BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 5 2006, pages 566–572
doi:10.1093/bioinformatics/btk019

Gene expression

Improving missing value estimation in microarray data with gene ontology

Johannes Tuikkala[1,3,*], Laura Elo[2,3,4], Olli S. Nevalainen[1,3,4] and Tero Aittokallio[2,3,4]

[1]Department of Information Technology, University of Turku, Lemminkäisenkatu 14A, FIN-20520, Finland, [2]Department of Mathematics, University of Turku, FIN-20014 Finland, [3]Turku Centre for Computer Science (TUCS), Lemminkäisenkatu 14A, FIN-20520, Finland and [4]Turku Centre for Biotechnology, Tykistökatu 6, FIN-20521, Finland

Received on June 14, 2005; revised and accepted on December 20, 2005
Advance Access publication December 23, 2005
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Gene expression microarray experiments produce datasets with frequent missing expression values. Accurate estimation of missing values is an important prerequisite for efficient data analysis as many statistical and machine learning techniques either require a complete dataset or their results are significantly dependent on the quality of such estimates. A limitation of the existing estimation methods for microarray data is that they use no external information but the estimation is based solely on the expression data. We hypothesized that utilizing a priori information on functional similarities available from public databases facilitates the missing value estimation.

Results: We investigated whether semantic similarity originating from gene ontology (GO) annotations could improve the selection of relevant genes for missing value estimation. The relative contribution of each information source was automatically estimated from the data using an adaptive weight selection procedure. Our experimental results in yeast cDNA microarray datasets indicated that by considering GO information in the *k*-nearest neighbor algorithm we can enhance its performance considerably, especially when the number of experimental conditions is small and the percentage of missing values is high. The increase of performance was less evident with a more sophisticated estimation method. We conclude that even a small proportion of annotated genes can provide improvements in data quality significant for the eventual interpretation of the microarray experiments.

Availability: Java and Matlab codes are available on request from the authors.

Supplementary material: Available online at http://users.utu.fi/jotatu/GOImpute.html

Contact: jotatu@utu.fi

INTRODUCTION

Gene expression microarrays provide a popular technique to monitor the relative expression of thousands of genes under a variety of experimental conditions (Schena *et al*., 1995; Lockhart *et al*., 2000). In spite of the enormous potential of this technique, there remain challenging problems associated with the acquisition and analysis of microarray data that can have a profound influence on the interpretation of the results. A particular drawback of the techniques is that running the microarray experiment can be technically rather error prone. Microarray class slide may contain dust and scratches or the spotting and hybridization processes can fail resulting in incomplete information for some spots on the slides. The microarray users typically filter out corrupted or suspicious spots during the image analysis phase. Therefore, the microarray data frequently contain missing values that may seriously disturb or even prevent the subsequent data analysis.

In order to understand why missing values can be such a problem, De Brevern *et al*. (2004) analyzed eight publicly available microarray datasets. They discovered that the proportion of missing values is typically at least 5% of all values, and in most datasets >60% of genes contain at least one missing value. It was also observed that missing values drastically reduce the performance of different data analysis techniques such as clustering of gene expression profiles (De Brevern *et al*., 2004). Owing to the high number of genes and experiments involved with missing values we cannot simply discard the ones with missing values or repeat the experiments, but we need to use some method to estimate (or impute) the missing values as accurately as possible before continuing the actual data analysis.

Estimation of missing data is a well-studied problem in the statistical literature and imputation methods have traditionally been used in several data analysis applications (Little and Rubin, 1987). Recently, such methods have been reinvented and extensively applied to the imputation of microarray data; see e.g. the comparative study by Feten *et al*. (2005). Several imputation techniques have been proposed for microarray data including *k*-nearest neighbor (*k*-NN) (Troyanskaya *et al*., 2001), local least squares (LLS) (Kim *et al*., 2005), Bayesian approach (Oba *et al*., 2003) and collateral missing value imputation (Sehgal *et al*., 2005). It has been recognized that while simple average expression of the gene gives sufficient estimates in data without correlation structure, more sophisticated imputation methods should be used for data with significant correlations among the genes or conditions. In the latter case, methods based on *k*-NN can provide robust estimates. Accordingly, the imputation process is typically divided into two steps. In the first step, a set of genes nearest to the gene with a missing value is selected. The second step involves the prediction of the missing value using observed values of the selected genes. The present work focuses mainly on defining a proper similarity measure in the first step of the imputation process.

