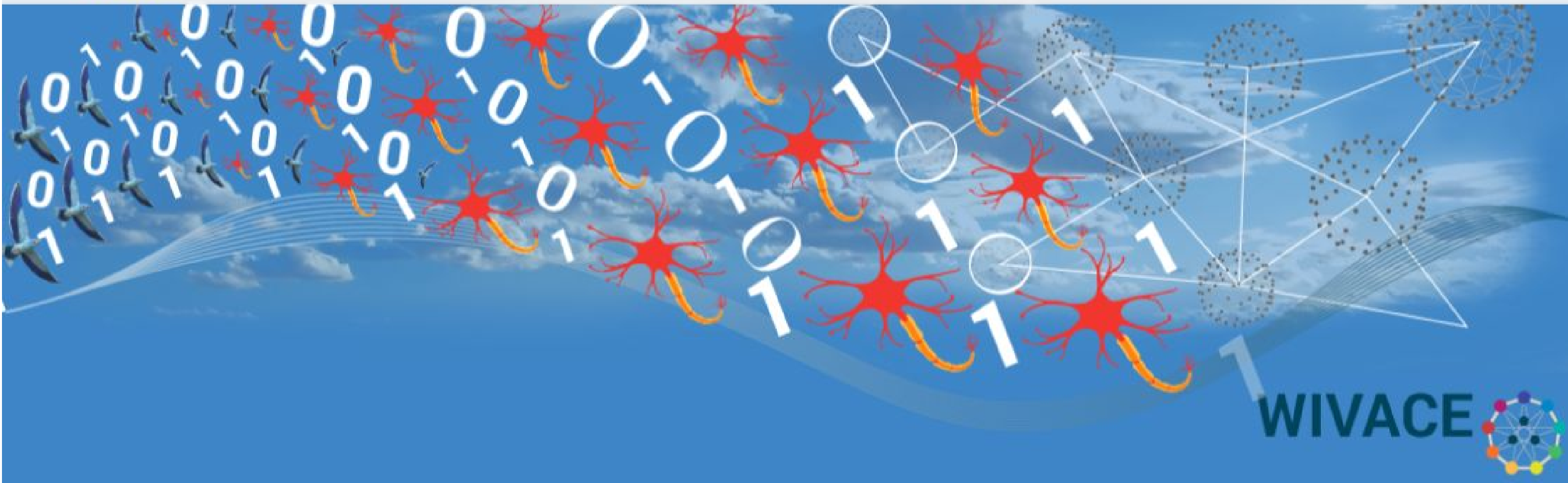
European Centre for Living Technology



WIVACE 2017
Book of Abstracts

Venice, 19-21 September 2017

For info please visit: http://wivace.org/2017

Università Ca'Foscari Venezia
Dipartimento di Scienze Ambientali, Informatica e Statistica

ccis

Springer

Ca'Foscari Zattere
Cultural Flow Zone

WIVACE 2017 - Book of Abstracts
Andrea Roli, Debora Slanzi, Marco Villani eds.

Università degli Studi di Modena e Reggio Emilia
Dip. di Scienze Sociali, Cognitive e Quantitative
Viale Allegri 9, 42121 Reggio Emilia (RE)

ISBN 978-88-903581-3-5 (online)

Index

Harnessing Open Science to Map Protein Fitness Landscapes

Erik Schultes

Dutch Techcentre for Life Sciences, NL
erik.schultes@dtls.nl

Abstract

Proteins exhibit extreme simplicity on one hand (linear chain polymers composed of only 20 amino acids) and extreme diversity on the other (complex molecular folding and manifold biological functions). Along with high throughput sequencing technology and scalable approaches to laboratory analysis, proteins are the ideal model system for investigating fundamental, yet real-world features of evolutionary adaptation. In principle, it is now possible to meaningfully sample and map protein fitness landscapes, although serious issues of data management and data integration must first be addressed. We have developed a Open web-based platform called the Sequenomics Data Exchange (SeDEx) where researchers can store, find, access, and reuse protein sequence-structure-function data. SeDEx users can also easily (automatically) compute protein structure predictions using an ensemble of community-sourced machine learning algorithms in a reusable and reproducible manner. These data can help validate experimentally determined structures and can also provide approximations for the much larger space of unknown (artificial) protein structures. Taken together, the data storage and structure prediction capabilities of the SeDEx provide an Open Science environment where the community can collaborate in mapping the large-scale features of protein fitness landscapes. Here, we present some key features of the SeDEx platform, algorithm performance results, and example experiments showing how protein fitness landscapes can be informatively sampled.

Applied Complex Systems Sciences

Rudolf M. Füchslin

Institute for Applied Mathematics and Physics
Zurich University of Applied Sciences, CH
and
European Centre for Living Technology, IT
rudolf.fuechslin@zhaw.ch

Abstract

Complex systems sciences are today a mature area of research. For many reasons, applications are now a key interest. These reasons range from the direction of European research policies to the fact that addressing real – world problems, among them seemingly highly specific questions arising in the daily operations of SMEs, can shed new light on what we can achieve by taking the vantage point of complex systems sciences and what we could address in the future.

The talk presents different applications, ranging from engineering over topics in health care to questions arising in modeling competing skills in human societies. In this latter example, which is inspired by a problem in anthropology and discussed in more detail, we analyze the dynamics of two skills both solving the same problem; one of them can be quickly learnt but solves the problem only in a limited way (with a small benefit) whereas the other skill is more difficult to learn but yields a larger benefit. We apply the PRESS – method (probability – reduced evolution of spatially resolved species) to study their competition using a mean field theory in a mathematically simple spatially resolved environment, namely an infinite dimensional simplex. As will be shown, such an environment has a quite natural interpretation and, as a function of the mobility of individuals, a surprisingly rich variety of different outcomes of this competition.

A World Beyond Physics

Stuart Alan Kauffman

The University of Pennsylvania, USA

Abstract

The emergence and evolution of life is based on physics but is beyond physics. Evolution is an historical process arising from the non-ergodicity of the universe above the level of atoms. Most complex things will never exist. Human hearts exist. Prebiotic chemistry saw the evolution of many organic molecules in complex reaction networks, and the formation of low energy structures such as membranes. Theory and experiments suggest that from this, the spontaneous emergence of self reproducing molecular systems could arise and evolve. Such "collectively autocatalytic systems" cyclically link non-equilibrium processes whose constrained release of energy constitutes "work" to construct the same constraints on those non-equilibrium processes. Cells yoke a set of non-equilibrium processes and constraints on the energy released as work to build their own constraints and reproduce.

Such systems are living, and can propagate their organization with heritable variations, so can be subject to natural selection. In this evolution, these proto-organisms emerge unprestatably, and afford novel niches enabling, not causing, further types of proto-organisms to emerge. With this, unprestatable new functions arise. The ever-changing phase space of evolution includes these functionalities. Since we cannot prestate these ever new functionalities, we can write no laws of motion for this evolution, which is therefor entailed by no laws at all, and thus not reducible to physics. Beyond entailing law, the evolving biosphere literally constructs itself and is the most complex system we know in the universe.

The brave new world of living and intelligent technologies

Steen Rasmussen

Fundamental Living Technology (FLinT)
Department for Physics & Chemistry, University of Southern Denmark, DK
steen@sdu.dk

Abstract

Our physical technologies (e.g. the bio, info, nano and cogno (BINC) technologies) are rapidly advancing beyond our social technologies (e.g. governance, laws, education, and social norms), which both generate crisis and great opportunities. Our challenges are in part to develop and implement these new technologies for the common good and in part to update our social technologies so that they can address the new physical reality. Humanity has previously been in similar situations e.g. when our societies were transformed from agrarian to industrial. Besides great technological and societal changes, the past Industrial Revolution also brought about new narratives about the world. At the core of our current technological and societal changes lies technologies that mimic life and intelligence, and these technologies are increasingly connected with all of us via the Internet. I'll review key scientific underpinnings both of the technologies and their current societal impacts, and I'll present a framework by which we as scientists can discuss both the scientific and the societal challenges and opportunities including bringing about new and much needed narratives about our rapidly transforming world.

Contrasting views of the origin of human cancers

Roberto Taramelli

Department of Theoretical and Applied Sciences
Università degli Studi dell'Insubria, IT
roberto.taramelli@uninsubria.it

Abstract

During the talk I'll summarize the basic concepts and approaches underlying the most popular theories which are put forward to explain the origin of human malignancies. Although it is widely recognized that human cancers are caused by a plethora of heterogeneous agents ranging from physical radiation, to viruses, inflammation and environment, we currently have only few main interpretation frames to rationalize human cancer development. The most popular frame (Somatic Mutation Theory-SMT) posits that cancer development is ascribed to the step-wise accumulation of genetic mutations in a cell which is acted upon by selective forces driving it toward malignancy including metastatic potential. This explanatory frame is grounded on strict deterministic and linear presuppositions and additionally rests on a strictly cell-centered phenomenon and only partially considers what happens outside the cell. The other theory (Tissue organization Field Theory-TOFT) entails more integrative and organicistic stances and considers cancer growth as an altered/diseased developmental process ("development gone awry" according to his advocates). According to this theory cell to cell abnormal interactions are crucial for cancer to develops and the latter takes place not at the cell level, as implicate by SMT, but at the tissue level of biological organization. In the last decade several important inconsistencies have been delineated within SMT weakening its explanatory power and fostering the development of another conceptual/experimental approach which considers cancers as highly evolved complex adaptive systems and therefore adopt concepts of complexity theory and nonlinear dynamics. Moreover this conceptual approach highlights phenotypic non genetic plasticity of cells as the main determinant of cancer heterogeneity and explains cancer cell behavior regardless of cell's genotype.

Multiple Network Motif Clustering with Genetic Algorithms

Clara Pizzuti[1] and Annalisa Socievole[1]

National Research Council of Italy (CNR), Institute for High Performance Computing
and Networking (ICAR), Via Pietro Bucci, 7-11C, 87036 Rende (CS), Italy
{clara.pizzuti, annalisa.socievole}@icar.cnr.it

Abstract. The definition of community, usually, relies on the concept of edge density. Network motifs, however, have been recognized as fundamental building blocks of networks and, similarly to edges, may give insights for uncovering communities in complex networks. In this work, we propose a novel approach for identifying communities of network motifs. Differently from previous approaches, our method focuses on searching communities where nodes simultaneously participate in several types of motifs. Based on a genetic algorithm, the method finds a number of communities by minimizing the concept of multiple motifs conductance. Simulations on a real-world network show that the proposed algorithm is able to better capture the real modular structure of the network, outperforming both motifs-based and classic community detection algorithms.

Keywords: community detection, network motifs, evolutionary techniques, genetic algorithm

1 Introduction

Complex networks contain small subgraphs named *network motifs* [9], which are pattern of interconnections recurring more frequently than expected in a random network. The frequency of a motif describes the number of times this motif appears within the network. High frequencies of certain motifs are possible due to the important functions they play in a network. For example, the *feed forward loop* and the *bifan* motifs shown in Figure 1(a) and Figure 1(d), respectively, have been found highly frequent into the genetic regulation networks of *E. coli* and *S. Cerevisiae*, as well as into the *C. elegans* neurons network. It is worth noting that multiple motifs usually coexist within a network. Figure 2 shows a subgraph of *Florida Bay food web* network [16], where different microorganisms interact through multiple motifs. We highlight here three types of network motifs: M_5, M_6 (Figure 1(b)) and M_8 (Figure 1(c)). In motif M_5, Water POC serves as energy source for Free Bacteria and Meroplankton, and Meroplankton for Free Bacteria. In motif M_6, Free Bacteria acts also as energy source for Arcatia tonsa, and both nodes are served by Input. Finally, interaction patterns like Input serving Water flagellates and Water ciliates (motif M_8) occur many times.

