Effects of Spatial Growth on Gene Expression Dynamics and on Regulatory Network Reconstruction

Jan T. Kim

School of Computing Sciences,
University of East Anglia, Norwich NR4 7TJ, United Kingdom
jtk@cmp.uea.ac.uk
http://www.cmp.uea.ac.uk/people/jtk

Abstract. Morphogenesis and the spatial structure of an organism have repercussions on gene expression. These effects can influence the results of regulatory network reconstruction. An integrated, flexible and extensible computational framework for modelling gene expression dynamics within spatially growing structures is developed and used as a test system for evaluating a reconstruction algorithm. With complex morphological structures, significant effects of spatial organisation on the reconstruction process are observed. The results also reveal that stronger regulatory interactions result in more frequent cases of indirect regulation, posing a challenge for accurate network reconstruction.

1 Introduction

Regulatory gene networks are a central mechanism of organising and realising complex biological processes and structures based on genetic information. Since initial, now classical models, such as the NK model [1], there has been a steady interest in understanding regulatory networks [2,3,4,5,6,7]. High throughput "post-genomic" techniques, specifically microarrays for measuring gene expression [8,9], currently lead to renewed interest in biological networks [10,11], and various suggestions for reconstructing regulatory networks from gene expression data [12,13,14,15]. However, understanding the relation between regulatory network structure and the resulting gene expression dynamics remains a major challenge [16]. Artificial Life simulations provide a means to advance scientific understanding of gene expression dynamics in biological systems.

Complex spatial structures, which are key features of almost all biological systems. The morphology of an organism is encoded by the genome, from where it is decoded by regulatory networks. Conversely, spatial structures can have a substantial impact on gene expression dynamics. The effects of spatial growth on gene expression have to be expected to be significant for network reconstruction. In this contribution, `transsys` simulations [4] are used to explore the impact of morphogenesis and of other parameters on network reconstruction using the algorithm by Rung *et.al.* [13].

M. Capcarrere et al. (Eds.): ECAL 2005, LNAI 3630, pp. 825–834, 2005.

2 Methods

2.1 Concept

`Transsys` networks were generated as a target of reconstruction. Various types of random network topology, as well as different characteristics of gene regulation were used for network generation. Target networks were integrated into an `L-transsys` Lindenmayer system. Development of this system was simulated for a fixed time interval, after which expression data on all genes of the target network is collected from the grown structure. For each gene in the target network, a knockout mutant was generated and gene expression values were collected. The resulting data set was used as input for regulatory network reconstruction. Reconstruction was evaluated by comparing the reconstructed network to the target network.

2.2 Modelling of Gene Expression in a Spatially Extending System

Generating Target Networks. Target networks were generated as random graphs. The number of nodes (genes) was set to 100 and the number of edges (regulatory interactions) was either set to 200 or to 500. Edges were drawn at random according to the following random network models:

NK graphs [1] are networks in which each of the N genes is regulated by K other genes, chosen at random. Thus, the incoming degree of all genes is K (either 2 or 5), while the outgoing degrees are Poisson distributed.

Random graphs are constructed by choosing each of the possible edge with equal probability. Differently from NK networks, both the incoming degree and the outgoing degree are characterised by a Poisson distribution.

Scale free graphs are characterised by a power law distribution of both incoming and outgoing degrees.

Activating and repressing edges were generated equiprobably. Networks of the same type and with the same edge density have identical topologies in this study.

For all genes, the default level of expression was set to $1.0 + \mathrm{rnd}(0.01)$, where $\mathrm{rnd}(0.01)$ denotes a random value from a uniform distribution over $[0, 0.01[$. A new random value is generated each time expression of the gene is computed. This source of variation is essential for allowing spatial gene expression patterns to form. The default expression level is subject to modification by activation or repression.

Activation and repression are described by two parameters, the maximum amount of regulation $a_{\max}$, and a_{spec}, which is the concentration of the regulator at which activation amounts to $a_{\max}/2$. These parameters were set to the same values for all edges in a network. a_{spec} was set to 0.1 in all simulations whereas $a_{\max}$ was chosen from $\{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$. The parameters for repressing edges, $r_{\max}$ and r_{spec}, were always set to the corresponding activation parameters.