To our knowledge, it is a common property of all the imputation methods that the estimation is based solely on the expression data

*To whom correspondence should be addressed.

itself. However, one could easily think that well-established a priori information, for example, on functional similarities of the genes should facilitate the missing value estimation, in particular when the expression data are somehow limited. We illustrate this idea using gene ontology (GO) database as a source of external information (Ashburner *et al*., 2000). While GO is typically employed subsequent to gene expression data analysis, e.g. for verifying clustering results of different algorithms (Gibbons *et al*., 2002) or defining enriched functional classes of differential gene expression (Beißbarth and Speed, 2004), here the GO annotations are used already as an integral part of the data imputation process together with the expression data.

To investigate the influence of GO annotation on the estimation accuracy, we combine the semantic similarity in the GO with the expression similarity in the k-NN imputation algorithm. To investigate the influence of the prediction method on the results, we use also a more advanced LLS algorithm. The results of the GO-based algorithms are compared with those of the two original algorithms at different rates of missing values in four datasets comprising both the time series and the non-time series data. The datasets are examined in terms of the ontology used [molecular function (MF) or biological process (BP)], the number of experimental conditions and the accuracy of gene annotation. In Supplementary material, we have collected the results also from some other investigations, including the effects of release date and reduced ontologies. In Discussion, we provide additional directions how to further improve this general estimation framework.

METHODS

Imputation algorithms

Expression data from a series of microarray experiments can be represented as an $M \times N$ matrix $\mathbf{G} = (g_{ij})_{i,j=1}^{M,N}$, where the entries of $\mathbf{G}$ are the expression ratios for M genes under N different conditions. The columns represent the conditions and the rows identify the genes $\mathbf{g}_i = (g_{ij})_{j=1}^{N}$ for $i = 1, 2, \ldots, M$. For simplicity of representation, we consider the situation where only one of the genes contains a single missing value, whereas the other genes are completely observed i.e. the data from all N conditions are known. The general case with multiple missing values in several genes can be handled by reducing the numbers M and N accordingly. We also suppose without losing generality that the missing value occurs in the first component of the first gene $\mathbf{g}_1$. The other cases can be considered by reordering the rows or the columns of $\mathbf{G}$.

Perhaps the most commonly used missing value estimation method is the weighted k-NN imputation (Troyanskaya *et al*., 2001). The k-NN algorithm first selects k genes nearest to the gene $\mathbf{g}_1$ according to the Euclidean distance d_i between the vectors $\mathbf{w}_1 = (g_{1j})_{j=2}^{N}$ and $w_i = (g_{ij})_{j=2}^{N}$ from the gene set $i = 2, 3, \ldots, M$. Let $\mathbf{b} = (b_i)_{i=1}^{k}$ denote the k-dimensional vector consisting of the first components of such k neighboring genes and $\mathbf{A}$ the $k \times N - 1$ matrix of the remaining entries in these genes. An estimate for the missing value g_{11} is then computed as the weighted average:

$$\hat{g}_{11} = \frac{\sum_{i=1}^{k} b_i/d_i}{\sum_{i=1}^{k} 1/d_i}. \quad (1)$$

One of the most promising new missing value estimation methods is the LLS imputation (Kim *et al*., 2005). The LLS algorithm uses the k-NN process to select k nearest genes and then predicts the missing value using the least squares formulation for the neighborhood genes $\mathbf{A}$ and the non-missing entries $\mathbf{w}_1$ of $\mathbf{g}_1$. In the simplified case with only a single missing value, the resulting estimate is computed as the linear combination:

$$\hat{g}_{11} = \mathbf{b}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}})^{\dagger}\mathbf{w}_1, \quad (2)$$

where $(\mathbf{A}^{\mathrm{T}})^{\dagger}$ is the pseudoinverse of the transposed matrix $\mathbf{A}^{\mathrm{T}}$.

Semantic similarity in GO

GO is a structured network of defined terms which describe gene product attributes (Ashburner *et al*., 2000). It consists of three independent ontologies: (1) MF considers the biochemical activity of the gene products at the molecular level; (2) BP refers to a biological objective to which the gene or gene product contributes and (3) cellular component deals with the place in the cell where a gene product is active. Terms of ontologies can be arranged as a directed acyclic graph (DAG), where each node can have several parent terms and several child terms but the parent–child relation does not form a cycle (i.e. no term is an ancestor of itself). There may be several different relationships between a term and its parent, but the most common arcs in a DAG describe 'is-a' and 'part-of' relationships. An annotated gene can be associated to one or more terms of the ontology and a term of the ontology can have one or more genes associated with it (Carey, 2003). All available known associations between GO accession ids and genes are listed in the annotation file (called corpus) of a given organism.