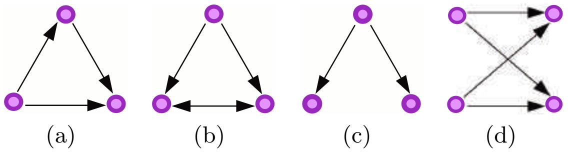


Fig. 1. (a) M_5 (feed-forward loop), (b) M_6, (c) M_8, and (d) M_{bifan} motifs.

Although network motifs have been recognized as "fundamental units of networks" [3], few studies explore the role these subpatterns have in community detection. Arenas et al. [1] show how motifs can be used to define a *motif-based modularity*, i.e. how motif-based modules present more motifs than a random division. Specifically, they extended the original definition of modularity introduced by Girvan and Newman [5] to deal with classes of motifs, and showed that the detected partitions are different with respect to those obtained by optimizing classical modularity. A spectral method based on the generalized modularity has been proposed by the same authors [15], and the differences between the obtained community structures on several networks are highlighted. In a recent work, Benson et al. [2] proposed a tensor spectral clustering method that cluster nodes according to the motif specified in input by the user. First, the higher-order structures involving multiple nodes are encoded by means of tensors (i.e., multidimensional matrices). Then, the method searches a partitioning that does not cut the motifs. Another work [3] of the same authors, described in detail in the next section, extends the concept of *conductance* [14] to network motifs for finding cluster of motifs with low *motif conductance*.

One of the main drawbacks of the aforementioned approaches is that the number of communities must be fixed in advance. In a previous work [12], we proposed *MotifGA*, an evolutionary motifs-based algorithm for community detection using *Genetic Algorithms (GAs)* and a type of motif as input for discovering a number of motif-based communities minimizing motif conductance. Here and in all the previous cited related works, motif-based clustering is thus performed fixing a type of motif and exploring the communities based only on that motif, without considering the coexistence of multiple motifs.

In this paper, we propose *M-MotifGA*, a genetic algorithm for detecting communities in complex networks simultaneously considering different motifs. The method evolves a population of individuals by minimizing the concept of *multiple motif conductance*, and finds a partition of the network into k communities, with k determined by the best local solution optimizing the multiple motif conductance as fitness function. A comparison with the approach of Benson et al. [3], with a variant of this approach we developed here for taking into account multiple motifs, and with the two best known community detection methods *Louvain* [4] and $Infomap$ [13] shows that *M-MotifGA* obtains results better than that found by the other state-of-the-art methods.

The paper is organized as follows. Section 2 introduces the concepts of conductance when motifs are considered and defines the problem we tackle. Section

3 describes our method. Section 4 details the dataset used to perform our experiments and the results obtained. Finally, Section 5 concludes the paper.

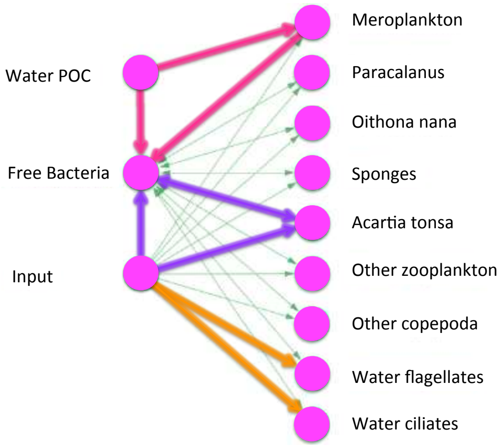


Fig. 2. Multiple motifs coexisting in a subgraph of Florida Bay food web network: M_5, M_6 and M_8. The edges composing these three motifs are highlighted in pink, purple and orange, respectively.

2 Network Motif Clustering

In this section, we start recalling the concepts of *network motif*, *conductance*, *motif conductance* and *multiple motifs conductance*. Then, we describe the method proposed by Benson et al. [3] and the introduction of multiple motifs within their method.

Given a graph $G = (V, E)$ with weights W, $n =| V |$ number of vertices, and $m =| E |$ number of edges, a *motif* of G is defined as an unordered subset $M = \{v_1, \ldots, v_r\}$ of nodes of G presenting a particular pattern of interconnections. Fig. 1 shows three types of motifs among three nodes (Fig. 1(a), 1(b), 1(c)) and a motif involving four nodes (Fig. 1(d)). Their labeling follows the same convention adopted in [3].

Given the diagonal degree matrix D of G defined as $D_{ii} = \sum_{j=1}^{n} W_{ij}$, and a set $S \subset V$ of nodes, the *cut* of S, denoted *cut*(S), is defined as the sum of edge weights having one endpoint in S and the other in $\overline{S} = V - S$:

$$cut(S) = \sum_{i \in S, j \in \overline{S}} W_{ij} \tag{1}$$

The *conductance* of S is defined as

$$\phi(S) = \frac{cut(S)}{min(vol(S), vol(\overline{S}))} \tag{2}$$

where $vol(S) = \sum_{i \in S} D_{ii}$ is the weighted sum of edge end points in S.

By substituting an edge with a motif instance of M, the conductance of S can be generalized to motifs as follows

$$\phi_M(S) = \frac{cut_M(S)}{min(vol_M(S), vol_M(\overline{S}))} \tag{3}$$

where $\phi_M(S)$ is defined *motif conductance*, $cut_M(S)$ is the number of instance motifs of M with at least a node in S and another in $\overline{S}$, and $vol_M(S)$ is the number of instances of M contained in S.

Problem definition (single motif). Fixed a network motif M, find a set of nodes S such that (1) they participate to as many instances of M as possible, and (2) cutting instances of M is avoided.

Benson et al. [3] proposed a method for partitioning V into S and $\overline{S}$ that, given a motif M, minimizes the motif conductance $\phi_M(S)$. The method works on the *motif adjacency matrix* W_M, where each element represents the number of times two nodes appear in an instance of M. When there are nodes that do not participate in any motif, these nodes are discarded from W_M. Then, the eigenvector corresponding to the second smallest eigenvalue of the normalized motif Laplacian matrix is computed. The components of the eigenvector generate an ordering of nodes, which produces nested sets of nodes. The set of nodes with the smallest motif conductance is proven to be a near-optimal partition. Further details on the approach can be found in [3]. For obtaining a partition with more than two communities, the method, named *Motif Recursive bi-partitioning* (*MRbi-part*), can be recursively executed on S and $\overline{S}$, until the desired number of clusters is obtained.

When considering M_1, M_2, ..., M_q motifs simultaneously, the *multiple motifs conductance* is defined as

$$\phi_{MM}(S) = \frac{\sum_{j=1}^{q} \alpha_j cut_{M_j}(S)}{min(\sum_{j=1}^{q} \alpha_j vol_{M_j}(S), \sum_{j=1}^{q} \alpha_j vol_{M_j}(\overline{S}))} \tag{4}$$

where each $\alpha_j \geq 0$ gives a weight to the impact of motif M_j on the network considered.

Problem definition (multiple motifs). Fixed a set of q network motifs M_1, M_2, ..., M_q, find a set of nodes S such that (1) they simultaneously participate to as many instances of all the considered motifs as possible, and (2) cutting instances of any M_j, $j = 1, ..., q$ are avoided.

In the next section, we propose to solve the problem of finding a division on a network based on multiple motifs by applying a Genetic Algorithm (GA) [6]. Specifically, the proposed algorithm minimizes the multiple motifs conductance computed on the motif adjacency matrices of the single motifs, associated with the graph G representing the network. For comparing our method to another method based on multiple motifs, we also modified *MRbi-part* algorithm extending their code with the multiple motifs conductance as measure to minimize. As such, we considered a *weighted motif adjacency matrix* $W_{Mw} = \sum_{j=1}^{q} \alpha_j W_{M_j}$ for running the method. We denominate this extension as *MRbi-part*$_{MM}$. It is worth noting that, differently from the methods of Benson et al., our method does not need the number of communities to find. This number is automatically provided by decoding the solution obtained by the method, i.e. the solution with the lowest local optimum value of conductance.

3 *M-MotifGA* Description

The algorithm we propose, named *M-MotifGA* is based on our previous work [12], where we proposed *MotifGA*, an approach for motif network clustering exploiting a genetic algorithm, that evolves a population of individuals by minimizing motif conductance. Similarly to *MotifGA*, *M-MotifGA* obtains the simultaneous partition of a network into k communities, with k determined by the best local solution optimizing the fitness function. However, differently from *MotifGA*, the fitness function used in *M-MotifGA* is the multiple motifs conductance.

A GA-based method basically evolves a population of individuals initialized at random, and performs variation and selection operators to increase the value of a criterion function, while exploring the search space during the optimization process. *M-MotifGA* uses the locus-based adjacency representation [11] for representing the problem, uniform crossover and neighbor-based mutation for evolving individuals. In the locus-based representation, an individual I is represented as a vector of n genes. Each gene can assume a value j in the range $\{1, \dots, n\}$: when a value j is assigned to the ith, nodes i and j are linked. A decoding step identifies all the connected components of the graph which correspond to the network division in communities. Uniform crossover generates a random binary mask of length equal to the number of nodes, and an offspring is obtained by selecting from the first parent the genes where the mask is a 0, and from the second parent the genes where the mask is a 1. Finally, the mutation operator randomly changes the value j of a i-th gene to one of its neighbors.

M-MotifGA receives in input the graph $G = (V, E)$ and the set of motifs M_1, M_2, ..., M_q, and performs the following steps:

1. compute the motif adjacency matrices W_{M_1}, W_{M_2}, ..., W_{M_q};
2. take the largest connected component $W_{M_j}^{max}$ of W_{M_j} for each motif M_j of the q motifs;
3. obtain the weighted graph $G_{M_j} = (V_{M_j}, E_{M_j})$ corresponding to $W_{M_j}^{max}$ for each $1 \leq j \leq q$;
4. compute the weighted graph $G_M = \sum_{j=1}^{q} G_{M_j}$;
5. run the GA on G_M for a number of iterations by using multiple motif conductance as fitness function to minimize, uniform crossover and neighbor mutation as variation operators;
6. obtain the partition $C = \{C_1, \dots, C_k\}$ corresponding to the solution with the lowest fitness value;
7. merge two communities if the number of inter-cluster connections is higher than the number of intra-cluster connections.

Note that the weighted graphs G_{M_j} associated with the largest connected components of the motif adjacency matrices may have different number of nodes. In this case, before Step 4, the algorithm computes the subset of nodes which are common to all the G_{M_j} graphs. Then, the matrices G_{M_j} will be reorganized in order to contains only the rows and the columns relative to that subset of nodes.

In the next section we present the results obtained by our algorithm and compare them with those returned by state-of-the-art methods. Moreover, we also investigate a variant of our approach, named *MS-MotifGA*, that uses as fitness function the sum the motif conductance of the single motifs, that $\phi_{MS}(S) = \phi_{M_1}(S) + \phi_{M_2}(S) + \ldots + \phi_{M_q}(S)$.

