \in \{11, \cup, 100\}$. Intuitively, this means that only 10 out of 100 genes clustered in H are, in fact, confidently annotated with the ontology term a. The incorporation of this information results in non-significant enrichment of H for a since: $\rho = 0.0425$ and $p = 0.42$ in 1000 permutation rounds. Therefore, it can be useful to incorporate uncertainty information into the enrichment analysis to also down-rank potentially spurious enrichment results.

The following illustration shows that the use of ordered categories ($k > 2$) can produce useful results when additional information, regarding the order, is added. Consider a hypothetical organism with 4000 genes. In a hypothetical network analysis, let the genes be categorized, for simplicity and without loss of generality, in a deterministic fashion in a natural order: hubs (H), regular nodes (N) and leaves (L). Let the term a be annotated deterministically (Additional Files) yielding the contingency table:

	G_a	G_a^c
H	15	2
N	5	378
L	180	3420

If one cannot express the difference between hubs and regular nodes in the enrichment analysis, the contingency table is forced to be described as:

	G_a	G_a^c
$H + N$	20	380
L	180	3420

The most connected nodes in the network are not enriched for a considering the consolidated table above using either the Fisher's Exact Test (p-value = 0.54) or ProbCD ($\rho = 0$, $p = 0.48$). However, using the original categorization order, ProbCD suggests a significant enrichment for a with $\rho = 0.98$ and $p < 10^{-4}$. The conclusion that the property a must be related to gene connectivity seems subjectively reasonable considering the numbers in the first contingency table. The rationale used for the hypothetical network analysis could be useful in other scenarios where there is a natural order that can provide extra information such as: highly expressed, expressed, and not expressed or up-regulated differentially expressed, not differentially expressed, and down-regulated differentially expressed.

The next illustration demonstrates the impact of considering the uncertainty in lists of genes, rather than in the annotations, on the enrichment analysis. In this example, the aim is to find which GO terms, annotating the yeast *Saccharomyces cerevisiae*, are statistically associated with periodic expression levels, measured by microarray technology [22]. Andersson and colleagues [22] devised an elaborate Bayesian model which produces the probability that a gene is periodically expressed during the cell-cycle. Since the final probability values are sufficient for our objectives in this work, we refer the interested reader to the original work by Andersson and colleagues [22] for more details. In this example, the annotation is considered to be deterministic and was downloaded from the GO project page (March 2007) [23].

To perform the usual enrichment analysis one needs to define a probability cutoff value in order to split the gene list in two: the periodic genes and the non-periodic genes. Consider initially the reasonable cutoff $\mathbb{P}$ ($gene_i$ is periodic) ≥ 70% and focus on a single GO term GO:0007090 (regulation of S phase of mitotic cell cycle), defined as "*a cell cycle process that modulates the frequency, rate or extent of the progression through the S phase of mitotic cell cycle*". Although this GO term is clearly associated with periodic gene expression, performing a usual enrichment analysis results in the conclusion that the periodic genes are not significantly enriched for GO:0007090 within usual significance cutoffs (*p*-value = 0.065).

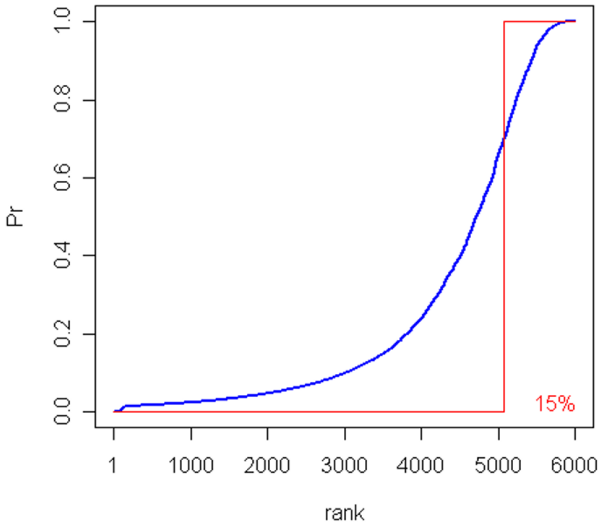

Figure 1
Probability of being periodic. The blue curve represents the probability of a gene being periodic (Pr) according to the model of [22]. The genes are sorted by probability values (rank) on the horizontal axis to facilitate the visualization. The red curve is the deterministic approximation using a 70% probability cutoff to consider a gene as periodic: $\mathbb{P}$ ($gene_i$ is periodic) ≥ 0.70 ⇒ $\mathbb{P}$ ($gene_i$ is periodic) = 1 and $\mathbb{P}$ ($gene_i$ is periodic) < 0.70 ⇒ $\mathbb{P}$ ($gene_i$ is periodic) = 0. This approximation labels 15% of the genes as periodic.