GO can be used to describe the semantic similarity between the terms and hence to provide a way to measure the functional similarity of annotated genes. Lord *et al*. (2003) proposed an information content-based measure of semantic similarity which considers a term in the ontology that is rather general containing less information than a term that is more specific and rare. They defined the information content $p(c)$ for each term c in the ontology as the 'probability' that the term occurs in the corpus being used. A term occurs if the term itself or any of its children occurs and $p(c)$ is the number of occurrences of the term divided by the total number of all different term occurrences. The semantic dissimilarity of two terms c_1 and c_2 is then measured by the information content $p(c)$ of minimum subsumer c of c_1 and c_2. If an unambiguous c is not found then the minimum $p(c)$ is selected from the set of minimum subsumers. Lord *et al*. (2003) used negative logarithm of minimum $p(c)$ as semantic similarity measure.

GO-based imputation

The semantic dissimilarity is used here as an external information on the functional similarity of two genes. According to earlier results (Allocco *et al*., 2004), we investigated the influence of both BP and MF ontologies on the imputation accuracy. The calculation of semantic dissimilarity starts by building an ontology tree T from GO–DAG so that nodes y which have several parents are duplicated, as previously described by Lee *et al*. (2004). The ontology tree is created from the ontology flat-file downloaded from the GO website (http://www.geneontology.org/). An annotation table A from the annotation file (corpus) is also created and used to fetch all GO accession ids that are associated with a given gene (i.e. in A there is an entry for each GO accession id and a list of genes associated with the accession id). Based on these data structures

```
SEMANTIC DISSIMILARITY
Input: Gene g1, Gene gi, GOTree T, Annotation table A
Output: Value of semantic dissimilarity d' ∈ [0, 1]

Let P = {} be set of information content values of minimum
subsumers
Find GO ids ids1 from A, which are associated with g1
Find GO ids ids2 from A, which are associated with gi
for all idi ∈ ids1
  for all idj ∈ ids2
    Find nodes n1 from T which GO id is idi
    Find nodes n2 from T which GO id is idj
    Y = set of shared ancestor nodes of n1 and n2
    P = P ∪ {min_{y∈Y}{y.p}}
  end for
end for
return mean(P)
```

Fig. 1. Our pseudocode of the Lords *et al*. (2003) algorithm for calculating the semantic dissimilarity between two genes $\mathbf{g}_1$ and $\mathbf{g}_i$.

the information content p-value $y.p$ for each node y in the tree is calculated.

The semantic dissimilarity $d'(\mathbf{g}_1, \mathbf{g}_i)$ between two genes $\mathbf{g}_1$ and $\mathbf{g}_i$ is calculated using the Semantic Dissimilarity algorithm presented in Figure 1. The smaller the d' the more similar the genes $\mathbf{g}_1$ and $\mathbf{g}_i$ are in their function. The algorithm calculates the semantic dissimilarity similarly as done in Lord *et al*. (2003). Briefly, the algorithm first finds the sets of GO accession ids (GO ids) for both genes from the annotation table A. All ids are iterated and for each id pair the set of shared ancestor nodes is found from the ontology tree T. For each id pair the minimum value of the information content of shared ancestor nodes is stored in the set P. Finally when all id pairs are checked we use the mean of P as the final value for semantic dissimilarity of genes $\mathbf{g}_1$ and $\mathbf{g}_i$ (Lord *et al*., 2003). If shared ancestor nodes are not found, then semantic dissimilarity value $d' = 1$ is used for the gene pair $\mathbf{g}_1$ and $\mathbf{g}_i$.

We apply the GO-annotations in the imputation algorithms to guide the selection process so that the set of genes $\{\mathbf{g}_i\}_{i=1}^{k}$ selected for predicting the missing value of gene $\mathbf{g}_1$ are not only close in their expression values but also in function. The semantic dissimilarity is incorporated into the imputation algorithms by replacing the expression level distance $d(\mathbf{g}_1, \mathbf{g}_i)$ with the combined distance c_i defined as

$$c_i = d'(\mathbf{g}_1, \mathbf{g}_i)^{\alpha} \cdot d(\mathbf{g}_1, \mathbf{g}_i), \tag{3}$$

where the positive weight parameter α controls how much the semantic dissimilarity value contributes to the combined distance between the genes $\mathbf{g}_1$ and $\mathbf{g}_i$. The weight $\alpha = 0$ means that the semantic dissimilarity has no contribution at all and the larger the α the more the semantic dissimilarity affects the selection of the genes. A small value of d' implies that genes $\mathbf{g}_1$ and $\mathbf{g}_i$ are semantically close to each other and their combined distance c_i is therefore reduced from the expression level distance accordingly. This procedure ensures that only the most specific terms of the ontology (small values of d') have a significant effect on the imputation result.