4 Experimental evaluation

To validate our algorithm, we performed several simulations using Matlab 2015b and the Global Optimization Toolbox. Specifically, we compared our algorithm with other well-known state-of-the-art algorithms in terms of NMI, ARI and F1 indexes [8]. The results for *M-MotifGA* have been averaged over 10 runs of the algorithm, setting the population size to 100, the number of generations to 200, the mutation rate to 0.2, and the crossover fraction to 0.8. Moreover, for computing the multiple motif conductance, we equally weighted all motifs using $\alpha_j = 1$. For *MRbi-part* we fixed to 4 the number of communities to find, as suggested by Benson et al. for this dataset, since an higher number of communities would give higher motif conductance values and thus, worse results for their algorithm. Specifically, we applied the motif recursive bipartitioning method twice in order to obtain the desired number of communities. The following subsections describe the dataset used and the algorithms taken into account for testing the effectiveness of *M-MotifGA*.

4.1 Dataset

We analyze the **Florida Bay food web** dataset containing the data of an ecosystem food web. Converting these data in a network graph, nodes can be considered organisms and species, and edges the directed carbon exchange between species. For clustering this network, we consider the motifs M_5, M_6 and M_8 shown in Fig. 1. $M5$, considered a building block for food webs, describes the hierarchical flow of energy between the species i and j which are energy sources for species k, and i is an energy source for both. M_6, on the contrary, models two species that exchange energy and compete to receive energy from a third specie. This motif has been shown to be prevalent within this network, resulting in a rich high-order modular structure. Finally, M_8 corresponds to a single specie feeding two non-interacting species.

The original Florida Bay food web network is composed by 128 nodes and 2106 edges. For detecting communities, we consider a subset of 62 nodes for which the ground truths are known. Specifically, two ground truths, denoted as

$GT1$ and $GT2$, are available and are relative to the two large connected components resulting from the analysis of the adjacency matrix of motif M_6. Table 1 shows the 50 nodes corresponding to the first component and the 12 nodes of the second component. The remaining 66 nodes are isolated. In $GT1$ nodes are classified into 11 different categories (*'demersal producer'*, *'seagrass producer'*, *'algae producer'*, *'microbial microfauna'*, *'zooplankton microfauna'*, *'sediment organism microfauna'*, *'macroinvertebrates'*, *'pelagic fishes'*, *'benthic fishes'*, *'demersal fishes'*, and *'detritus'*). In $GT2$, on the contrary, nodes are categorized into 7 groups: *'producer'*, *'microfauna'*, *'macroinvertebrates'*, *'pelagic fishes'*, *'benthic fishes'*, *'demersal fishes'*, and *'detritus'*. Basically, $GT2$ considers all the *producer* and *microfauna* subcategories of $GT1$ as unique macro categories.

The largest connected components of the adjacency matrices for motifs M_5 and M_8 have 127 and 128 nodes, respectively. Since both motif adjacency matrices contain the 62 nodes for which the ground truths are known, we consider only the submatrices corresponding to this set of 62 nodes when dealing with motifs $M5$ and $M8$.

4.2 Algorithms for community detection

We compare the two strategies of *M-MotifGA*, namely *MM-MotifGA*, when the fitness function used is $\phi_{MM}(S)$, and *MS-MotifGA*, when the fitness function is the sum of the single motif conductances, with the motifs-based *MRbi-part*, both in the case in which this last algorithm uses a single motif to detect communities and in the case of multiple motifs jointly used. We also compare *M-MotifGA* to two benchmark algorithms not using motifs: Louvain [4] and Infomap [13]. Louvain algorithm basically tries to optimize the modularity [10] of a partition through a greed optimization technique. First, small communities are searched by optimizing modularity locally. Then, a new network whose nodes are the communities are built and these steps are repeated until a hierarchy of high-modularity communities is obtained. Infomap, on the contrary, exploits the principles of information theory characterizing the problem of community detection as the problem of finding a description of minimum information of a random walk on the graph. Maximizing the Minimum Description Length [7] objective function, Infomap quickly provides an approximation of the optimal solution.

4.3 Results

Table 2 shows the NMI, ARI and F1 values obtained for the two ground truths of the Florida Bay food web results. For *MM-MotifGA* and *MS-MotifGA* we report both the average value and the standard deviation (in parenthesis) of the evaluation measures. For *MS-MotifGA*, we investigated as fitness function $\phi_{MS}(S) = \phi_{M_5}(S)+\phi_{M_6}(S)+\phi_{M_8}(S)$. On the ground truth $GT1$, *MM-MotifGA* outperforms all the other community detection schemes finding a number of communities ranging from 7 to 10. Similarly, *MS-MotifGA*, finds solutions with 7,

Table 1. Florida Bay food web ground truths (GT1 and GT2) for the two large connected components of the motif $M6$ adjacency matrix.

Node ID	Species	Component	GT1	GT2
8	’Benthic Phytoplankton’	1	Demersal Producer	Producer
9	’Thalassia’	1	Seagrass Producer	Producer
10	’Halodule’	1	Seagrass Producer	Producer
11	’Syringodium’	1	Seagrass Producer	Producer
13	’Drift Algae’	1	Algae Producer	Producer
14	’Epiphytes’	1	Algae Producer	Producer
24	’Benthic Flagellates’	1	Sediment Organism Microfauna	Microfauna
25	’Benthic Ciliates’	1	Sediment Organism Microfauna	Microfauna
26	’Meiofauna’	1	Sediment Organism Microfauna	Microfauna
29	’Other Cnidaridae’	1	Macroinvertebrates	Macroinvertebrates
30	’Echinoderma’	1	Macroinvertebrates	Macroinvertebrates
31	’Bivalves’	1	Macroinvertebrates	Macroinvertebrates
32	’Detritivorous Gastropods’	1	Macroinvertebrates	Macroinvertebrates
34	’Predatory Gastropods’	1	Macroinvertebrates	Macroinvertebrates
35	’Detritivorous Polychaetes’	1	Macroinvertebrates	Macroinvertebrates
36	’Predatory Polychaetes’	1	Macroinvertebrates	Macroinvertebrates
37	’Suspension Feeding Polych’	1	Macroinvertebrates	Macroinvertebrates
38	’Macrobenthos’	1	Macroinvertebrates	Macroinvertebrates
39	’Benthic Crustaceans’	1	Macroinvertebrates	Macroinvertebrates
40	’Detritivorous Amphipods’	1	Macroinvertebrates	Macroinvertebrates
41	’Herbivorous Amphipods’	1	Macroinvertebrates	Macroinvertebrates
42	’Isopods’	1	Macroinvertebrates	Macroinvertebrates
43	’Herbivorous Shrimp’	1	Macroinvertebrates	Macroinvertebrates
44	’Predatory Shrimp’	1	Macroinvertebrates	Macroinvertebrates
45	’Pink Shrimp’	1	Macroinvertebrates	Macroinvertebrates
48	’Detritivorous Crabs’	1	Macroinvertebrates	Macroinvertebrates
49	’Omnivorous Crabs’	1	Macroinvertebrates	Macroinvertebrates
50	’Predatory Crabs’	1	Macroinvertebrates	Macroinvertebrates
51	’Callinectus sapidus’	1	Macroinvertebrates	Macroinvertebrates
57	’Sardines’	1	Pelagic Fishes	Pelagic Fishes
58	’Anchovy’	1	Pelagic Fishes	Pelagic Fishes
59	’Bay Anchovy’	1	Pelagic Fishes	Pelagic Fishes
60	’Lizardfish’	1	Benthic Fishes	Benthic Fishes
61	’Catfish’	1	Benthic Fishes	Benthic Fishes
62	’Eels’	1	Demersal Fishes	Demersal Fishes
63	’Toadfish’	1	Benthic Fishes	Benthic Fishes
64	’Brotalus’	1	Demersal Fishes	Demersal Fishes
65	’Halfbeaks’	1	Pelagic Fishes	Pelagic Fishes
66	’Needlefish’	1	Pelagic Fishes	Pelagic Fishes
68	’Goldspotted killifish’	1	Demersal Fishes	Demersal Fishes
69	’Rainwater killifish’	1	Demersal Fishes	Demersal Fishes
72	’Silverside’	1	Pelagic Fishes	Pelagic Fishes
91	’Mullet’	1	Pelagic Fishes	Pelagic Fishes
93	’Blennies’	1	Benthic Fishes	Benthic Fishes
94	’Code Goby’	1	Benthic Fishes	Benthic Fishes
95	’Clown Goby’	1	Benthic Fishes	Benthic Fishes
96	’Flatfish’	1	Benthic Fishes	Benthic Fishes
99	’Other Pelagic Fishes’	1	Pelagic Fishes	Pelagic Fishes
100	’Omnivorous Ducks’	1	Demersal Fishes	Demersal Fishes
124	’Benthic POC’	1	Detritus	Detritus
15	’Free Bacteria’	2	Microbial Microfauna	Microfauna
16	’Water Flagellates’	2	Microbial Microfauna	Microfauna
17	’Water Cilitaes’	2	Microbial Microfauna	Microfauna
18	’Acartia Tonsa’	2	Zooplankton Microfauna	Microfauna
19	’Oithona nana’	2	Zooplankton Microfauna	Microfauna
20	’Paracalanus’	2	Zooplankton Microfauna	Microfauna
21	’Other Copepoda’	2	Zooplankton Microfauna	Microfauna
22	’Meroplankton’	2	Zooplankton Microfauna	Microfauna
23	’Other Zooplankton’	2	Zooplankton Microfauna	Microfauna
27	’Sponges’	2	Macroinvertebrates	Macroinvertebrates
123	’Water POC’	2	Detritus	Detritus
126	’Input’	2	Detritus	Detritus

Table 2. Florida Bay food web results.

		MM-MotifGA	*MS-MotifGA*	*MRbi-part*$_{M_5}$	*MRbi-part*$_{M_6}$	*MRbi-part*$_{M_8}$	*MRbi-part*$_{MM}$	**Louvain**	**Infomap**
$GT1$	NMI	0.9241 (0.0756)	0.8602 (0.0781)	0.4392	0.504	0.4197	0.3406	0.3879	0.4035
	ARI	0.8451 (0.1923)	0.6879 (0.2141)	0.1388	0.3005	0.1203	0.1291	0.2207	0.1423
	F1	0.8765 (0.1489)	0.754 (0.1646)	0.3149	0.4437	0.2949	0.2962	0.4068	0.31
$GT2$	NMI	0.8367 (0.1127)	0.6844 (0.1054)	0.3214	0.4822	0.3573	0.2829	0.3034	0.3471
	ARI	0.7039 (0.1798)	0.3886 (0.1106)	0.1045	0.3265	0.1332	0.1241	0.2229	0.1592
	F1	0.7756 (0.1329)	0.5549 (0.0727)	0.3087	0.4802	0.3244	0.3101	0.434	0.3416

8, 9 and 10 communities outperforming all the other methods. Considering *MS-MotifGA*, however, we observe that the strategy to sum the single motif conductances as fitness function to optimize, results in NMI, ARI and F1 values lower than *MM-MotifGA*. As such, we conclude that if we explore separately the three motifs and then we recombine them by summing their conductances to obtain the function to optimize, the algorithm does not take into account the intersection there could be between motifs in terms of edges. This intersection, as in the case of motifs M_5 and M_8, and M_6 and M_8 for example, considered when jointly analyzing multiple motifs is able to provide more meaningful communities as the results show. Analyzing all the algorithms by Benson et al. where the number of communities has been set to 4, the communities found result in significantly lover values of the evaluation measures we considered. It is worth noting that considering only M_5, M_6 or M_8 for clustering nodes does not produce satisfying results compared to our multiple-motifs strategies. Moreover, when jointly considering all the motifs as in *MRbi-part*$_{MM}$, the algorithm performs even worse than the single-motif strategies *MRbi-part*$_{M_5}$, *MRbi-part*$_{M_6}$, and *MRbi-part*$_{M_8}$. As such, we conclude that when the number of communities needs to be fixed in input as for *MRbi-part*, detecting clusters of multiple motifs using multiple motif conductance as function to be minimized may lead to suboptimal results. Finally, comparing our method to *Louvain* and $Infomap$ which do not exploit motifs, we observe that also these methods are not able to find a good matching with the ground truth. Focusing on the largest community of the ground truth (i.e., the *macroinvertebrates*) including 21 nodes, for example, we observe that *MM-MotifGA* perfectly matches it in all the runs of the algorithm. *Louvain* distributes the nodes into 4 different communities: 2 groups with 7 nodes are inserted into different communities, 3 nodes into another one, and the remaining nodes into another community. Finally, $Infomap$ inserts all the 21 nodes into a unique but larger community including other nodes.

