

Probabilistic Consistency Analysis for Gene Selection

Sach Mukherjee* and Stephen J. Roberts
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, U.K.
{sach,sjrob}@robots.ox.ac.uk

Abstract

A great deal of recent research has focused on the problem of selecting differentially expressed genes from microarray data ('gene selection'). Recent theoretical work has shown that the effectiveness of a gene selection algorithm can be captured as a probability called 'selection accuracy'. Unfortunately, in practice, there tends to be relatively little known about the very features upon which selection accuracy depends, making it difficult to choose a suitable method. In this paper we present a 'consistency analysis' which allows the inference of posterior distributions over selection accuracy from data. The utility of our approach lies in the fact that it can be used to assess gene selection algorithms in a practical but principled manner, and thus choose an appropriate method for given experimental data.

1. Introduction

The widespread use of microarrays has meant that gene selection has become one of the most practically important problems in statistical bioinformatics. While numerous methods have been proposed in the literature [4], recent theoretical work [2] has shown that subtle features of the underlying biological system can have serious effects on selection accuracy, to the extent that many widely-used methods may produce quite spurious results. Incorrect results lead to a subsequent waste of time and resources, often with no real 'sanity-check' until late in the investigative life-cycle. It is therefore critically important to choose an algorithm appropriate for given experimental data. In principle, selection accuracy is jointly determined by the form of the selection function and the statistical properties of the biological system under study. Unfortunately, in practice, these very statistical properties tend to be unknown. The absence of labeled data (i.e. datasets with genes flagged as rel-

* to whom correspondence should be addressed

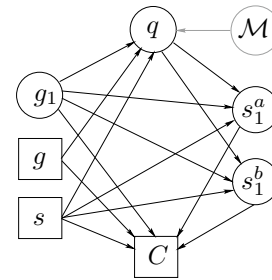


Figure 1. Graphical model for consistency.

evant/irrelevant) further means that there is no obvious way to assess selection accuracy. In this paper we address these issues by exploiting a simple but powerful notion of *consistency*, which turns out to be an effective guide to selection accuracy and algorithm choice.

2. Consistency

Definition: Consider two sets of microarray data \mathcal{D}_a and \mathcal{D}_b pertaining to the same biological problem. Each dataset contains expression levels for the same genes; a ranking function r produces two orderings of those genes from the two datasets. Let the s highest-ranked genes in each case be selected as results and denoted by SET_a and SET_b respectively. Sample consistency C is then defined as:

$$C(r, s, \mathcal{D}_a, \mathcal{D}_b) \stackrel{def}{=} |SET_a \cap SET_b| \quad (1)$$

Sample consistency is thus the number of genes in common between the two sets, and depends on ranking function, data and number of genes selected. When only a single dataset is available, sample consistency can be obtained by iteratively partitioning the data into disjoint halves.

A simple example will provide an intuitive sense of the relationship between consistency and probability of success. Consider a scenario where the total number of genes runs into the thousands, with only a few dozen being truly

FUNCTION	NORMALIZED CONSISTENCY	SELECTION ACCURACY	'Confidence' for FUNCTION vs.:		
			DOM	SAM	TS
Difference of means (DOM)	0.14	0.50	-	0.53	0.85
SAM statistic (SAM)	0.13	0.49	-	-	0.84
t-statistic (TS)	0.06	0.29	-	-	-

Table 1. Results of consistency analysis on leukaemia microarray data.

relevant. Suppose also that some proportion of the genes selected by an algorithm are false positives. Then, provided a good number of these false positives are chosen more-or-less at random from the large pool of irrelevant genes, the variability in their *identities* will tend to be high, compared to the corresponding variation among the relevant genes selected. Hence, the greater the proportion of relevant genes among those selected (i.e. the higher the selection accuracy), the more agreement there will tend to be between result-sets. It can be shown (under quite benign conditions) that the expected value of sample consistency *must be positively correlated* with underlying (and unobservable) selection accuracy. This result that can also be easily verified by simulation.

Probabilistic analysis: Figure 1 shows a *probabilistic graphical model* [3] for sample consistency C . Nodes represent random variables, and edges the statistical dependencies between them. The Figure shows that selection accuracy q (for a given ranking function) depends only on model \mathcal{M} , total number of genes g , number of relevant genes g_1 and number of genes selected s . Sample consistency, in turn, depends only on g , g_1 , s and the numbers of relevant genes selected from two datasets (s_1^a and s_1^b respectively). Space does not permit a full description of the model, but the important point to note is that C depends on model \mathcal{M} only via selection accuracy q : in other words, sample consistency is conditionally independent of the model, given selection accuracy. Since we are interested only in inferring selection accuracy given observed C , we can avoid having to explicitly deal with model \mathcal{M} altogether. This enables us to 'side-step' the problem we mentioned at the beginning of not having sufficient knowledge about the underlying model. Using Bayes' theorem, the posterior density over selection accuracy q , given sample consistency C , is:

$$p(q | C) = \frac{P(C | q)p(q)}{\int_0^1 P(C | q)p(q) dq} \quad (2)$$

Suitable priors are chosen for q and g_1 and the likelihood term $P(C | q)$ computed by marginalizing over the discrete random variables s_1^a , s_1^b and g_1 .

3. Results

We applied consistency analysis to a widely-studied microarray dataset pertaining to leukaemia [1]. Following pre-processing, consistencies were computed for each of the three selection methods for which theoretical results were presented in [2]. The graphical model described above was then used to infer posterior distributions over selection accuracy. Results are summarized in Table 2: the simple 'difference of means' method (this is essentially a fold analysis) has the highest consistency and (MAP estimated) selection accuracy. Our explicitly probabilistic approach means that we can easily compute a level of confidence in any decision made as to whether one algorithm or another is more suitable for the given data. We obtain a confidence score by asking how certain we can be that one method is *more effective* than another: that is, by computing the posterior probability that its selection accuracy is higher. Confidence scores (for each pair of methods) are also shown in Table 1. It is clear that the difference in observed consistencies between difference of means and SAM is insignificant; however, it is equally clear that the t-statistic is badly suited to this data. Looking at these results in light of the theoretical work mentioned above, we conjecture that relevant genes in this dataset may have *higher variances* than irrelevant ones.

In conclusion, this paper has briefly outlined the use of consistency analysis in the assessment and choice of gene selection algorithms for microarray data. Full theoretical details and empirical results will follow; supplementary information is available at www.robots.ox.ac.uk/~sach/research.html.

References

- [1] T. R. Golub *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–37, 1999.
- [2] S. Mukherjee and S. J. Roberts. A Theoretical Analysis of Gene Selection. In *Proceedings of the IEEE Computer Society Bioinformatics Conference*. IEEE Press, 2004. To appear.
- [3] K. Murphy. An introduction to graphical models. Technical report, Intel Research, 2001.
- [4] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.