Identifying Valid Solutions for the Inference of Regulatory Networks

Christian Spieth
Centre for Bioinformatics Tübingen (ZBIT)
University of Tübingen
Sand 1, D-72076 Tübingen, Germany
spieth@informatik.uni-tuebingen.de

Felix Streichert
Centre for Bioinformatics Tübingen (ZBIT)
University of Tübingen
Sand 1, D-72076 Tübingen, Germany
streiche@informatik.uni-tuebingen.de

Nora Speer
Centre for Bioinformatics Tübingen (ZBIT)
University of Tübingen
Sand 1, D-72076 Tübingen, Germany
nspeer@informatik.uni-tuebingen.de

Andreas Zell
Centre for Bioinformatics Tübingen (ZBIT)
University of Tübingen
Sand 1, D-72076 Tübingen, Germany
zell@informatik.uni-tuebingen.de

ABSTRACT

In this paper, we address the problem of finding gene regulatory networks from experimental DNA microarray data. The problem often is multi-modal and therefore appropriate optimization strategies become necessary. We propose to use a clustering based niching evolutionary algorithm to maintain diversity in the optimization population to prevent premature convergence and to raise the probability of finding the global optimum by identifying multiple alternative networks than standard algorithms. With this set of alternatives, the identification of the true solution has then to be addressed in a second post-processing step.

Categories and Subject Descriptors: J.3 [LIFE AND MEDICAL SCIENCES]: Biology and genetics

General Terms: Algorithms.

Keywords: Systems Biology, Gene Regulation, Inference, Evolutionary Algorithm.

1. INTRODUCTION

The construction of a network model, which can describe the behavior of the biochemical system, is an important but very difficult task addressed in recent bioinformatics. The purpose of such a gene regulatory network (GRN) is to represent the rules of regulation defining the gene expression. It is regarded as an abstract mapping of the more complicated biochemical network, which includes other components such as proteins, metabolites, etc. The knowledge of the genetic network may then be used as the guidance for further biological experiments to explore higher level of interaction. In literature, several mathematical models can be found that address the problem of analyzing gene regulation. A good overview of related work can be found in [1].

In this work, we suggest to use niching EAs. Using this type of algorithm, we are able to show that niching algorithms reliably meet the requirements of step one by generating a limited set of alternative solutions, which contains the true system with a high probability.

The remainder of this paper is structured as follows. Section 2 describes the proposed algorithm and the mathematical model used in the optimization process. The results are listed in section 3 and the conclusions and an outlook are given in section 4.

GECCO'05, June 25–29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

2. METHOD

There are several approaches to mathematically model a regulatory system. S-Systems (SS) are one possibility, which has been suggested by [4] and can be described by a set of nonlinear differential equations. Another possible model type are linear weight matrices (WM), which originally were introduced by [6].

We compared four different evolutionary algorithms. And the experiment settings of the multi-start algorithms were repeated 20 times to gain sound multi-run statistics.

- 10-start hill climber (10S-HC): multi-start hill-climber with $1,000,000$ fitness evaluations per run.
- 10-start GA (10S-GA): real-value encoding GA, population size 200, tournament selection strategy with tournament group size of 8, 3-point crossover-operator wit $p_c = 1.0$ and $p_m = 0.1$, $5,000$ generations.
- 10-start (5,20)-ES (10S-ES): ES with Covariance Matrix Adaptation (CMA) mutation operator, $p_c = 0.0$ and $p_m = 1.0$.
- Clustering Based Niching ES (CBN-ES) [5]

3. RESULTS

To test the proposed method, we created artificial microarray data sets with a randomly created model with 20 genes. This data set was then reverse engineered by the compared algorithms. Because GRNs are sparse systems in nature, we created regulatory networks randomly with a maximum cardinality of $k \leq 3$.

To evaluate the different algorithms, we counted the number of different solutions found in the inference process and

Table 1: Inference results for a 10-dimensional regulatory net.

