Graphical Selection of Perturbation Thresholds

Anthony Almudevar[1]

Short Abstract — **Gene perturbation experiments are commonly used in the reconstruction of gene regulatory networks. Given the statistical nature of the data, an important question is the determination of the p-value threshold which will be used to define a perturbation effect. We propose to use graphical methods rather than multiple testing procedures to determine this threshold. In particular, for each possible threshold, the structure of the resulting network can be compared to randomly generated graphs, and the threshold chosen by observing significant deviations from chance. The procedure will be illustrated using simulated networks, as well as on perturbation experiments performed on the yeast genome.**

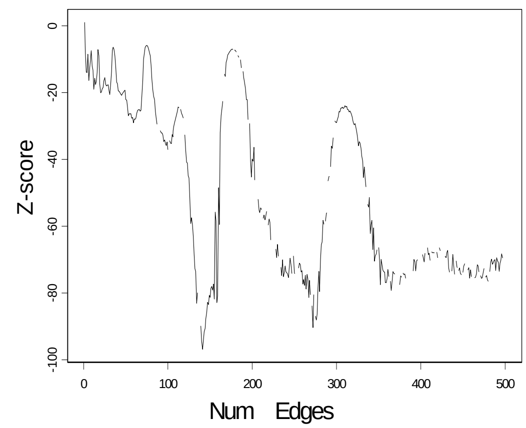
I. Introduction

THE reconstruction of gene regulatory networks using perturbation data has been successfully implemented in recent years [1-3]. One outstanding issue is related to the statistical nature of the data. The determination of a perturbation effect (target gene expression altered by perturbation of control gene) takes the form of a hypothesis test. The acceptance of the p-value as small implies the existence of a directed edge or path (control to target) in the network graph. Network reconstruction thus depends on the selection of a p-value threshold, which may be selected according to principles of multiple hypothesis testing. However, the number of perturbations detected, and hence the complexity of the reconstructed network, will depend as much on sample size as on biological fact.

As an alternative we may accept the ordering of perturbations implied by the p-values, and hence the hierarchical sequence of the resulting networks obtained by varying the p-value threshold. The decision as to which of the networks to accept will be made with reference to the structure of the networks rather than the magnitudes of the p-values.

II. Methodology

Principles of coding theory have been applied to the area of graphical modeling [4], particularly in the context of the Minimum Description Length principle [5] for complex modeling. We propose an algorithm in which for each network of the hierarchical sequence the length of a code required to code the underlying graph is estimated. A reference distribution is then estimated by randomly generating graphs with the same number of edges, estimating the required code length of each. We expect that a gene regulatory network will contain more structure than is explainable by chance [6], resulting in a statistically shorter code length compared to the reference distribution. A threshold may then be selected on the basis of significant deviation from the random graphs.



The procedure will be applied to a selection of simulated networks, as well as to deletion experiments performed on 270 genes from the yeast genome [7] (each possible control-target pair is tested). The above figure shows the number of standard deviations from randomness that was observed by each network in the hierarchical sequence up to 500 edges. The procedure flags the network with 141 edges as being of particular interest.

In this presentation we will show these findings in more detail, and discuss their implications.

Acknowledgements: This work was funded by NIH grant GM075299.

[1]Department of Biostatistics and Computational Biology, University of Rochester, Rochester NYn. E-mail: anthony_anthony@urmc.rochester.edu.

References

[1] Akutsu T, Maruyama O, Kuhara S and Miyano S (1998) A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpression. *The Ninth Workshop on Genome Informatics*, 151-160.

[2] Ideker TE, Thorsson V and Karp R (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pacific Symposium on Biocomputing*, **5**, 302-313.

[3] Wagner A (2004) Reconstructing pathways in large genetic networks from genetic perturbations. *J Comp Biol*, **11**, 53-60.

[4] Friedman N and Goldszmidt M (1998). Learning Bayesian networks with local structure. In *Learning in Graphical Models* (Jordon MI, Ed), MIT Press, Cambridge, MA, pp 412-459.

[5] Rissanen J (1978). Modeling by the shortest data description. *Automatica*, **14**, 465-471.

[6] Wagner A (2002), Estimating coarse network structure from large-scale gene perturbation data. *Genome Research*, **12**, 309-315.

[7] Hughes TR, *et al.* (2000). Functional Discovery via a compendium of expression profiles. *Cell*, **102**, 109-126.