Suspecting that this non-intuitive result could be due to the probability threshold chosen to select periodic genes, illustrated in the Figure 1, one could repeat the same analysis above building the contingency table considering the cutoffs $\mathbb{P}$ ($gene_i$ is periodic) ≥ 50%, 95%, 99% or 99.99%. The result of this repeated analysis is also non-intuitive since the *p*-values are: 0.12, 1.0, 1.0 and 1.0 for 50%, 95%, 99% and 99.99% cutoffs, respectively, meaning that increasing the stringency to define a gene as periodic only decreases the significance of the enrichment for GO:0007090.

Using ProbCD, one can consider the actual probability of being periodic (blue curve in Figure 1) in the enrichment analysis instead of using the deterministic approximation (red curve in Figure 1). This results in a relatively high statistical association between periodicity and the term "regulation of S phase of mitotic cell cycle" (ρ = 0.78) with high significance (p = 0.009 in 1000 simulation rounds). Judging subjectively by the definition of GO:0007090, ProbCD returned a meaningful result.

Other similar cases can be easily identified. For example, the GO term GO:0000083 (G1/S-specific transcription in mitotic cell cycle) exhibits erratic behavior depending on the chosen cutoff for the probability of being periodic: *p*-value of 0.15, 0.10, 0.01, 0.096 and 1.0 for 50%, 75%, 95%, 99% and 99.99% cutoffs, respectively. The probability stringency used to build the contingency table and the subsequent significance test are not necessarily correlated. ProbCD yielded a significant (p = 0.006) moderate association (ρ = 0.48) for GO:0000083. Other examples include GO:0045787 (positive regulation of progression through cell cycle), defined as "*any process that activates or increases the frequency, rate or extent of progression through the cell cycle*", which would be called significant using the regular enrichment method only if the right probability cutoff $\mathbb{P}$ ($gene_i$ is periodic) ≥ 95% is guessed initially: *p*-value of 0.047, 0.024, 0.0058, 0.086 and 0.024 for 50%, 75%, 95%, 99% and 99.99%, respectively.

The above analysis process is repeated for all GO terms, with the results available as Additional Files and summarized in Figure 2. This figure suggests that there is a large variability in the possible final outcome of an enrichment analysis depending on the probability cutoff used to build the associated contingency table. This variability is avoided by ProbCD because it directly takes into account the uncertainty in the data instead of introducing a discretization step (Figure 1).

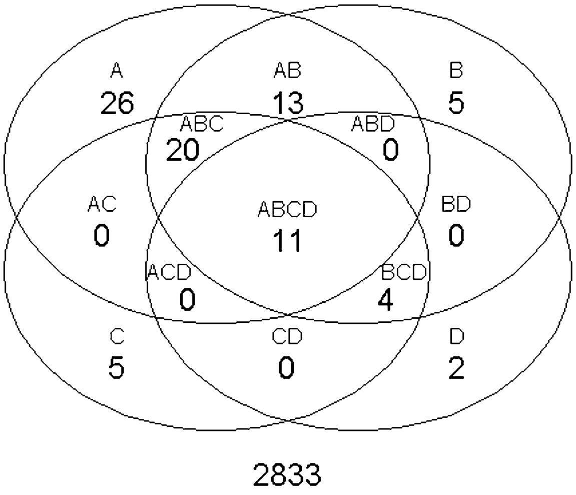

Figure 2
Venn diagram of over-represented terms. The Venn diagram shows the number of GO terms considered significantly over-represented (p-value ≤ 0.01) by the Fisher Exact Test using four different probability cutoffs $\mathbb{P}$ ($gene_i$ is periodic) ≥ A, B, C or D ⇒ periodic: A = 0.70, B = 0.95, C = 0.99 and D = 0.9999.