On the ground truth $GT2$, we obtain similar results. *MM-MotifGA* still outperforms all the other methods resulting in the highest NMI, ARI and F1 values finding solutions with 5 or 6 communities. Overall, for all the algorithms, we observe NMI, ARI and F1 values for $GT2$ are lower than the values obtained for $GT1$. This behavior was also observed in our previous work [12] and it is probably due to the merging of some specie categories done on $GT2$ to create macrocategories which do not perfectly reflect the modular structure of the network.

5 Conclusion

In this paper, we have proposed *M-MotifGA*, a method for discovering communities composed by multiple motifs. Based on a genetic algorithm, our method simultaneously considers different motifs for searching a partition with a number of communities minimizing the multiple motif conductance as fitness function. Simulations on the Florida bay food web network show that *M-MotifGA* results in NMI, F1 and ARI values sensibly higher than both the single-motif and the multiple-motif based analyzed strategies, Louvain and Infomap. Specifically, we have observed that for better matching the underlying real communities, not only multiple motifs should be simultaneously considered, but also fixing the number of communities to obtain as in Benson et al. [3] works, does not fully exploit the benefits of considering multiple motifs. As future work, we plan to extend our experiments to other datasets to further validate our method. We also intend to explore how community detection can be performed when several motifs appear at different network layers in multi-layered network structures.

Bibliography

[1] A. Arenas, A. Fernández, S. Fortunato, and S. Gómez. Motif-based communities in complex network. *Journal of Physics A: Mathematical and Theoretical*, 42(22):224001, 2008.

[2] Austin R. Benson, David F. Gleich, and Jure Leskovec. Tensor spectral clustering for partitioning higher-order network structures. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 118–126, 2015.

[3] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[4] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefevre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008, 2008.

[5] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. In *Proc. National. Academy of Science. USA 99*, pages 7821–7826, 2002.

[6] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[7] Peter D Grünwald, In Jae Myung, and Mark A Pitt. *Advances in minimum description length: Theory and applications.* MIT press, 2005.

[8] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 353(298):824–827, 2002.

[10] Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review*, E69:026113, 2004.

[11] Y.J. Park and M.S. Song. A genetic algorithm for clustering problems. In *Proc. of 3rd Annual Conference on Genetic Algorithms, Morgan Kaufmann Publishers*, pages 2–9, 1989.

[12] Clara Pizzuti and Annalisa Socievole. An evolutionary motifs-based algorithm for community detection. In *Proc. of 8th IEEE International Conference on Information, Intelligences, Systems and Applications (IISA 2017)*, 2017.

[13] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[14] Satu Elisa Schaeffer. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, August 2007.

[15] B. Serrour, A. Arenas, and S. Gómez. Detecting communities of triangles in complex networks using spectral optimization. *Computer Communications*, 34(5), 2011.

[16] RE Ulanowicz, C Bondavalli, and MS Egnotovich. Network analysis of trophic dynamics in south florida ecosystem, fy 97: The florida bay ecosystem. *Annual Report to the United States Geological Service Biological Resources Division Ref. No.[UMCES] CBL*, pages 98–123, 1998.

Algebraic perspectives of solutions spaces in combinatorial optimization

Marco Baioletti[1], Alfredo Milani[1,2], and Valentino Santucci[1]

[1] Department of Mathematics and Computer Science, University of Perugia, Italy
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong
{marco.baioletti,alfredo.milani,valentino.santucci}@unipg.it

1 Introduction

Motivated from the algebraic evolutionary algorithms proposed in [4] and [1], here we introduce novel algebraic perspectives for the search space of a large class of combinatorial optimization problems.

By moving from some simple concepts of group theory, we propose a framework that allows: (i) to use algebraic concepts in order to formally define what is a search move on a discrete space of solutions, (ii) to provide a rationale of the algebraic concepts by means of simple geometric arguments, and (iii) to derive a formal languages point-of-view in order to link algebraic and geometric views, other than to extend the framework to more general search spaces.

The rest of the abstract is organized as follows. Section 2 describes some algebraic background. Abstract algebraic perspectives of combinatorial search spaces are introduced in Section 3, while concrete spaces are depicted in Section 4. Section 5 describes applications of the proposed framework, while conclusion are drawn in Section 6.

2 Algebraic Background

A group $(X, \circ)$ is an algebraic structure consisting of a set of elements X and a binary operation $\circ : X \times X \to X$ satisfying the properties of closure, associativity, existence of a unique neutral element e X, and existence of a unique inverse element x^{-1} for every $x \in X$. If the operation $\circ$ is also commutative, then the group is Abelian. The group $(X, \circ)$ is finitely presented if there exists a presentation (G, R) such that: (i) $G \subseteq X$ is a finite set of elements, called generators, such that every element of X can be expressed as the composition (under the group operation $\circ$) of finitely many elements of G, i.e., for any $x \in X$ it is possible to write $x = g_1 \circ \cdots \circ g_k$ for some $k \in \mathbb{N}_0$ and $g_i \in G$ for $1 \leq i \leq k$, while (ii) R is a finite set of equivalence relations (made using the group operation $\circ$) among the generators in G. Therefore, we can denote a finitely presented group as $(X, \circ, G, R)$ with the properties described.

Geometrically, every finitely presented group $(X, \circ, G, R)$ can be interpreted as an arc-colored digraph $C(X, \circ, G, R)$ such that: (i) the vertices are uniquely identified by the elements of X, (ii) the arc-colors are identified by the generators in G, (iii) for any $x \in X$ and $g \in G$, the arc $x \rightarrow (x \circ g)$ is colored by g, and (iv) any relation in R corresponds to n cycles in the graph (one for each vertex). This graph is called Cayley graph and has the properties of being regular (every vertex has the same degree), strongly connected (for every ordered pair of vertices there is a path connecting them) and vertex-transitive (informally, it is not possible to recognize a vertex by simply looking at the colors of its incoming and outgoing arcs).

Finally, a group can be observed also from the point of view of formal languages. Indeed, the presentation (G, R) of the group $(X, \circ)$ allows to interpret: (i) the generators in G as a set of symbols, (ii) the elements in X as strings over the alphabet G, (iii) the operation $\circ$ as a concatenation of strings, and (iv) the relations in R as rewriting rules for equivalent strings.

3 Combinatorial Search Spaces

Generally, meta-heuristic algorithms for combinatorial optimization iteratively improve a set of candidate solutions (possibly one, as in the case of local search based algorithms) by performing moves from one candidate solution to another. Therefore, the search moves implicitly define neighborhood relations among the candidate solutions in the search space. These moves can be classified in simple and composite moves. Simple moves are the ones usually adopted by the local search based algorithms (e.g., bit-flip moves), while composite moves can be considered as a composition of simple moves (e.g., “jumps” in the space such as those performed by genetic mutation or crossover operators).

Using the concepts described in Section 2, it is possible to define a relationship between a discrete search space and a finitely presented group $(X, \circ, G, R)$ where: (i) the group elements in X are exactly all the candidate solutions in the search space, and (ii) the generators in G, through the group operation $\circ$ and the relations in R, exactly identify the simple search moves. As a consequence, composite search moves can be represented by combining generators, i.e., as strings of generators. Now, since a string of generators is also equivalent to a group element, thus to a candidate solution, we can use the same representation to dichotomously identify both the candidate solutions and the search moves.

Geometrically, the Cayley graph of the group $(X, \circ, G, R)$ connects the candidate solutions by labeling the arcs with the simple search moves needed to reach a solution from another. As a consequence, composite moves can be seen as a path on the Cayley graph. This interpretation allows to naturally define a distance among the candidate solutions, that is, the shortest path distance on the Cayley graph. Moreover, given two solutions $x, y \in X$, all the paths from x to y are represented by sequences of arc-colors. Notably, all these color sequences, by the properties of $\circ$ and the relations in R, evaluate to the same element $y^{-1} \circ x$. This allows to uniquely define the difference between two candidate solutions as $x \ominus y := y^{-1} \circ x$, and, since among all the paths from x to y there are also the shortest paths from x to y, it appears that the group element $y^{-1} \circ x$ can be decomposed to a minimal string of generators with length exactly equal to the distance between x and y. Similarly, other "vectori" operations such as addition and multiplication by scalar can be consistently defined.

Practically, this framework makes possible to represent movements between combinatorial solutions using algebraic or, more generally, linguistic operations.

4 Concrete Search Spaces

Most of the search spaces induced by the representations typically used to handle combinatorial optimization problems can be described by using a finitely presented group $(X, \circ, G, R)$. The main examples are:

- n-length bitstrings where: X is the set of all strings of n bits, $\circ$ is the bitwise *xor* operator, and the generators in G are all the strings with exactly one 1-bit and $n-1$ 0-bit. Therefore, every generator, when composed with a generic bitstring $x \in X$, corresponds to flipping one bit of x. The induced distance is the classical Hamming distance.
- n-length integer vectors where: X is the set of all n-length integer vectors, i.e., $X = \mathbb{Z}^n$, $\circ$ is the integer vector addition, and the generators in G are all the vectors with $n-1$ components equal to 0 and one component equal to ± 1. The induced distance is the classical Manhattan distance.
- permutations of the set $[n] = \{1, \dots, n\}$ where: X is the set of all the permutations of $[n]$, $\circ$ is the usual permutation composition operator, and the generators in G are all the permutations that differ from the identity permutation for one pair of adjacent items that are swapped. Therefore, every generator, when composed with a generic permutation $x \in X$, corresponds to swapping adjacent items of x. In this case, the induced distance is the popular Kendall-τ distance. It worths to note that permutations admit also different generating sets like, for example, those representing items' insertion and interchange moves (see [2]).

In order to build composite move operators, two algorithms are needed: the implementation of the composition operator, and an algorithm that decompose a generic elements in a minimum sequence of generators. In the cases above, it is possible to derive these algorithms. Furthermore, the linguistic perspective also allows to build operative procedures for more general search spaces. The idea is to represent elements directly as strings of generators. Hence, there is no need for decomposers, and composition reduces to string concatenation followed by a minimal rewriting operation (Knuth-Bendix algorithm [3] solves this problem for any finitely presented group).

5 Applications to Evolutionary Computation

The algebraic framework previously discussed has three main applications:

- defining combinatorial variants of many continuous evolutionary algorithms (e.g., differential evolution [4] and particle swarm optimization [1]) for combinatorial search spaces by linking the continuous and combinatorial geometry of the algorithm moves;

- interpreting many existing combinatorial genetic operators using algebraic operators induced by the proposed framework;
- defining new combinatorial operators.

Furthermore, the proposed algebraic perspectives make possible to have a unified view of different combinatorial spaces. Hence, from a theoretical and computational point-of-view, concrete spaces can be studied by analyzing the algebraic properties induced by their finite presentation.