The decay rate of all factors in the target networks was set to the same value. Simulations were run with decay rates of 0.8 and 0.2. For the diffusibility parameter, values of 0.1 and 0.3 were tested.

Fig. 1. Final growth stages for the single shoot structure (left) and the *Arabidopsis thaliana* model (right). Red and blue spheres represent represent meristems, i.e. growth centres from which new morphological elements are generated.

Embedding Networks into `L-transsys` Systems. The process of morphogenesis within which gene expression dynamics controlled by the target network takes place is externally specified in the scenarios studied here. Three morphological structures, depicted in Fig. 1, are used. The cell structure consists of just one symbol, the cell, and no L-system rules. Therefore, no spatially extended structure develops. This serves as a control. The single shoot structure starts out with one meristem, which produces a phytomer consisting of a leaf and an internode (a stem piece) every 20 time steps. The *Arabidopsis* structure is a rather coarse-grained `L-transsys` model of *Arabidopsis* growth, proceeding through stages of rosette leaf growth with decussate and spiral phyllotaxis, bolting, and flower formation.

All three structures are specified by a growth controlling `transsys` program and an `L-transsys` specification. The factors and genes of the target network are inserted into the `transsys` program. The target network does not have any effect on morphogenesis, but growth of the plant structure does have effects on gene expression dynamics. This ensures that all measurements of knockout mutants are based on identical morphological structures.

This approach simulates reconstruction of a target network that does not organise morphogenesis, but may be informed by it. It was chosen here to enable attribution of differences to individual morphological structures, rather than to collections of mutant structures with complex and unfavourable statistical properties. For example, if morphogenesis was controlled by the target network, there may not be any growth in a significant fraction of knockout mutants. Such non-growing mutants would be equivalent to the single cell structure, and consequently, differences between the single cell and the more complex morphological structures would be blurred.

Simulation of Gene Expression Measurement. 504 cases, resulting from combination of 3 structures, 3 random graph types, 2 edge densities, 7 regulatory strength settings, 2 decay settings and 2 diffusibility settings, were assayed. Each structure was grown for 250 time steps, starting out with a single symbol. After

the final time step, the expression level of all factors of the target network, averaged over all symbols, was measured. The resulting vector of gene expression levels corresponds to one microarray experiment in molecular biology.

2.3 Reconstruction of Regulatory Networks

The method for regulatory network reconstruction introduced by Rung *et.al.* [13] uses a set of mutants, called knockout mutants, each of which has one gene disrupted such that it encodes a non-functional gene product. For each knockout mutant, and for the wild type as a control, expression of all genes is measured. The results are assembled into an expression data matrix in which the elements r_{ij} denote the logarithm of the ratio of the expression level of gene i in the mutant with gene j disabled to the expression level of gene i in the wild type. Subsequently, normalised values $\tilde{r}_{ij} = r_{ij}/\hat{\sigma}_{ij}$ are computed, where $\hat{\sigma}_{ij}$ are estimated standard deviations. As measurement is undistorted in the simulations, this step was effectively omitted by setting all $\hat{\sigma}_{ij} = 1$.

The regulatory network graph is then reconstructed by starting with the genes as isolated nodes and placing an edge from gene j to gene i if $|\tilde{r}_{ij}| \geq \gamma$, where γ is a user-supplied threshold. For negative values of $\tilde{r}_{ij}$ (gene i expressed at lower level in absence of gene j), an activating effect is predicted.

2.4 Analysis of Reconstruction

The threshold γ controls the sensitivity and the specificity of the reconstruction algorithm by Rung *et.al.*, low values providing a high sensitivity but low specificity while high threshold settings give good specificity at the expense of a low sensitivity. Since choice of the threshold value is arbitrary in the sense that it is not systematically deduced from the expression data, ROC (Receiver-Operator-Characteristic) curves were used to assess the performance of reconstruction. A ROC curve is computed by reconstructing networks with different threshold settings, ranging from 0 to $\max_{i,j} |\tilde{r}_{ij}|$. For each threshold, sensitivity and specificity are determined. Connecting these points yields the ROC curve. The area under the curve indicates the potential of of the reconstruction procedure. A value of 1 means that perfect reconstruction is possible while a value of 0.5 indicates no potential. By integrating over all possible threshold values, this approach allows assessing the reconstructive potential independently of γ.