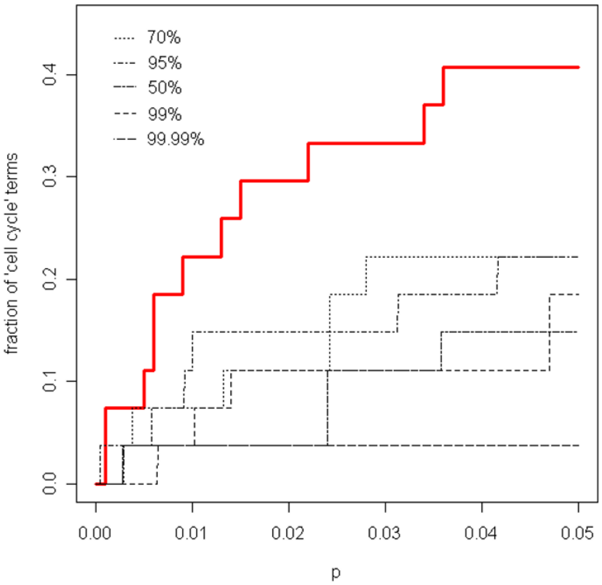

Figure 3
Fraction of "cell-cycle" GO terms selected as a function of the *p*-value. The curves show the fraction of GO terms containing the word "cell-cycle" in their definition that are considered significant as a function of the significance cutoff (p). The red curve is obtained with ProbCD and all others are obtained with one of the probability cutoffs: 50%, 70%, 95%, 99% or 99.99%.

Figure 3 shows that ProbCD considers more terms (vertical axis in Figure 3) containing the word "cell cycle", likely associated to periodically expressed genes, as significant if compared to the usual enrichment analysis in a wide range of significance values (p in Figure 3). Although this is not a proof, since one cannot be certain about which "cell cycle"-marked terms should be enriched, this is a reasonable indication that one can, in fact, avoid the discretization step when building the enrichment problem using ProbCD and obtain meaningful results.

Discussion and Conclusion

The usual enrichment analysis is a particular case in this probabilistic framework and can be obtained by ProbCD ignoring the difference between evidence sources in gene annotation and defining fixed gene lists, which would correspond to the deterministic setting: $q_j = \mathbb{P}$ ($gene_j \in G_t$) = 1 or 0 and $P_{i,\,j} = \mathbb{P}$ ($gene_j \in outcome_i$) = 1 or 0.

Even if a probabilistic annotation is not readily available for a given organism, it could be interesting to perform enrichment analysis taking into account some form of weighting on available annotations according to their reliability. For a concrete example, the GO Consortium [24] provides annotations accompanied with evidence codes related to the kind/level of evidence available for a given GO annotation [25], such as *IEA: Inferred from Electronic Annotation*, *IMP: Inferred from Mutant Phenotype*, *RCA: inferred from Reviewed Computational Analysis* or *IDA: Inferred from Direct Assay*. It is known that some evidence sources are more reliable than others and this knowledge can be used, in a Bayesian sense, as subjective probabilities.

Once an annotation is considered in a probabilistic framework, it could reflect a dependence on the context. One can consider cases in which $\mathbb{P}$ ($gene_j \in G_t$|disease) ≫ $\mathbb{P}$ ($gene_j \in G_t$), defining context-dependent gene annotations derived, for instance, from automatic literature mining [26].

Our intention is to complement existing approaches, rather than substitute them. Toward this aim, we built ProbCD to be as modular as possible in order to be incorporated into existent software or pipelines [19], composed of ontology pre-processing [27] or powerful visualization capabilities [28,29].

It is important to note that ProbCD is also applicable to other categorical data analysis contexts in which the construction of contingency tables is subject to uncertainty, a recurrent theme in science.

Availability and requirements

- Project Name: ProbCD

- Project Home Page: http://xerad.systemsbiology.net/ProbCD

- Operating Systems: platform independent

- Programming Languages: R

- License: GNU Lesser General Public License 3.0

Authors' contributions

RZNV implemented the project. IS supervised the project. All authors read and approved the final manuscript.

Additional material

Additional file 1

ProbCD source-code and Examples*. Source-code used to build the ProbCD package. Future upgrades will be available at the project website [20]. Dataset and results for the three examples presented in the Results section.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-383-S1.zip]

Acknowledgements

We thank Drs. John Boyle, Vesteinn Thorsson, Nathan Price and other ISB colleagues for insightful discussions. We also thank Dr. Carlos A. de B. Pereira (USP) for introducing us to the works of L.A. Goodman and W.H. Kruskal. This work is partially supported by NIH grants U54-AI54253, U19-AI057266 and P50-GMO-76547.