6 Conclusion

Summarizing, this abstract sketched three perspectives of combinatorial search spaces from the point-of-views of algebra, geometry and formal languages. This interpretation applies to a variety of practical search spaces. Using well understood algebraic concepts in literature (such as product groups and rewriting systems) it is also possible to define finitely presented groups for more complex spaces. The main result is the single and dichotomous representation for both candidate solutions and search moves. This builds an analogy with continuous search spaces (numerical vectors in $\mathbb{R}^n$), thus providing, among the others, a mean to generalize meta-heuristic algorithms for continuous optimization to combinatorial problems. As future research avenues, we think that these algebraic perspectives also allow to derive some theoretical analysis of search algorithms for combinatorial spaces.

References

1. Baioletti, M., Milani, A., Santucci, V.: Algebraic particle swarm optimization for the permutations search space. In: Proc. of IEEE Congress on Evolutionary Computation CEC 2017 (in press)
2. Baioletti, M., Milani, A., Santucci, V.: An extension of algebraic differential evolution for the linear ordering problem with cumulative costs. In: Parallel Problem Solving from Nature - PPSN XIV - 14th International Conference, Edinburgh, UK, September 17-21, 2016, Proceedings. pp. 123–133 (2016), http://dx.doi.org/10.1007/978-3-319-45823-6_12
3. Knuth, D.E.: The genesis of attribute grammars. In: Attribute Grammars and Their Applications, pp. 1–12. Springer (1990)
4. Santucci, V., Baioletti, M., Milani, A.: Algebraic differential evolution algorithm for the permutation flowshop scheduling problem with total flowtime criterion. IEEE Transactions on Evolutionary Computation 20(5), 682–694 (2016), http://dx.doi.org/10.1109/TEVC.2015.2507785

Quantum Neural Networks Implementing Deutsch-Jozsa Algorithm

D. Nicolay and T. Carletti

Department of Mathematics and Namur Institute for Complex Systems (naXys), University of Namur, Belgium

Abstract. This paper explores the possibility to construct quantum algorithms thanks to neural networks endowed with quantum gates trained to achieve prescribed goals. First tentatives are performed on the well known Deutsch and Deutsch-Jozsa problems. Preliminary results are promising. In fact, we are able to prove that a solution is detected for different initializations of the problem using a standard evolutionary learning process.

1 Introduction

Quantum computation has generated important interest for the two last decades. In fact, the demand for better performance of computers strongly increases and quantum computation could be the answer to overcome the limitations of current computing. But, even in the case of relatively simple problems, the search for a quantum algorithm is not trivial, this fact is clearly illustrated by the parcelled development of the solution for the well known problem of Deutsch [3].

In this work, we study the possibility to combine tools of classical computation in the quantum framework. Our starting point is the idea to make use of neural networks endowed with quantum gates to develop appropriate quantum algorithms. These neural networks are trained for their specific goals by evolutionary optimization methods, namely genetic algorithms. The task we consider for this preliminary study is the Deutsch problem already mentioned. We show that this methodology leads to promising results, as appropriate solutions are detected for different configurations of the problem.

2 Background to Quantum Computing

The bit is the fundamental unit of classical computation. Quantum computation is developed upon a similar concept, the quantum bit, also called qubit. These qubits have basic states $|0\rangle$ and $|1\rangle$, which correspond to logical states 0 and 1 for classical bits. But, contrary to the latter ones, qubits can also be in a superposition of states $|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle$ where α and β are complex numbers constrained by the normalization condition $|\alpha|^2 + |\beta|^2 = 1$. As qubits are quantum objects, this superposition of states is not observable. Once the qubit is measured, the superposition is lost and the system will be found in the state $|0\rangle$ with probability $|\alpha|^2$ and $|1\rangle$ with probability $|\beta|^2$.

In the same way, we can define systems with n-qubit as $|x_n x_{n-1} \dots x_1\rangle$ where $x_i \in \{0,1\}$ for $i = 1, \dots, n$. Such states can be written as a tensor product of qubits but quantum computation is much more rich. Indeed, thanks to the superposition, a 2-qubit can be in the state $\alpha |00\rangle + \beta |11\rangle$ which can not be constructed using tensor products of qubits. This property of quantum system is called the entanglement and is proper to quantum systems. Quantum gates, working on a qubit or an n-qubit system, are obtained using unitary operators, hence they are reversible and they respect the normalization condition. They are the basic building blocks, combined to form quantum circuits. Among them, we can cite the identity (I) and the NOT (X) operators whose effects are similar to those observed in classical computation. To these ones, we can add the Hadamard transformation (H) defined by $H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$ The most used 2-qubit operator is the controlled-not operator (C), also called the 2-qubit XOR gate, whose effect consists in changing the value of the second qubit if and only if the first one is equal to $|1\rangle$. It has been proved [1] that this 2-qubit gate combined with all qubit gates form an universal set for quantum computation.

3 Deutsch Problem and QNN Model

The Deutsch [3] problem is a basic problem in quantum computing. It consists in deciding if a binary function $f : \{0,1\} \to \{0,1\}$ is constant using only one function evaluation. It is clear that this is not possible in the classical framework, where two function evaluations are needed. To achieve this goal, we have a quantum black box, called oracle, at our disposal. This oracle calculates one of the four possible functions, i.e. forming all the possible couples $f(u) = v$ with $u, v \in \{0,1\}$, by applying an unitary operator U_f defined as $U_f(|x\rangle |y\rangle) = |x\rangle |y \oplus f(x)\rangle$ where $|x\rangle$ and $|y\rangle$ are the states of the system. The quantum circuit representing the solution of this problem is presented in Fig. 1. The sequence of operations described in Fig. 1 leads to the final state $|\psi\rangle$:

$$|\psi\rangle = \begin{cases} \pm |0\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right] & \text{if } f(0) = f(1) \\ \pm |1\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right] & \text{if } f(0) \neq f(1) \end{cases}$$

A measure of the first state is then sufficient to evaluate if the function is constant ($|0\rangle$) or balanced ($|1\rangle$).

Even if this problem is relatively simple, let us remark that finding its solution is not trivial. Indeed, the algorithm originally proposed by Deutsch [3] was probabilistic. It was successful with a probability of one half. In [5], Deutsch and Jozsa developed a deterministic algorithm but it required two oracle calls to succeed. The current solution, with only one function evaluation, has been proposed by Cleve et al. [2]. This shows that even in relatively simple cases, there is a need for a general strategy allowing to construct the algorithm associated to the problem at hand. This is exactly the ultimate goal of our research.

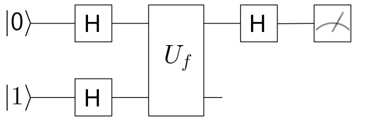


Fig. 1. Quantum circuit for the resolution of the Deutsch problem. The first state is initialized to $|0\rangle$ while the second one is set to $|1\rangle$. Then, an Hadamard gate is applied to the two input states before calling the oracle. An Hadamard gate is finally applied on the first state.

Our model of quantum neural networks is based on the model proposed by Deutsch [4]. The idea consists in building a network whose nodes are quantum gates and connections bring quantum information thanks to qubits. The network is obviously feedforward and the number of nodes is equal to two in every layer. Let us note that, for our trial problem, we do not consider a set of universal gates. The nodes can only be assigned to one of the three qubit gates I, X and H or to the oracle.

Quantum neural networks are evolved to solve the considered problem by a genetic algorithm (GA) [6]. The training environment contains the functions to classify. The fitness of each individual is defined by the fraction of correct classifications. As the optimization is heuristic, all experiments have been replicated 10 times. The results presented are means on these 10 simulations [1].

4 Results and Discussion

The first tests on Deutsch problem have been performed with an initialization of the first state to $|0\rangle$ and the second one to $|1\rangle$. All simulations led to the correct solution. The only difference observed among the different proposed solutions concerns the operator applied on the second state in the last layer, which is not important as only the first state is measured to answer the asked question.

Then, different parameters have been altered to observe the consequences on the learning and the final algorithm. These parameters are the number of layers in the network, the initialization of the states and the value to measure to be constant or balanced. When the number of layers is increased, we observe that a solution is always found even if the number of possible networks increase exponentially. Indeed, the number of admissible solutions also increase exponentially according to the number of layers.

[1] The selection is performed by a roulette wheel selection. The genetic operators are the 1-point crossover and the uniform mutation. Their respective rates are 0.9 and 0.01. The population size is 100 and the maximum number of generations is 5000. The survival of best individuals is ensured by elitism.

If we exchange the initialization of the two states, we have to consider a network of at least 4 layers to find a solution. And, most of the time, the solution consists in replacing the network in the previous initialization, which means that a NOT operator is applied to each state in the first layer. Results are similar if we alter the initialization by setting both states to $|0\rangle$ or $|1\rangle$. If we switch the states to measure to have a constant ($|1\rangle$) or balanced ($|0\rangle$) function, we can find a solution whatever we take as initialization of our states. The smaller network is obtained if both states are initialized to $|1\rangle$. In the other cases, the solution is made of 4 layers. We have also tried to look for a solution if we measure the second state instead of the first one but it has not worked whatever the considered initialization and configuration.

5 Conclusion

Quantum computation provokes considerable interest as it can be an answer to the limitations of current computers. Nevertheless, it remains difficult to elaborate quantum algorithms even for relatively simple problems. Our aim is to study the possibility to develop a general framework based on neural networks endowed with quantum gates and evolutionary computation to tackle this difficulty.

Results are promising as we prove that solutions can be found for different configurations of the Deutsch problem. Similar analyses will be performed on the problem of Deutsch-Josza, which is a generalization of the Deutsch problem. By the following, we could envisage to enlarge our set of gates to make it universal and to use our approach to develop quantum algorithms for more complicated problems. It could also be used to implement quantum operators that interact with n-qubit systems such as the Toffoli gate, which is a generalization of the controlled-not operator.

References

1. Barenco, A., Bennett, C.H., Cleve, R., DiVincenzo, D.P., Margolus, N., Shor, P., Sleator, T., Smolin, J.A., Weinfurter, H.: Elementary gates for quantum computation. Phys. Rev. A 52, 3457–3467 (Nov 1995)
2. Cleve, R., Ekert, A., Macchiavello, C., Mosca, M.: Quantum algorithms revisited. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 454(1969), 339–354 (1998)
3. Deutsch, D.: Quantum theory, the church-turing principle and the universal quantum computer. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 400(1818), 97–117 (1985)
4. Deutsch, D.: Quantum computational networks. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. vol. 425, pp. 73–90. The Royal Society (1989)
5. Deutsch, D., Jozsa, R.: Rapid solution of problems by quantum computation. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 439(1907), 553–558 (1992)
6. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley (1989)

Biological inspired metrics for alignment free sequences analysis

S. Piotto[1], L. Di Biasi[1], L. Sessa[1], P. Iannelli[1] and S. Concilio[2]

[1]Department of Pharmacy, University of Salerno, Via Giovanni Paolo II, 132 - 84084 - Fisciano (SA), Italy
[2]Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132 - 84084 Fisciano (SA), Italy

The analysis of peptide or genome sequences is generally accomplished by pair wise sequence alignment. Several algorithms have been developed to add a proper penalty to any incorrect pairing, and such analysis have been successfully applied to generate phylogenetic relationships. Unfortunately, in spite of numerous experiments, a working grammar of natural antimicrobial peptides is still missing. This failure is due to the short length of peptides along with their large compositional variability. Antimicrobial peptides (AMPs) are generally linear peptides with less than 60 amino acids extracted from a variety of organisms and they are, therefore, extremely variable in composition. AMPs are also closely related to the transmembrane portions of membrane proteins.