Testing procedure

All the data used in the evaluating of the imputation algorithms were constructed by first removing the genes (rows) with one or more missing expression values from the datasets, thus yielding originally complete data matrices. Starting with this observation matrix we then generated datasets with missing values by randomly setting certain percent of values as missing (between 1 and 20%). Multiple missing values were allowed in the same gene.

As an evaluation criterion, we calculated the root mean squared difference between the original and imputed values of the missing entries, divided by the root mean squared original values in these entries (referred to as NRMS error). An advantage of such an error function is that the zero imputation obtains always unit error, providing a useful reference value for comparisons across different datasets.

The strength of the correlation structure between the genes in a particular dataset was investigated using the eigenvalues of the covariance matrix of the expression data, as was previously described (Feten *et al*., 2005). Equal distribution of the eigenvalues indicates weak correlation structure, whereas one or several relatively large eigenvalues is an indication of stronger correlation structure in the dataset.

The selection of the neighborhood size k was done by evaluating the imputation accuracy of the pure k-NN and the GO-based k-NN methods with different values of k. We observed that 20 neighbors were enough for each of the datasets and thus the value of $k = 20$ was used in each test run (see Supplementary figure 1). The neighborhood size of the LLS-based imputation algorithms was fixed to 150 in accordance with the studies of Kim *et al*. (2005).

The selection of α was done using an adaptive selection procedure in each dataset to be imputed. First, it selects a fixed proportion of genes from the generated data and marks one non-missing value of each selected gene as 'temporally missing'. Then, it estimates these values separately for each α value using the GO-based estimation. The α that produces the smallest overall NRMS error is selected for the actual missing value estimation process. We used 20% of the genes for the α selection procedure according to our experiments (see Supplementary figures 2–4).

To further study the effects of the number of experimental conditions or the number of annotated genes, we also randomly removed columns from the data matrix or the annotations from the corpus file, respectively. The evolution of GO was studied using the older BP ontology files. A total of 10 random missing value datasets were generated for each test situation and for each missing value percentage. The results are reported as mean error along with the standard error of the mean (SEM).

RESULTS

Datasets

The datasets used for testing the imputation accuracies consisted of public yeast cDNA microarray data downloaded from the *Saccharomyces* Genome Database (SGD) website (http://sgdlite.princeton.edu/). The corpus is the SGD annotation file from the GO website (http://www.geneontology.org/). A summary of the characteristics of the datasets is shown in Table 1. The strength of the correlation structure of a dataset C was determined as the ratio between the first eigenvalue of the covariance matrix and the sum of all eigenvalues as shown in Supplementary figure 5.

The first dataset (diauxic) is from a study of temporal gene expression during the metabolic shift from fermentation to respiration in *Saccharomyces cerevisiae* (DeRisi *et al*., 1997). The second

Table 1. Summary of the test datasets

Name	M	M'	N	$M\%$	$A\%$	C
Diauxic	6068	5875	7	0.46	90.60	0.73
Elutriation	6075	5766	14	0.38	90.53	0.40
Histone	6181	6169	7	0.02	90.32	0.61
Phosphate	6013	5783	8	0.77	91.25	0.41

M is the number of genes in the original dataset and M' after the genes with missing values are removed. N is the number of conditions in the microarray experiment. $M\%$ is the original percentage of missing values and $A\%$ is the percentage of genes that have an annotation in the GO file. The last column is the strength of the correlation structure of the datasets. The diauxic, elutriation and histone datasets are time-series data, whereas the phosphate dataset is non-time series data.

dataset (elutriation) is the elutriation part from the Spellman *et al*. (1998) yeast cell-cycle microarray material. The third dataset (histone) is from a study of the nucleosomes and silencing factor effects on global gene expression (Wyrick *et al*., 1999). The last dataset (phosphate) is from a phosphate accumulation and polyphosphate metabolism study in *S.cerevisiae* (Ogawa *et al*., 2000).

Effect of the ontology used

Test results from the two imputation algorithms at different rates of missing values with BP and MF ontologies are represented in Figure 2. The GO-based *k*-NN method produces more accurate imputation results than the pure *k*-NN imputation in the diauxic and histone datasets, especially at large missing value rates. Even if the imputation accuracy in the elutriation dataset is not improved with GO information when all the 14 conditions are used, the GO-based *k*-NN outperforms the pure *k*-NN method when the number of conditions is reduced (see Section 3.3). The LLS imputation algorithm gives generally much better results than the *k*-NN and the usage of GO could not improve these results any further.

The overall performance of imputations based on the BP ontology seemed to be somewhat higher than imputations based on the MF, especially when the missing value percentage is high (20%). However, in phosphate dataset, the MF ontology is generally better than the BP. This is the only non-time series dataset investigated in this study. In other cases, the adaptive α selection procedure ensures that the GO-based imputation methods are at least as good as original methods (Fig. 2). The correlation structures of the datasets under study were generally rather strong and no clear indication was found whether the strength of the correlation structure influences the imputation accuracies of the different imputation methods.