To further investigate the reconstruction process, the length of the shortest connecting path, denoted by p_{ij}, was computed for all pairs (i, j) of genes in the target network. For this purpose, no difference between activating and repressing regulatory connections was made, both types were treated as directed edges with length 1. For perfect reconstruction to be possible, there has to exist a γ such that $\forall(i, j) : p_{ij} = 1 \Leftrightarrow \gamma \leq |\tilde{r}_{ij}|$. A scatter plot of all pairs $(p_{ij}, \tilde{r}_{ij})$ reveals whether this condition is satisfied, and provides further insight into the effects which the different variants of network structures and the parameters controlling gene expression dynamics have on the performance of network reconstruction.

2.5 Software

Generation of knockout mutants and collection of gene expression measurements was implemented in `Python`,[1] based on the `transsys` framework [4]. `R` [17] was used for programming data analysis and visualisation. The code underlying the results presented here will be made available on the `transsys` website,[2] which also provides further information on technical aspects of `transsys`.

3 Results and Discussion

3.1 Effects of Spatial Structure

Fig. 2 shows ROC curves for a random graph network, expressed in the single cell and in the *Arabidopsis* structure. A clear difference between reconstruction is observed. In the single cell case, almost perfect specificity is possible up to a sensitivity of 0.95, while a significant decline of specificity is seen with the *Arabidopsis* structure even at sensitivity levels below 0.5.

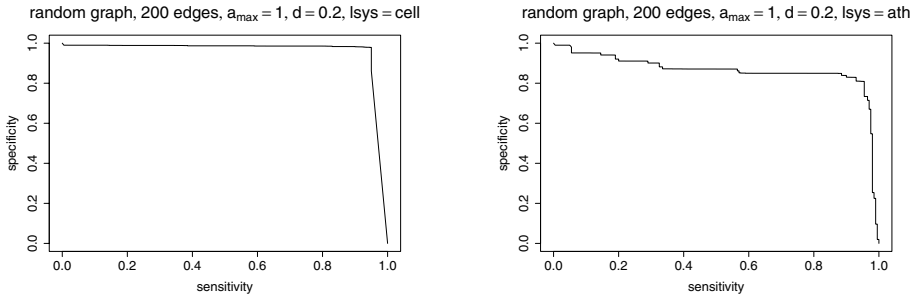

Fig. 2. ROC curves showing the reconstructive potential for a random graph with $N = 100$ genes and 200 edges, regulatory strength $a_{\max} = 1$ and a decay rate of 0.2. Left: results with a single cell, right: results with the *Arabidopsis* structure.

The scatter plot of p_{ij} vs. $\tilde{r}_{ij}$, shown in Fig. 3, reveals that this decline in specificity is due to a substantial increase of variance in the $\tilde{r}_{ij}$ values. There are gene pairs i, j separated by up to 9 network links (p_{ij} up to 9) and $\tilde{r}_{ij} > 2$ observed in the *Arabidopsis* structure, while in the single cell, $-0.297 \leq \tilde{r}_{ij} \leq 0.297$ for all gene pairs with $p_{ij} > 4$. There is a clear trend that effects of a gene knockout on the expression level are more pronounced if the disabled gene is close within the regulatory network, but effects on genes that are distant in the network are possible as well.

While there are substantial differences between reconstruction based on the single cell and the *Arabidopsis* structure, the results obtained with the shoot and the cell structures were not significantly different, as summarised in