This work aimed to provide new methods for predicting AMPs by means of combination of sequence alignment and alignment-free methods with feature selection methods. The methods are based on the definition of "distances" among sets of relevant proteomes. Here we present a successful attempt of mining critical information out from the largest existing collection of antibiotic peptides, Yadamp [1]. Here we describe the metric for sequence distance calculation. The results are structured into 3 parts: 1. Validation of the method. Proteins correlation is compared with phylogenetic analysis performed by means of traditional alignment methods. 2. This method is applied to sets of AMPs. The analysis permitted to establish important correlation between activity and organism specificity, which can be of overwhelming importance to design novel peptides with reduced toxicity. 3. The distance analysis is used to correlate the antimicrobial activity of known sequences with the proteme and transproteome profiles. 4. Finally, the distance results are applied to perform a new rational drug design based on Yada [2], a tool for molecular docking that combine structural information (namely the crystallographic structure of protein receptors) with phylogenetic information.

1. Piotto, S.P., Sessa, L., Concilio, S., Iannelli, P.: YADAMP: yet another database of antimicrobial peptides. International journal of antimicrobial agents **39**(4), 346-351 (2012)

2. Piotto, S., Di Biasi, L., Fino, R., Parisi, R., Sessa, L.: Yada: a novel tool for molecular docking calculations. Journal of computer-aided molecular design **30**(9), 753-759 (2016)

Optimizing the Individuals Maturation for Maximizing the Evolutionary Learning

T. Calenda[1], A. Vitale[2], A. Di Stefano[2], V. Cutello[1] and M. Pavone[1]

[1] Department of Mathematics and Computer Science
University of Catania
V.le A. Doria 6, I-95125 Catania, Italy
{cutello, mpavone}@dmi.unict.it

[2] Department of Electric, Electronics and Computer Science
University of Catania
v.le A. Doria 6, I-95125, Catania

1 Extended Abstract

As it is well known in the natural computing field, one of the major successful factors in evolutionary algorithms is the design and development of the exploration and exploitation mechanisms. A good balancing between these two phases is crucial since it strictly affects the efficiency and robustness of the evolutionary algorithms performances. While the aim of the exploration mechanism is to search for new solutions in new regions by using the mutation operator, the second mechanism has the purpose to exploit in the best possible way all information gathered using the selection process. Both phases, hence, help the algorithm in discovering, gaining and learning new information, and, subsequently, in exploiting all gained promising regions so to generate better populations.

However, what allows to take advantage of the acquired information is truly given by how long each individual lives and in doing so influencing the evolution and maturation of the population. Besides, this lifetime affects, also, the exploration phase, allowing having a better and deep search process. Thus, the time an individual remains in the population becomes crucial in the performances of any evolutionary algorithm, and it is strictly related to the good balancing between the exploration and exploitation processes. Indeed, letting individuals live for a long time produces a dispersive search, and, then, an unfruitful learning, with the final outcome of increasing the probability to easily get trapped in local optima due to the low diversity that is generated. On the other hand, allowing a short lifetime often does not help to have enough overall learning of the knowledge discovered, and it neither allows a careful search within the solutions space, producing instead high diversity into the population, which, in turn, negatively affects the convergence towards a global optimum.

A first research work on this aspect was conducted in [3], where the authors presented an experimental study whose main aim was to understand the right lifetime of any individual/solution in order to perform a proper exploration within the search space, as well as a fair exploitation of the gained information. Such experimental analysis was conducted on an immunological algorithm, whose core components are the cloning, hypermutation and aging operators.

Table 1. Age assignment options.

Type	Symbol	Description
0	$[0 : 0]$	age zero
1	$[0 : \tau_B]$	randomly chosen in the range $[0 : \tau_B]$
2	$[0 : (2/3\ \tau_B)]$	randomly in the range $[0 : (2/3\ \tau_B)]$
3	$[0 : inherited]$	randomly in the range $[0 : inherited]$
4	$[0 : (2/3\ inherited)]$	randomly in the range $[0 : (2/3\ inherited)]$
5	$inherited$ or $[0 : 0]$	inherited; but if constructive mutations occur then type 0
6	$inherited$ or $[0 : \tau_B]$	inherited; but if constructive mutations occur then type 1
7	$inherited$ or $[0 : (2/3\ \tau_B)]$	inherited; but if constructive mutations occur then type 2
8	$inherited$ or $[0 : inherited]$	inherited; but if constructive mutations occur then type 3
9	$inherited$ or $[0 : (2/3\ inherited)]$	inherited; but if constructive mutations occur then type 4
10	$inherited - 1$	same age of parents less one

In the cited research work, eleven different options about the lifetime of each individual were studied (see table 1), with the main goal to answer the three main questions: (i) *"is the lifespan related to the number of offspring generated?"*; (ii) *"is the lifespan related to the population size?"*; and in case of negative answer to the two previous ones, (iii) *"how long must the lifespan of an offspring be to carry out a proper exploration?"*. Once these questions were answered , an efficiency ranking was produced, from which clearly emerged that a too short lifetime (parent's age less 1 - "$type10$" of table 1) is always the worst choice; whilst the best 4 are, respectively: "$type0$;" "$type4$;" "$type3$;" and "$type2$;". Thus, following the above described study, in this research work we want to check if the achievements produced on the immunological algorithm (IA) are still valid, and work, on a genetic algorithms (GA). Of course, what we do not expect to get the same efficiency ranking, but rather we would like to check if the top 4 for IA still appear as the top 4 for GA, even if in different ranking order, and, moreover, if the worst for IA continues to still be the worst for GA.

The tackled Problem: to validate and generalize the obtained results, it is crucial to develop an algorithm which is not tailored to a specific problem, by keeping it unaware of any knowledge about the domain. As it is well-known in literature, to tackle and solve generic and complex combinatorial optimization problems, any evolutionary algorithm must incorporate local search methodologies, used as refinement and improvement of the fitness function, and this means that they have to add knowledge about the features of the problem and application domain. This, consequently, makes the algorithm unsuitable and inapplicable to any other problem. To overcome this limitation and make the outcomes as general as possible, in this study we tackle the classic *One–Max* (or *One–Counting*) problem [5, 2] (as done in [3]). *One–Max* is a well-known toy problem, used to understand the dynamics and searching ability of a stochastic algorithm

[4]. Although it is not of immediate scientific interest, it represents a really useful tool in order to well understand the main features of the algorithm, for example: what is the best tuning of the parameters for a given algorithm; which search operator is more effective in the corresponding search space; how is the convergence speed, or the convergence reliability of a given algorithm; or what variant of the algorithm works better [1]. It is worth emphasizing that a toy problem gives us a *failure bound*, because a failure occurs in toy problems at least as often as it does in more difficult problems. *One-Max* is simply defined as the task to maximize the number of 1 of a bit-string $\boldsymbol{x}$ of length ℓ:

$$f(\boldsymbol{x}) = \sum_{i=1}^{\ell} x_i,$$

with $x_i \in \{0, 1\}$. In order to validate our studies and our outcomes we have set $\ell = 10,000$ in all experiments.

Genetic Algorithm: in this work a classical genetic algorithm has been developed, which is based on the *uniform crossover*, and *flip mutation* operator, where each individual has probability $P_m = 0.4$ to have one randomly chosen bit to be flipped. However, in order to adapt the GA on this experimental study, to each individual/offspring we assigned an age that determines its lifetime into the population until it reaches the maximum age (τ_B), which is determined by a user-defined parameter. In a nutshell, each chromosome is allowed to remain for a number of generations determined by the assigned ageuntil it reaches the value τ_B. Whenever an offspring is generated, it is is assigned a given age, chosen from Table 1, which is incremented by 1 at each generation. Instead, to every individual of the initial population it is always assigned age zero, regardless of the age assignment chosen. It is important to highlight that in general, Crossover and Mutation do not affect the age of any individual, except for the $5 - 9$ options of the age assignment types, where the assigned age is updated only if its fitness value improves after the performance of the genetic operators. Once the maximum age allowed is exceeded for a chromosome, this will be removed from the population via the aging operator, regardless of its fitness function value. However, an exception is allowed only for the best solution found so far (elitist variant). This variant helps the algorithm keep track of the most promising region - which would otherwise be lost - and whose exploitation might be useful in solving some specific kinds of problems. The last operator performed is the selection operator, which identifies the best elements from the offspring set and the old parents, guaranteeing monotonicity in the evolution dynamics. Nevertheless, due to the aging operator, it could happen that the number of individuals which survived (d_s) is less than the input population size (d). In this case, the selection operator randomly generates $d - d_s$ new individuals. The age assignment together with the aging operator, have the purpose to reduce premature convergences, and keep high diversity into the population. It is worth emphasizing again that the choice of what age value to assign plays a central role on the performances of the GA designed (and of any evolutionary algorithm), since, from it, it depends the evolution and maturation of the solutions. What age to assign is then the focus of this research work.

Results and Conclusions: as described above, several experiments have been performed with the main aim to check if the outcomes highlighted and obtained on the immunological algorithm are still valid on a Genetic Algorithm. If so, this allows us to give a rough indication on the needed lifetime to a solution to have a proper balancing between exploration and exploitation in order to maximize the evolutionary learning. For these experiments we took into account the same experimental protocol used in [3], but reducing the string length to $\ell = 2000$. Unfortunately, for higher values than $\ell = 2000$ the GA is not able to reach the optimal solution. All age assignment options in table 1, have been studied varying the population size ($pop\_size = 50, 100$) and $\tau_B = 5, 10, 15, 20, 50, 100, 200$, fixing as termination criteria $Tmax = 10^5$ (maximum number of fitness function evaluations), and, finally, each experiment was computed on 100 independent runs. Furthermore, GA was studied in both its variants: *elitist* and *no_elitist*.

By analyzing all the results it is possible to assert that the worst age assignment types on IA continue to be the worst even on GA, and in particular in the last two positions appear $''type6''$ and $''type10''$ respectively. Likely their bad performances are due to the high diversity they produce, not allowing a relevant lifetime to perform a good exploration. Moreover, we may also assert that the top 4 options produced by IA, in most cases, are still in the top 4 of the efficiency ranking produced by GA, although never in the same order.

References

1. V. Cutello, A. G. De Michele, M. Pavone: "*Escaping Local Optima via Parallelization and Migration*", VI International Workshop on Nature Inspired Cooperative Strategies for Optimization (NICSO), Studies in Computational Intelligence, vol. 512, pp. 141–152, 2013.
2. V. Cutello, G. Narzisi, G. Nicosia, M. Pavone: "*Clonal Selection Algorithms: A Comparative Case Study using Effective Mutation Potentials*", 4th International Conference on Artificial Immune Systems (ICARIS), LNCS 3627, pp. 13–28, 2005.
3. A. Di Stefano, A. Vitale, V. Cutello, M. Pavone: "*Document How long should offspring lifespan be in order to obtain a proper exploration?*", 2016 IEEE Symposium Series on Computational Intelligence (SSCI), INSPEC number 16670548 2016, pp. 1–8, 2016.
4. A. Prugel-Bennett, A. Rogers: "*Modelling Genetic Algorithm Dynamics*", Theoretical Aspects of Evolutionary Computing, pp. 59-85, 2001.
5. J. D. Schaffer, L. J. Eshelman: "*On crossover as an evolutionary viable strategy*", 4th International Conference on Genetic Algorithms, pp. 61–68, 1991.