In the case of diauxic, histone and phosphate datasets the estimation accuracy is rather unsteady when the missing value percentage is small, as indicated by the relative large SEM values in Figure 2. This is because the prediction error differs considerably for some genes and the difference is noticeable only when small amount of the genes have missing observations (Feten *et al*., 2005). Another interesting observation is the behavior of the LLS methods in the phosphate and histone datasets: when the missing value percentage increases from 10 to 15% the NRMS error unexpectedly decreases from 0.78 to 0.76, which is outside the error intervals. A similar phenomenon is also visible in the original LLS-study by Kim *et al*. (2005), but the reason is unknown.

Effect of the number of conditions

We further studied how the number of conditions affects the imputation accuracy of the pure *k*-NN and the GO-based *k*-NN methods. This was investigated in the elutriation dataset which originally has the largest value of conditions ($N = 14$). Figure 3 shows that the imputation accuracy of the *k*-NN methods degrades markedly as the number of conditions decreases. The benefit gained from GO annotations is more evident when less conditions are available, especially at larger missing value rates (Fig. 3B). In particular, the GO-based imputation outperforms the *k*-NN imputation for each missing value percentage if the number of conditions is ≤ 6. We observed similar behavior also in another dataset with many experimental conditions ($N = 18$), which confirms these results (see Supplementary figure 6).

Effect of the accuracy of annotation

We first tested the imputation accuracy of the GO-based *k*-NN method with some older BP ontology files. It turned out that the evolution of the BP ontology has significant influence on the imputation accuracy of the GO-based *k*-NN method only if the percentage of missing values is high (see Supplementary figure 7). We next studied how the increase in the number of annotated genes affect the imputation accuracy of the GO-based *k*-NN method in the diauxic dataset. It can be noticed from Figure 4 that the extent of annotations clearly contributes to the imputation accuracy of the GO-based *k*-NN method. The imputation accuracy was enhanced especially at higher missing value rates, where even a small proportion of annotated genes provides improvements distinctly outside the error intervals (Fig. 4).

DISCUSSION

We have described the integration of external information in terms of GO annotations into the *k*-NN and LLS imputation algorithms. The experimental results suggested that GO improves the imputation accuracy, especially when the number of experimental conditions is small or the proportion of annotated genes is large. In each experiment, the benefits gained from using GO were emphasized at higher rates of missing values. We therefore recommend the use of GO information with an imputation if the number of conditions is <10 and in particular if the percentage of missing values is sufficiently large (say >10%). The choice of the ontology type (either BP or MF) had no marked influence on the prediction accuracies. The prediction method itself, on the other hand, had the greatest influence on the imputation results. One of our future goals is therefore to improve the prediction of LLS with GO in Equation (2), similar to what is done in the *k*-NN prediction (1).

Beyond the results presented above, we have also carried out a number of additional experiments with the proposed imputation method. For instance, we studied the usage of the GO Slims (generic or yeast) instead of the full versions of GO (BP or MF). Although the number of terms in Slims is much smaller than in the full ontology, the imputation results were not so dramatically changed, except in the diauxic dataset (see Supplementary figure 8). Next, we reduced the MF ontology file by deleting the subgraphs that are related to the transferase or the kinase activity. The objective was to study the effect of such general terms on the imputation method. The results show that there were no significant differences in the imputation accuracy as compared with the original ontology