[1] `http://www.python.org/`
[2] `http://www.cmp.uea.ac.uk/~jtk/transsys/`

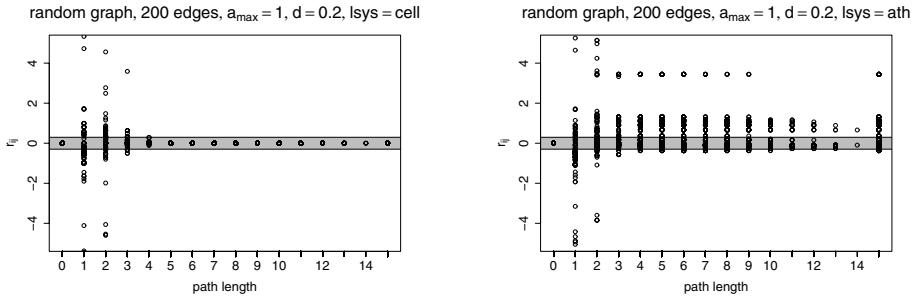

Fig. 3. Scatter plots of path length p_{ij} vs. normalised log ratios $\tilde{r}_{ij}$ for the same target network as in Fig. 2. Left: results with a single cell, right: results with the *Arabidopsis* structure. The gray area shows the range $[-\gamma, \gamma]$ of $\tilde{r}_{ij}$ where no edge is predicted, for $\gamma = 0.297$, the optimal γ value for the *Arabidopsis* structure.

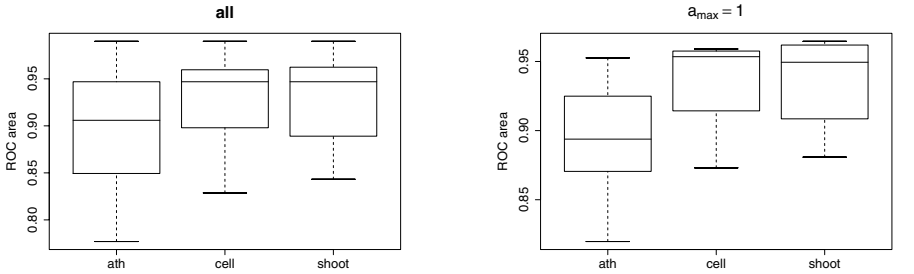

Fig. 4. Box-plots summarising reconstruction performance for the cell, shoot and *Arabidopsis* structures. Boxes show interquartile range of the area under the ROC curve, the horizontal line within each box shows the median value. The differences between the structures are more pronounced with stronger regulatory interactions, as exemplified by $a_{\max} = 1$ (right plot).

Fig. 4. The box-plots show a significantly lower reconstructive potential with the *Arabidopsis* structure, as exemplified by the case discussed above, is generally observed with strong regulatory effects, provided by $a_{\max} = 1$.

Reconstruction performance was generally similar with the single cell and the single shoot structure. This observation indicates that more complex spatial structures have more pronounced effects on regulatory dynamics and network reconstruction.

3.2 Effects of Regulation Strength

Regulation strength has a major impact on reconstruction. The scatter plots shown in Fig. 5 show that with weak regulation with $a_{\max}$ up to 0.4, there are no significant indirect regulatory effects: For all gene pairs with $p_{ij} > 1$, the corresponding value of $\tilde{r}_{ij}$ does not substantially deviate from 0. Consequently, very accurate reconstruction can be achieved.

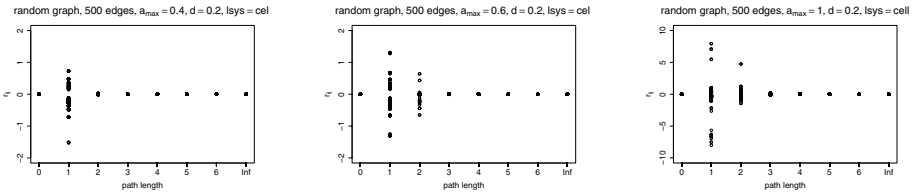

Fig. 5. Scatter plots of path length p_{ij} vs. normalised log ratios $\tilde{r}_{ij}$ for an random graph network with $N = 100$ genes and 500 edges. Left: $a_{\max} = 0.4$, middle: $a_{\max} = 0.6$, right: $a_{\max} = 1$. Notice the different scale of the $\tilde{r}_{ij}$ axis in the right plot.