Model	Algorithm	Sol,	Hits	Hit ratio
WM	GA	1	4	20.0%
	ES	1	5	25.0%
	10S-HC	4	2	10.0%
	10S-GA	6	6	30.0%
	10S-ES	5	7	35.0%
	CBN-ES	**11**	**12**	**60.0**%
SS	GA	1	3	15.0%
	ES	1	3	15.0%
	10S-HC	7	3	15.0%
	10S-GA	13	5	25.0%
	10S-ES	12	5	25.0%
	CBN-ES	**19**	**11**	**55.0**%

the number of hits, i.e. the number of times the algorithm was able to find the true system. A solution was counted as a hit if the euclidian distance between the parameters of the evolved model and the true system was smaller than a threshold $d_{hit} = 1.0$. Typically, the standard GA and ES return only one solution. In case of the multi-start experiments, we used the same DBScan clustering method [3] as a postprocessing step to identify unique solutions. The CBN-ES on the other hand only proposes converged subpopulations as potential solutions.

The results of the inference process is given in Tab. 1 for each of the model and optimization algorithm, respectively. The CBN outperforms standard GA and ES, the multi-start EA, and the multi-start hill climber. The proposed method identifies the correct solution in around 55% to 60% of the runs. The multi-start EAs are again superior to the standard methods as can be seen from both the hit ratio and the averaged fitness values.

4. DISCUSSION

The problem of inferring GRNs is a very difficult process due to the limited data available and the large number of unknown variables in the system. One of the problems found in the literature is that conventional methods repeatedly run into local optima, thus not necessarily being able to find the optimal solution. Therefore, we introduced an algorithm that increases the probability of finding the correct regulatory network by inferring experimental microarray data in this paper. We showed that standard evolutionary algorithms suffer from the problem of finding solutions within the solution space that comply with the data but do not resemble the original system. The proposed cluster-based niching algorithm efficiently preserve the diversity of network candidates in the optimization process and results in multiple alternative solutions, thus finding the correct network.

We showed that the CBN was able to find better solutions with respect to the fitness than the standard methods independent of the mathematical model used for the simulation of the regulatory network. Further on, the GA based inference algorithm performed slightly better than the one with an ES implementation. This is most likely because GAs have the advantage of a larger initial population size and thus likely having a better coverage of the solution space. However, this issue has to be addressed in future work to be verified. Both standard EAs outperform the naive hill climber due to the multi-modal nature of the solution space. One potential reason for the CBN to be superior to the other algorithms is that if an individual has converged it is then re-initialized and thus more solutions can be found. This halting-window strategy can be implemented for GA and ES as well and will be examined in a future publication. Another possibility to overcome the problem of multi-start EAs to explore the same optima several times is an extension called clearing procedure, which was introduced by Petrowski in [2].

The proposed method yielded more than one valid network solution and although this seems like a disadvantage on the first glimpse, only this enables us to actually find the correct solution in an under-determined and ambiguous environment. To overcome the resulting problem of picking the true solution from the list of valid solutions, one has to use the CBN together with other techniques to clearly identify the overall optimal network model. This postprocessing step can either be done by incorporating additional microarray data sets to decrease the ambiguity in the data or by using a priori known biological or biochemical constraints to eliminate some of the solutions found during optimization. Furthermore, biologists are able to select from the list of the resulting network solutions to exclude some solutions that are not biologically plausible.

Acknowledgement

This work was supported by the National Genome Research Network (NGFN) of the Federal Ministry of Education and Research in Germany under contract number 0313323.

5. REFERENCES

[1] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, January 2002.

[2] A. Petrowski. A clearing procedure as a niching method for genetic algorithms. In *Proceedings of the Congress on Evolutionary Computation (CEC1996)*, pages 798–803, 1996.

[3] J. Sander, M. Ester, H.-P. Kriegel, and X. Xiaowei. Density-based clustering in spatial databases, the algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.

[4] M. A. Savageau. 20 years of S-systems. In E. Voit, editor, *Canonical Nonlinear Modeling. S-systems Approach to Understand Complexity*, pages 1–44, New York, 1991. Van Nostrand Reinhold.

[5] F. Streichert, G. Stein, H. Ulmer, and A. Zell. A clustering based niching ea for multimodal search spaces. In *Proceedings of the 6th International Conference Evolution Artificielle (EA 2003)*, volume 2936 of *LNCS*, pages 293–304. Springer-Verlag, 2003.

[6] D. Weaver, C. Workman, and G. Stormo. Modeling regulatory networks with weight matrices. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 4, pages 112–123, 1999.