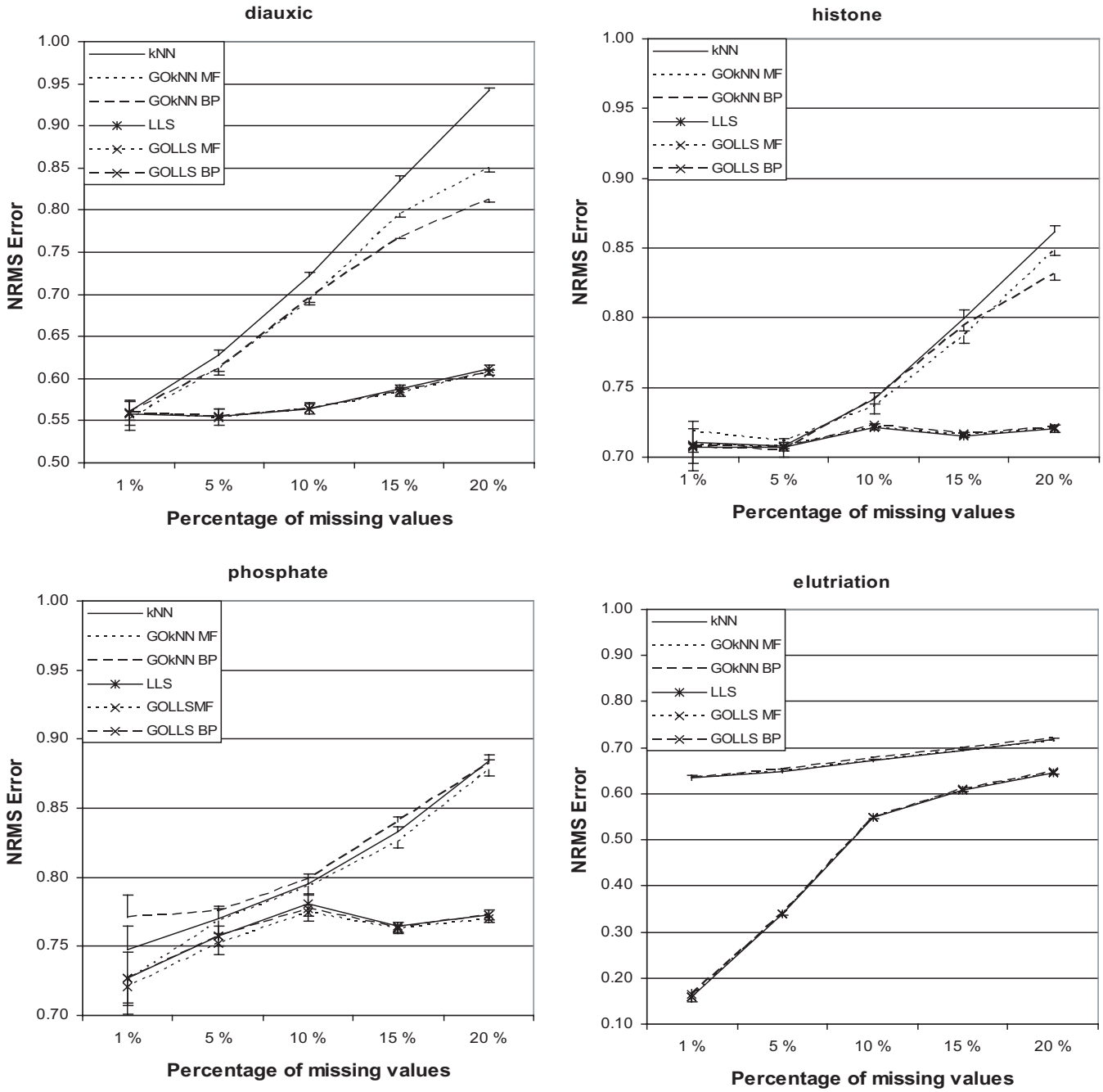

Fig. 2. Comparison of the NRMS errors of the imputation methods at different percentages of missing values. Error bars indicate SEM. GOkNN, *k*-NN with GO. LLS, pure LLS method. GOLLS, LLS with GO. kNN, pure *k*-NN method. MF, molecular function ontology. BP, biological process ontology. Some results are so close to each other that the lines are not separable. Note that the scales on *y*-axis are different in each graph. The maximum value is fixed to one, which is the error obtained with zero imputation.

(see Supplementary figure 9). This owes to the fact that, in addition to these coarse-level terms, most genes also have more specific terms, which are taken into consideration when calculating the semantic similarity.

We investigated the distribution of the MF terms with the largest impact on the imputation accuracy by locating the 100 best and the 100 worst terms in the MF ontology. It turned out that the transferase activity sub-ontology was one of the sources of bad terms (see Supplementary figure 10). We also investigated the effect of constraining the GO annotations using the traceable author statement or inferred from direct assay evidence codes only. With MF ontology, this impairs the imputation accuracy on average, whereas with BP ontology, the result is vice versa. However, these differences as compared with all annotations were rather small