Evolving Genotype Phenotype Mappings as Dynamical Systems

Jan Paredis

Department of Data Science and Knowledge Engineering, Maastricht University,
P.O. Box 616, 6200 MD Maastricht, The Netherlands
j.paredis@maastrichtuniversity.nl

Abstract. This paper investigates the evolution of Genotype Phenotype Mappings (GPMs). Here, the GPMs are represented as dynamical systems. It is investigated to which regions of a parameterized space of GPMs evolution leads. These regions are called Regions Of Maximum Adaptability (ROMAs). These ROMAs are stable but hard to predict.

Keywords: Coevolution, Dynamics, Genotype Phenotype Mappings

1 Introduction

The Human Genome Project [1] successfully determined the human DNA sequence. This project gave rise to two unexpected findings relevant for the current paper. First, the human genome was shorter than expected. Moreover, a large part was neutral: it did not encode anything. This led to the conclusion that the complexity resides in the expression of the DNA, i.e. the genotype phenotype Mapping (GPM).

Wagner et al. [2] put forward the hypothesis that GPMs are under genetic control and that evolutionary algorithms (EAs) can be used to investigate this. One of the advantages of such an approach is that experiments can be done that are impossible in nature (e.g. because of the lack of control over system parameters). The research proposed here is an instantiation of their proposal in a simple artificial coevolutionary context. The aim of the current research is not to develop biologically plausible models. Nature is far too complex for that. Here, a simple model is used to study the dynamics of GPMs. The idea to model the GPMs as dynamical systems was suggested in [3].

In the current paper, the Coevolutionary Genetic Algorithm (CGA), introduced in [4], is used. This algorithm was inspired by the seminal work of Hillis [5]. In the past, the CGA has mainly been used as a tool for optimalisation, see e.g. [6]. Now, the dynamics of the CGA is studied. The coevolutionary interactions in nature are often complex. The goal of this paper consists of the design of a SIMPLE coevolutionary application and GPM which - despite their simplicity - still exhibits realistic, complex dynamics.

The current paper is a sequel to [7], [8], and [9] in which gradually more complex GPMs are evolved. Moreover, [7] introduced the concept *region of maximum adaptability* (ROMA). It is the region in a parameterized space of GPMs evolution leads to. Actually, sections 2 and 3 of the current paper originate from [8]. Just like in the earlier reseach a pursuer evader (PE) model is used. This model was chosen because it leads to ongoing evolution.

The structure of the paper is as follows. After this introduction the CGA is described. Next, pursuer-evader (PE) sytems are discussed. Section four describes GPMs as dynamical systems followed by a section with empirical results of the evolution of GPMs. Section 6 discusses the relvance of the results. Finally, future research is described and conclusions are drawn.

2 A Coevolutionary Genetic Algorithm

Here, the basic CGA is described, as a first step it creates, two populations (called pop1 and pop2). Typically, the individuals in these initial populations are (uniformly) randomly generated. Next, the fitness of these individuals is calculated. This fitness depends on the particular application, but it is the result of a number - here 10 - of ENCOUNTERS of an individual with individuals of the other population. These encounters result in a pay-off which is stored in the history of the individual. The actual fitness is the average of these (10) history elements. Because these encounters represent predator-prey interactions, success for one individual (in an encounter) is failure for the other one. Hence, the value of an encounter is stored in the history of one individual involved in the encounter. The other individual stores the negative of this value in its history. Once all initial fitnesses are calculated, both populations are sorted on fitness: the individual with the highest fitness on top the least fit one at the bottom.

Next, the main *cycle* of a CGA is executed. The pseudo-code of this cycle is given below. First, 20 encounters are executed between SELECTed individuals. This selection is linearly biased towards highly ranked individuals: similar to GENITOR [10] the top individual is 1.5 times more likely to be selected than the median individual. Next, the pay-off of this encounter is calculated and stored in the history, removing the payoff of the least recent encounter from the history. Hence, the history is implemented as a queue. Finally, the fitness (the average of the history) of both individuals involved in the encounter is re-calculated. Possibly, this changes the ranking of the individual in its population. Note that the predator prey interaction results in a negative pay-off for the individual of the second population.

After these 20 encounters the CGA produces one offspring for each population: it SELECTs two parents. A new individual is generated from these parents through the application of MUTATion (probability of mutating a gene is 0.1) and (uniform) CROSSOVER. The fitness is calculated by executing 10 encounters between the new individual and SELECTed members of the other population (again using the negative payoff for individuals which belong to the second population). In case this fitness is higher than the fitness of the bottom individual then the new individual is placed in the population at its appropriate rank. All individuals with a lower fitness go one position down and the bottom individual is deleted. This basic cycle is repeated a large number of times (e.g. 20000 cy-

```
DO 20 TIMES
ind1 := SELECT(pop1)
ind2 := SELECT(pop2)
payoff := ENCOUNTER(ind1,ind2)
UPDATE-HISTORY-AND-FITNESS(ind1,payoff)
UPDATE-HISTORY-AND-FITNESS(ind2,-payoff)
ENDDO

p1 := SELECT(pop1)        ; pop1 parent1
p2 := SELECT(pop1)        ; pop1 parent2
child := MUTATE-CROSSOVER(p1,p2)
f := FITNESS(child)
INSERT(child,f,pop1)
p1 := SELECT(pop2)        ; pop2  parent1
p2 := SELECT(pop2)        ; pop2  parent2
child := MUTATE-CROSSOVER(p1,p2)
f := FITNESS(child)
INSERT(child,f,pop2)
```

cles). The sampling process to calculate (and update) the fitness is called lifetime fitness evaluation (LTFE). In the current paper, all parameter settings and genetic operators are identical to those described in (Paredis [6]) unless mentioned otherwise.

3 Pursuer-Evader Dynamics

In this particular application, each individual consists of two genes: real numbers in the interval [0,1]. The pay-off of an encounter between two individuals consists of the cartesian distance between the two pairs of genes. The first population maximizes the distance to the individuals of the other population. The negative payoff of the members of the second population results in a minimization of the distance to the individuals of the first population. This because in both populations fitness is maximized.

Each individual can be represented as one point on the plane [0,1] x [0,1]. Furthermore, in order to allow for an unbounded evolution, this plane is considered to be a torus. Hence, the distance is the minimum of the two possible distances (one crossing (an) "edge(s)"). Furthermore, mutation can cross the "edges" as easy as it can move in the plane. Or, in other words, 0.95 is equally likely to be mutated into, for example, 0.085 or 0.05. Finally, a standard uniform crossover is used: new offspring receives each gene from one of its parents randomly and independently.

The dynamics of this application is fairly simple. The initial (random) populations are scattered randomly over the plane. In the first experiment described below equal population sizes consisting of 50 individuals are used. Fairly soon (typically in less than thousand cycles) during evolution two clusters appear (one for each population) where one cluster chases (pursuer) the other (evader). Figure 1 provides a snapshot of such a chase. From time to time different behavior is observed. Sometimes the pursuers catch up on the evaders. At this moment the cluster of evaders breaks up. Most of the time the evader cluster breaks up in two or four sub clusters, which are located symmetrically with respect to the pursuers. These sub clusters virtually immobilize the pursuers while the evader sub clusters move radially and finally become one cluster again. Due to sampling errors and finite population sizes the evaders cluster (i.e. unite) again before the sub clusters have gone all the way. Once the evaders are clustered again, the "standard" pursuing of two clusters continues. Obviously, the symmetrical case may occur as well: the pursuers breaking up to immobilize the evaders for a while, as the snapshot in figure 2 depicts.

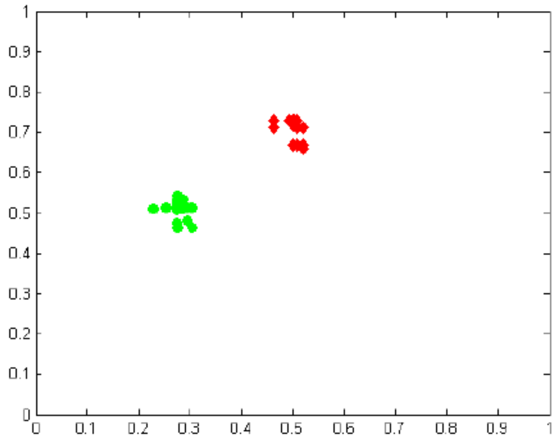


Fig. 1. A cluster of pursuers (black diamonds) pursuing a cluster of evaders (grey cirles).

When the two populations have different population sizes then their respective speed changes. This is because at each cycle both populations reproduce once. Hence, the smaller population evolves the fastest, i.e. moves faster on the plane. In case the pursuer population is smaller, the pursuers regularly catch up with the evaders. When this happens the evaders split up, again immobilizing the pursuers until the evaders form one cluster again. Then the chase resumes. In the other case, the evader population is the smallest population. Here, the evader population successfully keeps ahead of the pursuer population. Occasion-

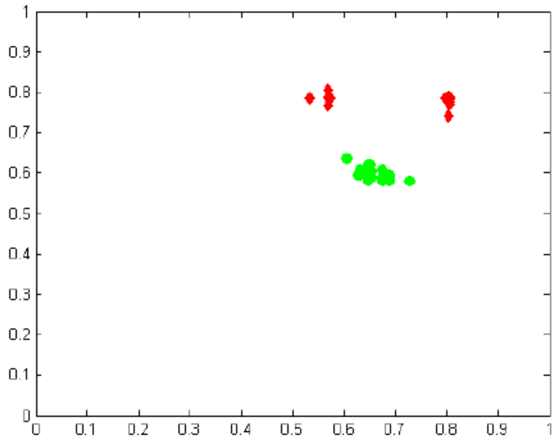


Fig. 2. A cluster of pursuers (black diamonds) splits up in two parts temporarily immobilizing a cluster of evaders (grey cirles).

ally, the evaders even have to slow down in order not to get too close to the pursuers (remember: the world consists of a torus).

4 GPMs as Dynamical Systems

As a starting point, a simple dynamical system with a 2 dimensional phase space is used here to represent a GPM. It contains point attractors. The number of attractors and the size of their basins is under genetic control. The equations below - taken from [11] - describe the dynamics of this system. Figure 3 shows the corresponding phase space.

$$\dot{x} = sin(2 * \pi * x) * cos(2 * \pi * y) \quad (1)$$

$$\dot{y} = sin(2 * \pi * y) * cos(2 * \pi * x) \quad (2)$$

Now, two additionnal parameters - r1 and r2 - are added to transform the dynamics. These r's constitute a parameterized space of GPMs. Both parameters are in the interval [0,1] and are under genetic control: i.e. each individual - evader or pursuer - have these parameters in addition to the x and y positions on the torus as their genetic representation. The equations above are now changed into:

$$\dot{x} = sin(200 * \pi * r1 * x) * cos(200 * \pi * r2 * y) \quad (3)$$

$$\dot{y} = sin(200 * \pi * r2 * y) * cos(200 * \pi * r1 * x) \quad (4)$$

The constant 200 provides un upper bound to the number of attractors.

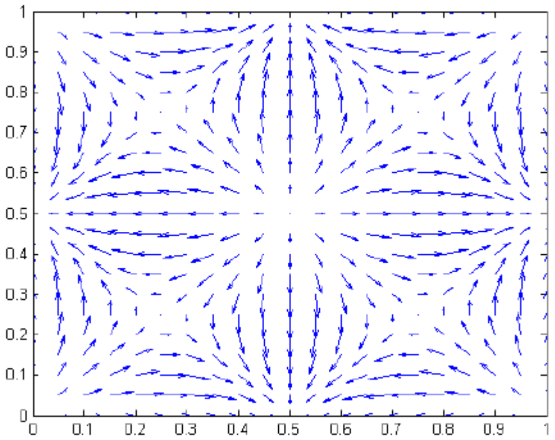


Fig. 3. Phase space diagram for equations (1) and (2).

