Evolutionary Search for Improved Path Diagrams

Kim Laurio[1], Thomas Svensson[2], Mats Jirstrand[3], Patric Nilsson[1], Jonas Gamalielsson[1], and Björn Olsson[1]

[1] Systems Biology Research Group, University of Skövde, Sweden
[2] Biovitrum AB, Göteborg, Sweden
[3] Fraunhofer-Chalmers Research Center for Industrial Mathematics, Göteborg, Sweden

Abstract. A path diagram relates observed, pairwise, variable correlations to a functional structure which describes the hypothesized causal relations between the variables. Here we combine path diagrams, heuristics and evolutionary search into a system which seeks to improve existing gene regulatory models. Our evaluation shows that once a correct model has been identified it receives a lower prediction error compared to incorrect models, indicating the overall feasibility of this approach. However, with smaller samples the observed correlations gradually become more misleading, and the evolutionary search increasingly converges on suboptimal models. Future work will incorporate publicly available sources of experimentally verified biological facts to computationally suggest model modifications which might improve the model's fitness.

1 Introduction

Several algorithms have been proposed and evaluated for the problem of inferring gene regulatory networks from observations. For an overview, see (Wessels, van Someren et al. 2001). However, the focus has often been on the reconstruction of the entire network from scratch. This work investigates a particular combination of path analysis (Wright 1934) and evolutionary search for comparison and improvement of existing pathway models of gene regulatory networks. Instead of trying to develop algorithms which start from scratch in every new model building effort, we are more interested in developing methods which help us analyze existing models in the light of new data. Our aim is to develop methods which can automatically check if new data rejects a model, or some model features. Such methods could be used for continuously monitoring and updating a database of pathways, and alerting users when new data arrives which contradicts the models they rely on. Another envisioned application is a system that accepts as input a rough model of a set of variables and their relations, and proceeds to refine it based on both existing knowledge and new observations. Current research on biological ontologies and semantic web technology indicates that this is receiving increased attention within the biological sciences (Bodenreider and Stevens 2006).

Regulatory interactions among genes can (to some extent) be represented as models in the form of path diagrams, as exemplified in figure 1. A path diagram is a directed graph which may contain cycles. A path diagram is also a graphical model

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 114 – 121, 2007.

describing a theory about causal relationships among measured and unmeasured variables (Loehlin 1998). Double-headed arrows represent residual correlations between variables and arrows are drawn between variables where one is considered to be a function of the other (Wright 1934). Each arrow (single- or double-headed) has an attached floating-point value - called a path coefficient - which represents the strength of the interaction.

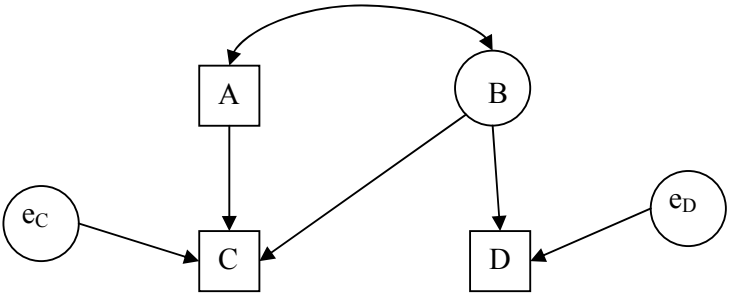

Fig. 1. A path diagram representing the relations between the measured variables *A*, *C* and *D* and the unmeasured variables *B*, e_C and e_D. Here *A* and *B* are correlated due to some unspecified interaction. Variable *A* influences *C* and variable *B* influences both *C* and *D*. Variables *C* and *D* are further influenced by the unspecified and unmeasured, error factors e_C and e_D.

The coefficients define a set of model-implied correlation values between any pair of variables in the diagram, according to a set of rules defined by (Wright 1934). The correlation between any two variables is the sum of the values of all compound paths between them. The value of a compound path is the product of the involved coefficients. A compound path, according to Wright's rules, consists of a sequence of distinct variables, is formed by doing at most one traversal along a double-headed arrow, and includes no backwards traversal after the first forward traversal. In figure 1 the correlation between *C* and *D* is the sum of the compound paths $C \leftarrow A \leftrightarrow B \rightarrow D$ and $C \leftarrow B \rightarrow D$. The coefficient for $B \rightarrow D$ is used in both these compound paths, hence it must be optimized to fit both. For details on path diagram creation, their usage and constraints on their usage, the reader is referred to (Loehlin 1998).

Using path diagrams, alternative theories can be formulated as alternative models, and these can easily be compared to each other as well as to observations on the basis of the model-implied correlations. Given sufficient training data the path coefficients can be optimized to minimize the differences between the observed correlations and the model-implied ones. Our search is using this as a fitness function.

In contrast, (Bay, Shrager et al. 2003) compare a set of model-implied vanishing *partial* correlations to observations. They do it by trying to improve a regulatory network structure by identifying which variables are directly linked, and which are indirectly linked to each other. They further restrict their networks to acyclic graphs. Once their network structure has stabilized they do a second pass to estimate signs for the links between variables. In our approach we do not separate the structure learning and path coefficient estimation into separate stages. The result of the path coefficient estimation is instead used as a hint for better fitting structures during the construction of the next population. Neither is the approach we use restricted to acyclic graphs.