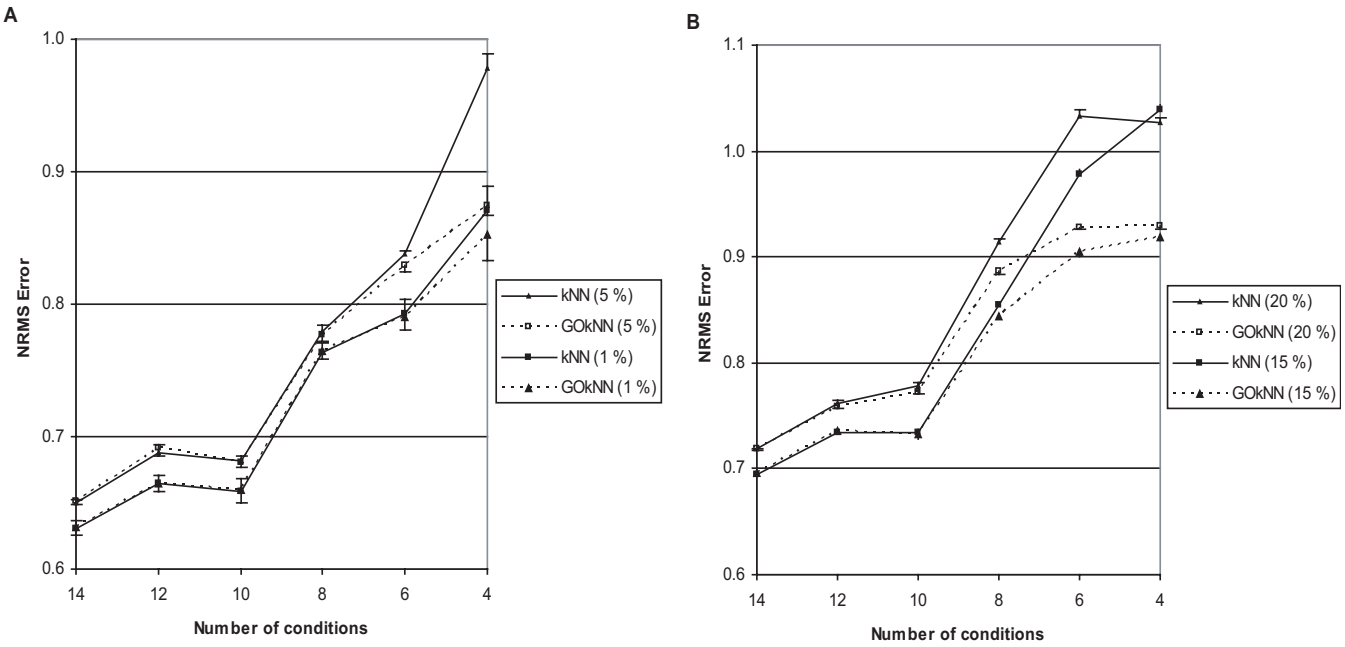

Fig. 3. Comparison of the NRMS errors of the *k*-NN and the GO-based *k*-NN methods for four different missing value percentages when the number of conditions varies. Elutriation dataset and biological process ontology are used here.

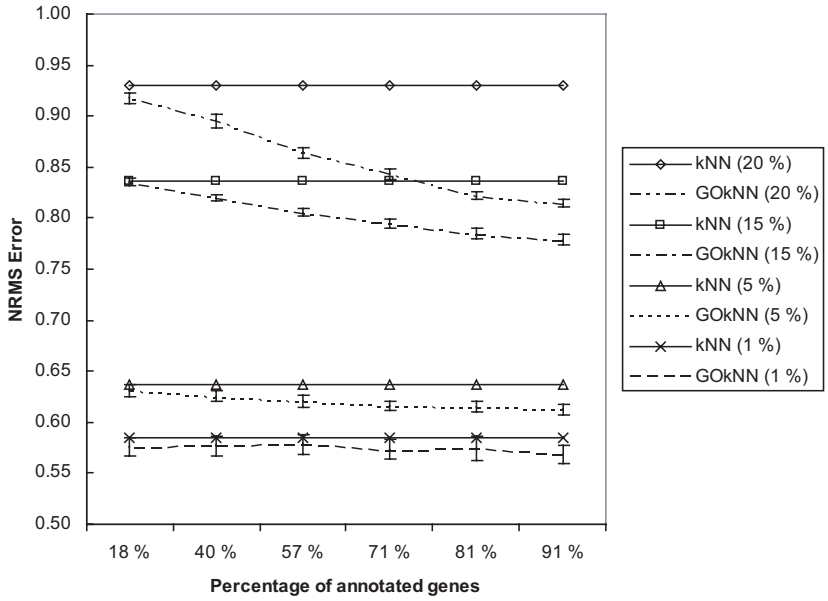

Fig. 4. Comparison of the NRMS errors of the *k*-NN and the GO-based *k*-NN methods for four different missing value percentages when the percentage of annotated genes varies. Diauxic dataset and biological process ontology are used here.

(see Supplementary figure 11). The eventual choice whether to use the full ontology (BP or MF) or somehow reduced versions of GO (Slims or evidence codes) depends on the data accuracy requirements and the computational power available (see Supplementary figure 12).