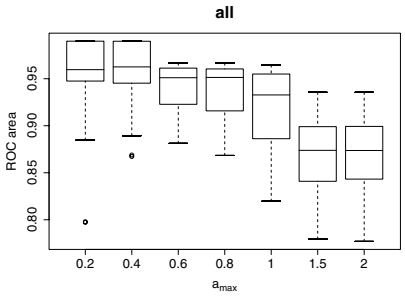

Fig. 6. Box-plot summarising reconstruction performance as a function of regulatory strength $a_{\max}$

With $a_{\max} = 0.6$, significant indirect regulatory effects occur, making perfect reconstruction impossible, as the algorithm by Rung *et.al.* cannot distinguish direct from indirect effects. With $a_{\max} = 1$, regulatory effects increase by an order of magnitude and more overlap between direct and indirect effects that result in $\gamma \leq \tilde{r}_{ij}$ results. Overall, this results in a decline of reconstruction potential as $a_{\max}$, and hence the extent of indirect regulation, increases, as summarised in Fig. 6.

3.3 Effects of Network Structure

The effects of network structure on reconstruction are summarised by the exploits in Fig. 7. This analysis was restricted to the samples with stronger regulation ($a_{\max} \geq 1$) because this reveals effects that are obscured by the large number of cases of near perfect reconstruction if the entire data set is included.

For the graphs with 200 edges, the median reconstruction potential achieved for the three networks exhibits different levels of variance, but the median reconstruction potential is very similar. In contrast to this, the networks with 500 edges result in different reconstruction potentials; the random graph can be reconstructed significantly better than the other types.

Considering that only one representative of each type and density has been evaluated, the results presented here are not sufficient for a deeper analysis of the effects caused by network structure. It is, however, interesting to note that

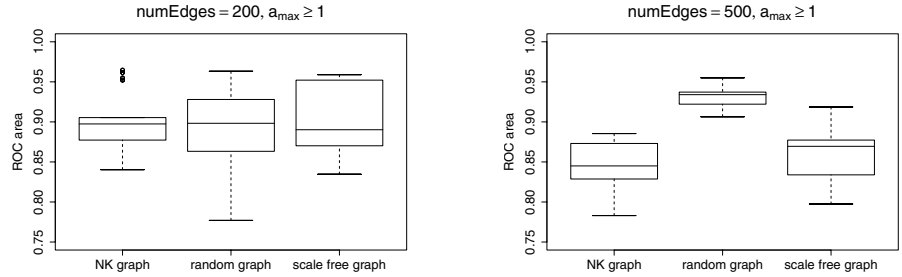

Fig. 7. Box-plots summarising reconstruction performance for sparse (200 edges, left plot) and the dense (500 edges, right) variants different types of networks (NK graph, random graph and scale free graph), for networks with $a_{\max} \geq 1$

effects of network structure on reconstruction performance depend quite strongly on parameters of gene expression such as strength of regulation.

4 Conclusion and Outlook

Artificial Life simulations provide a basis for evaluating methods to reconstruct regulatory networks based on gene expression measurements. Here the `transsys` framework was used to investigate the reconstruction method by Rung *et.al.*

This algorithm assumes that significant changes in expression levels resulting from a gene knockout indicate a direct target gene. The results presented here show that indirect regulation is more frequent in systems with stronger regulatory effects. It would therefore be important to develop criteria for estimating the extent of indirect regulation, and to further develop methods to identify cases of indirect regulation to refine reconstruction by improving specificity.

The results presented here indicate that embedding gene expression within a complex, growing spatial structure results in the formation of patterns that are different from those observed with the same dynamical system within a spatially unstructured environment. This is, in fact, a classical topic in Artificial Life [18,19]. In the interest of focusing on the effects of spatial growth on gene expression dynamic and on network reconstruction, the converse effects of gene expression on morphogenesis have been excluded in this study. This model approximates the case of subnetworks that realise functions other than morphogenesis, but it does not adequately capture networks that organise morphogenesis. This important case will be addressed by studying networks evolved to control morphogenesis. An evolutionary model, based on `LindEvol` [3], is currently being developed for this purpose.

The framework presented here provides points of departure for various further studies. More detailed studies of the impact of decay and diffusion on expression dynamics and network reconstruction are currently underway, and the spectrum of network topologies and dynamical parameters will be further extended. Studies of noise effects will use simulation of measurement errors, as described in [20], and employ the standard deviation of gene expression within the spatial

structures as a measure for biological noise. This will allow to include the normalisation step used by Rung *et.al.* in the evaluation, and make the framework more useful for testing other reconstruction methods.