References

1. Dopazo J: **Functional Interpretation of Microarray Experiments.** *OMICS: A Journal of Integrative Biology* 2006, **10(3):**.
2. Rivals I, Personnaz L, Taing L, Potier M: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2007, **23(4):**401-407.
3. Fisher R: **On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P.** *Journal of the Royal Statistical Society* 1922, **85:**87-94.
4. Goodman L, Kruskal W: **Measures of Association for Cross Classifications.** *Journal of the American Statistical Association* 1954, **49(268):**732-764.
5. Vencio R, Koide T, Gomes S, Pereira C: **BayGO: Bayesian analysis of ontology term enrichment in microarray data.** *BMC Bioinformatics* 2006, **7:**86.
6. Jiang Z, Gentleman R: **Extensions to gene set enrichment.** *Bioinformatics* 2007, **23(3):**306.
7. Goeman J, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics* 2007, **23(8):**980.
8. Joshi T, Chen Y, Becker J, Alexandrov N, Xu D: **Genome-Scale Gene Function Prediction Using Multiple Sources of High-Throughput Data in Yeast Saccharomyces cerevisiae.** *Omics A Journal of Integrative Biology* 2004, **8(4):**322-333.
9. Levy E, Ouzounis C, Gilks W, Audit B: **Probabilistic annotation of protein sequences based on functional classifications.** *BMC Bioinformatics* 2005, **6:**302.
10. Engelhardt B, Jordan M, Muratore K, Brenner S: **Protein molecular function prediction by Bayesian phylogenomics.** *PLoS Comput Biol* 2005, **1(5):**.
11. Martin D, Berriman M, Barton G: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.** *BMC Bioinformatics* 2004, **5:**178.
12. Engelhardt B, Jordan M, Brenner S: **A graphical model for predicting protein molecular function.** *Proceedings of the 23rd international conference on Machine learning* 2006:297-304.
13. Carroll S, Pavlovic V: **Protein classification using probabilistic chain graphs and the Gene Ontology structure.** *Bioinformatics* 2006, **22(15):**1871.
14. Vinayagam A, del Val C, Schubert F, Eils R, Glatting K, Suhai S, König R: **GOPET: A tool for automated predictions of Gene Ontology terms.** *BMC Bioinformatics* 2006, **7:**161.
15. Jones C, Brown A, Baumann U: **Estimating the annotation error rate of curated GO database sequence annotations.** *BMC Bioinformatics* 2007, **8:**170.
16. Zhang W, Shmulevich I: *Computational and Statistical Approaches to Genomics* 2nd edition. New York, NY, USA: Springer; 2006.
17. Zhang W, Shmulevich I, Astola J: *Microarray Quality Control* Wiley-Liss; 2004.
18. **The R Project for Statistical Computing** [http://www.r-project.org]
19. Shannon P, Reiss D, Bonneau R, Baliga N: **Gaggle: An open-source software system for integrating bioinformatics software and data sources.** *BMC Bioinformatics* 2006, **7:**176.
20. **ProbCD Home Page** [http://xerad.systemsbiology.net/ProbCD]
21. **Bernoulli scheme – Wikipedia, The Free Encyclopedia** [http://en.wikipedia.org/w/index.php?title=Bernoulli scheme&o%ldid=64557593]
22. Andersson C, Isaksson A, Gustafsson M: **Bayesian detection of periodic mRNA time profiles without use of training examples.** *BMC Bioinformatics* 2006, **7:**63.
23. **Gene Ontology Current Annotations** [http://www.geneontology.org/GO.current.annotations.shtml]
24. **The Gene Ontology Consortium** [http://www.geneontology.org]
25. **Guide to GO Evidence Codes** [http://www.geneontology.org/GO.evidence.shtml]
26. Aubry M, Monnier A, Chicault C, de Tayrac M, Galibert M, Burgun A, Mosser J: **Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets.** *BMC Bioinformatics* 2006, **7:**241.
27. Lewin A, Grieve I: **Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data.** *BMC Bioinformatics* 2006, **7:**426.
28. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks.** *Bioinformatics* 2005, **21(16):**3448-3449.
29. Sealfon R, Hibbs M, Huttenhower C, Myers C, Troyanskaya O: **GOLEM: an interactive graph-based gene-ontology navigation and analysis tool.** *BMC Bioinformatics* 2006, **7:**443.