The proposed method is generic in the sense that it describes a general imputation framework, with an emphasis placed on the selection of the neighboring genes, whereas several different methods can be used at other phases of the imputation process. First of all, any prediction algorithm can be used to estimate the missing

values subsequent to finding the set of neighboring genes. Besides the *k*-NN and LSS algorithms used here, a wide range of alternative prediction methods have recently been suggested, based on singular value decomposition (Liu *et al*., 2003), linear or non-linear regression (Zhou *et al*., 2003), Bayesian principal component analysis (Oba *et al*., 2003) or expectation maximization algorithm (Bø *et al*., 2004). Interesting modifications of the existing algorithms are the sequential imputation methods, which can be especially useful at high missing value rates (Kim *et al*., 2004). As all these imputation methods are based on the expression data, the proposed method cannot make reliable gene expression predictions without this principal source of information.

Second, there are several different techniques for selecting the set of informative genes for prediction. We tested different ways to combine the semantic similarity with expression similarity in Equation (3), and found out that the best way is to use d' as an additional weight in Euclidean distance between genes. However, the semantic similarity could be incorporated also in other expression similarity metrics such as Pearson's correlation coefficient used in many gene expression imputation algorithms (Bø *et al*., 2004; Kim *et al*., 2005). Moreover, the GO-guided neighborhood selection should be helpful even in more sophisticated selection techniques, such as Bayesian gene selection (Zhou *et al*., 2003) or Gaussian mixture clustering (Ouyang *et al*., 2004). Comprehensive comparison of different combinations of selection and prediction methods is, however, outside the scope of the present work.

Third, there exist alternatives to define the semantic similarity in GO as well. Gibbons and Roth (2002) constructed a table indicating whether or not a gene is known to possess an attribute. Such a table could be used as a measure of dissimilarity of two genes using e.g. Manhattan distance. Lee *et al*. (2004) constructed a tree from the GO DAG structure so that nodes with several parents were duplicated. They calculated the similarity of two nodes in the ontology by finding their lowest common ancestor and using its weighted distance from the root. However, the various semantic similarity measures are quite different and there is no good comparison between them in the literature. We used the semantic similarity measure of Lord *et al*. (2003) because it is well-suited to our problem and it has been properly tested against sequence similarities of gene products.

Finally, some other external information on the functional relatedness of the genes instead of GO could be used, for instance, the similarity of their protein sequences (Raghava and Han, 2005). It is also probable that more advanced techniques will be developed to define gene functions, increasing further the need for integrated computational analysis of expression data. However, irrespective of the nature of external information employed in the imputation process, it comprises only the starting point in the analysis of microarray experiments. As the eventual goal is not to predict the expression values themselves, but rather to facilitate revealing the most meaningful interpretations, methods that repeat the actual data analysis with multiple imputed values or can even handle the missing values as a part of the analysis process are likely to be developed in the gaze of bioinformatics research in the coming years.

ACKNOWLEDGEMENTS

The work of L.E. and T.A. was supported by the Academy of Finland (grant 203 632) and the graduate school in Computational Biology, Bioinformatics, and Biometry (ComBi).

Conflict of Interest: none declared.

REFERENCES

Allocco,D.J. *et al*. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinfromatics*, **5**, 18.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Beißbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Bø,T.H. *et al*. (2004) LSimpute: accurate estimation of missing values in microarray data with least squares method. *Nucleic Acids Res*., **32**, e34.

Carey,V.J. (2003) Ontology concepts and tools for statistical genomics. *J. Multivariate Anal*., **90**, 213–228.

De Brevern,A.G. *et al*. (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, **5**, 114.

DeRisi,J.L. *et al*. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Feten,G. *et al*. (2005) Prediction of missing values in microarray and use of mixed models to evaluate the predictors. *Stat. Appl. Genet. Mol. Biol*., **4**, 10.

Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res*., **12**, 1574–1581.

Kim,H. *et al*. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**, 187–198.

Kim,K.Y. *et al*. (2004) Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, **5**, 160.

Lee,S.G. *et al*. (2004) A graph-theoretic modelling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381–388.

Liu,L. *et al*. (2003) Robust singular value decomposition analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 13167–13172.

Little,R.J.A. and Rubin,D.B. (1987) *Statistical Analysis with Missing Data*. Wiley, NY.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Lord,P.W. *et al*. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Oba,S. *et al*. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Ogawa,N. *et al*. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell*, **11**, 4309–4321.

Ouyang,M. *et al*. (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, **20**, 917–923.

Raghava,G.P. and Han,J.H. (2005) Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics*, **6**, 59.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Sehgal,M.S. *et al*. (2005) Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, **21**, 2417–2423.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarray. *Bioinformatics*, **17**, 520–525.

Wyrick,J.J. *et al*. (1999) Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature*, **402**, 418–421.

Zhou,X. *et al*. (2003) Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics*, **19**, 2302–2307.