References

1. Kauffman, S.A.: Developmental logic and its evolution. BioEssays **6** (1987) 82–87
2. Reil, T.: Dynamics of gene expression in an artificial genome – implications for biological and artificial ontogeny. In Floreano, D., Nicoud, J.D., Mondada, F., eds.: Advances in Artificial Life. Lecture Notes in Artificial Intelligence, Berlin Heidelberg, Springer-Verlag (1999) 457–466
3. Kim, J.T.: Lindevol: Artificial models for natural plant evolution. Künstliche Intelligenz (2000) 26–32
4. Kim, J.T.: `transsys`: A generic formalism for modelling regulatory networks in morphogenesis. In Kelemen, J., Sosík, P., eds.: Advances in Artificial Life (Proceedings of the 6th European Conference on Artificial Life). Volume 2159 of Lecture Notes in Artificial Intelligence., Berlin Heidelberg, Springer Verlag (2001) 242–251
5. Banzhaf, W.: On the dynamics of an artificial regulatory network. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J.T., Ziegler, J., eds.: Advances in Artificial Life (ECAL 2003). Volume 2801 of Lecture Notes in Artificial Intelligence., Berlin Heidelberg, Springer Verlag (2003) 217–227
6. Bongard, J.: Evolving modular genetic regulatory networks. In Fogel, D.B., El-Sharkawi, M.A., Yao, X., Greenwood, G., Iba, H., Marrow, P., Shackleton, M., eds.: Proceedings of the IEEE 2002 Congress on Evolutionary Computation (CEC2002), Piscataway, NJ, IEEE Press (2002) 1872–1877
7. Bornholdt, S., Rohlf, T.: Topological evolution of dynamical networks: Global criticality from local dynamics. Physical Review Letters **84** (2000) 6114–6117
8. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science **298** (2002) 799–804
9. Yu, H., Luscombe, N.M., Quian, J., Gerstein, M.: Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends in Genetics **19** (2003) 422–427
10. Bray, D.: Molecular networks: The top-down view. Science **301** (2003) 1864–1865
11. Barabási, A.L., Oltvai, Z.N.: Network biology: Understanding the cell's functional organization. Nature Reviews Genetics **5** (2004) 101–113
12. Akutsu, T., Miyano, S., Kuhara, S.: Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics **16** (2000) 727–734
13. Rung, J., Schlitt, T., Brazma, A., Freivalds, K., Vilo, J.: Building and analysing genome-wide gene disruption networks. Bioinformatics **18** (2002) S202–S210
14. Repsilber, D., Liljenström, H., Andersson, S.G.: Reverse engineering of regulatory networks: Simulation studies on a genetic algorithm approach for ranking hypotheses. BioSystems **66** (2002) 31–41
15. Bongard, J., Lipson, H.: Automating genetic network inference with minimal physical experimentation using coevolution. In ?, ed.: Proceedings of the 2004 Genetic and Evolutionary Computation Conference (GECCO), Berlin Heidelberg, Springer Verlag (2004) 333–345

16. Variano, E.A., McCoy, J.H., Lipson, H.: Networks, dynamics and modularity. Physical Review Letters **92** (2004) 188701
17. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2004) ISBN 3-900051-07-0.
18. Boerlijst, M., Hogeweg, P.: Self-structuring and selection: Spiral waves as a substrate for prebiotic evolution. In Langton, C.G., Taylor, C., Farmer, J.D., Rasmussen, S., eds.: Artificial Life II. Volume X of Santa Fe Institute Studies in the Sciences of Complexity, Proceedings., Redwood City, CA, Addison-Wesley (1992) 255–276
19. Kumar, S., Bentley, P.J., eds.: On Growth, Form and Computers. Elsevier Academic Press, Amsterdam (2003)
20. Repsilber, D., Kim, J.T.: Developing and testing methods for microarray data analysis using an artificial life framework. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J.T., Ziegler, J., eds.: Advances in Artificial Life (ECAL 2003). Volume 2801 of Lecture Notes in Artificial Intelligence., Berlin Heidelberg, Springer Verlag (2003) 686–695