5 Empirical Results

Now, the algorithm is run 100 times with a population of 50 evaders and 20 pursuers. The notation 50-20 is used for this setup. Earlier research [8] showed that pressure towards the ROMAs is stronger for larger populations. This because the reproduction rate is the same for both populations. Hence the selection pressure on the evaders is largest. For this reason, the distribution of the r's of the evaders will be depicted here, because they are the most outspoken. Figure 4 shows the distribution of the r's of the evader population at the end of each of the 100 runs. Clearly, the r's are drawn towards both axes. So both axes are the ROMAs. It is important to note that it is difficult to understand - let alone predict - what the ROMAs will be. But, on the other hand, these ROMas are stable: repeating the experiments results in the same ROMAs.

Each point in figure 4 represents a dynamical system. Here, we will describe two examples of dynamical systems in the ROMA. Figure 5 represents the system where both r's are equal to 0.02. It consists of 12 equally spaced point attractors. In the second system, shown in figure 6, r1 is equal to 0.02 as well, r2, on the other hand, is equal to 0.98. Hence, both systems are part of the ROMA. A couple of observations are relevant here. Both systems have different dynamics but their geometrical look is similar: the attractors and saddle point are at the same positions.Their type (attractor or saddle point) might change be different. Furthermore, if r1 and r2 are unequal and you switch them then a similar phase space is obtained only all vectors are reversed.

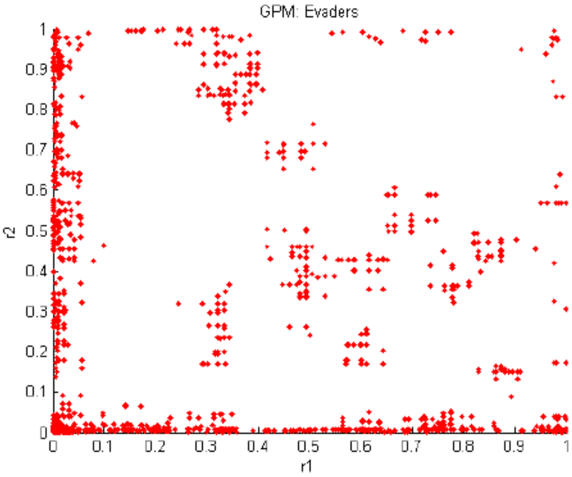


Fig. 4. Distribution of r's of 100 runs of 50-20.

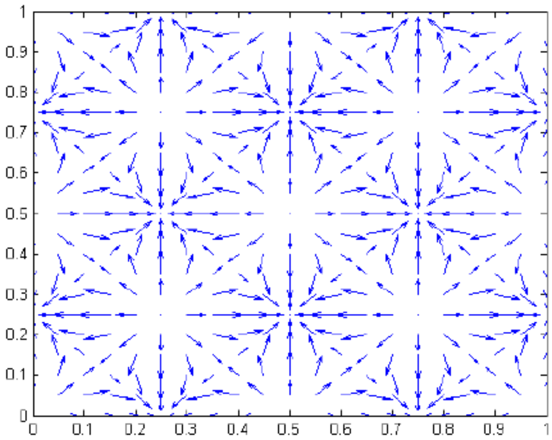


Fig. 5. Phase space of the system described by Eqns. (3) and (4) with r1 = r2 = 0.02.

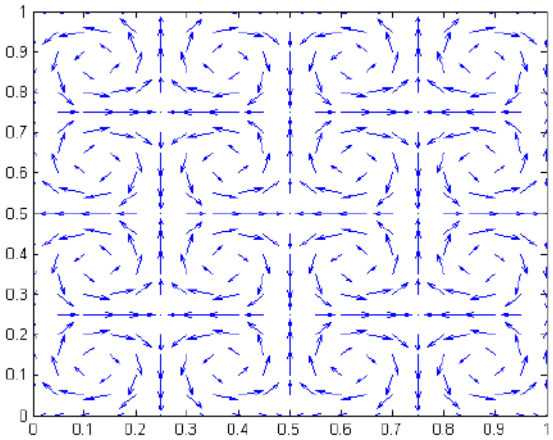


Fig. 6. Phase space of the system described by Eqns. (3) and (4) with r1 = 0.02 and r2 = 0.98.

6 Discussion

The fact that the basic processes (variation, selection, and reproduction with inheritance) are used, the general principle discussed here - increased selection pressure provides a push towards ROMAs - is likely to carry over to nature. Furthermore, the results presented here correspond with [9]. More specific, the many to one mapping, called discretisation, in [9] also provides a push towards the axes.

7 Future Research

A logical follow up is to explore different types of paramterized spaces of dynamical systems. This research will be included in the final version of this paper.

8 Conclusion

This paper investigates the evolution of GPMs as dynamical systems. It confirms the existence of ROMAs - regions in a paramterized space of GPMs - evolution leads to. These ROMAs are stable but hard to predict.

References

1. International Human Genome Sequencing Consortium: Finishing the Euchromatic Sequence of the Human Genome. Nature (431) 7011, pp. 931–945, Nature Publishing Group (2014).

2. Wagner, Gunter P and Altenberg, Lee: Perspective: Complex adaptations and the evolution of evolvability. Evolution, pp. 967–976, JSTOR (1996).
3. Jaeger, J. and Monk, N.: Bioattractors: dynamical systems theory and the evolution of regulatory processes, The Journal of physiology, 592(11), pp.2267–2281, (2014).
4. Paredis, J.: Steps towards co-evolutionary classification neural networks. Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems, pp. 102–108, MIT Press (1994).
5. Hillis, W. D.: Co-evolving Parasites improve Simulated Evolution as an Optimization Procedure. Physica D: Nonlinear Phenomena, (42)1-3, pp. 228–234, Elsevier (1990).
6. Paredis, J.: Coevolutionary Computation, Artificial life, 2(4), pp. 355–375, MIT Press (1995).
7. Paredis, J.: (Co)evolution Leads towards ROMAs, Proceedings of the European Conference on Artificial Intelligence 2014 (ECAI-14), pp. 1079-1080, IOS Press (2014), doi: 10.3233/978-1-61499-419-0-1079.
8. Paredis, J.: Where does (Co)evolution Lead to?, Proceedings of the European Conference on Artificial Life, pp. 138-145, MIT Press (2015).
9. Paredis, J.: Exploring the Evolution of Genotype Phenotype Mappings, Proceedings IEEE Congress on Evolutionay Computation (2017).
10. Whitley, D.: The GENITOR algorithm and selection pressure: Why rank-based allocation of reproductive trials is best, Proceedings of the third International Conference on Genetic Algorithms, vol. 1, pp. 116–121 , Morgan Kaufmann(1989).
11. Jordan, D. W., and Smith, P.: Nonlinear ordinary Differential Equations: Problems and Solutions, Oxford University Press, USA (2007).

Evolving multi-objective optimization in high dimensional systems

Debora Slanzi[1,2], Valentina Mameli[1], Marina Khoroshiltseva[1], and Irene Poli[1,2]

[1] European Centre for Living Technology, S. Marco 2940, 30124 Venice, Italy
[2] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venice, Italy
{debora.slanzi, valentina.mameli, marina.khoroshiltseva, irenpoli}@unive.it

Introduction

Discovering optimal values in high dimensional systems can be a challenging problem, in particular when the number of experimental tests (or observations) is small. Moreover the optimal values can involve different properties of the systems, introducing multiple (and possible conflicting) objective functions to be optimized simultaneously. This framing of the problem can make the search of the optimal values difficult.

In general, a multi-objective optimization problem can be described in the following way:
consider a vector valued objective function $f : C \to \mathbb{R}^k$ from a set $C \subseteq \mathbb{R}^d$ to real numbers $\mathbb{R}^k$, with $f(c) = (f_1(c), \ldots, f_k(c))^T$, where d is the dimension of each element of C; search the element $c_0 \in C$ such that $f(c_0) \leq f(c)$ for all $c \in C$ (minimization) or such that $f(c_0) \geq f(c)$ for all $c \in C$ (maximization).
Frequently, in multi-objective optimization, there does not exist a feasible solution, c_0, which minimizes (or maximizes) all objective functions simultaneously. Therefore, the goal is to achieve Pareto optimal solutions, that is, solutions that cannot be improved in any of the objectives without degrading at least one of the other objectives. In this research we will introduce a methodological approach to address multi-objective optimization in the context above described and related to a molecular system of interest for drug discovery.

1 Evolutionary inference for discovering the system optimal values

In order to develop an efficient approach able to achieve the optimal values of a system with a very small set of experimental tests, we developed a methodological approach based on evolutionary statistical inference for high dimensional experimental spaces and big data analysis. This approach, which we call m-EDO (multi-objective Evolutionary data Design for Optimization), drives the evolution towards the target by estimating and combining predictions from different stochastic models, such as Lasso Regression, Stepwise Regression, Boosting, Neural Networks; see for example [2] and references therein. m-EDO is based on experimental data and is designed to discover the best solution through testing only an extremely small number of candidate solutions, making very efficient and effective the discovery process.

2 Lead optimization in a molecular system

A key problem that the drug discovery research field confronts is to identify small molecules, modulators of protein function, which are likely to be therapeutically useful. Common practices rely on the screening of vast libraries of small molecules (often 1-2 million molecules) in order to identify a molecule, known as a lead molecule, which specifically inhibits or activates the protein function. Such a molecule interacts with the required target, but generally lacks other essential attributes required for a drug candidate. Discovering the optimal lead molecule can then be framed as a multi-objective optimization problem. In this research we address the lead optimization of MMP-12 Inhibitors, using the combinatorial library and biological data made available (public domain) by [1]. This library consists of 2500 molecules characterized by their composition and by their experimental responses. The composition here considered is represented by a set of 22272 fragments, that we describe as binary variables (presence/absence). The high number of fragments give rise to the high dimensionality of the molecular system. For this system the experimental response variables here considered are: *Activity* at the target protein; *Solubility*; *Safety*; *ClogP*; *Molecular Weight*. The aim of this study is to develop a multi-objective optimization procedure based on experimental data (no simulation), and involving a very small number of experimental tests, to avoid waste of research time and resources.
We built m-EDO using the molecular library provided by Pickett et al. (2011) as a source of response variables for selected compositions. We assume that the compositions to test in the lab should be less than 140 (out of the 2500). Knowing the whole experimental space (complete library) allowed us to evaluate the performance of the approach in searching the best response values. These values of the response variables represent the target of our study, and are reported in the following:

- *Activity*, Y_1: the maximum value of Y_1 is **8**, which corresponds to the optimal value. The **99-th percentile** of the response variable distribution is **7.5** (**maximization of** Y_1).
- *Solubility*, Y_2: the maximum value of Y_2 is **-1.766**, which corresponds to the optimal value. The **99-th percentile** of the response variable distribution is **-2.415** (**maximization of** Y_2).
- *Safety*, Y_3: the maximum value of Y_3 is **3.6262**, which corresponds to the optimal value. The **99-th percentile** of the response variable distribution is **3.2309** (**maximization of** Y_3).
- *ClogP*, Y_4: the minimum value of Y_4 is **-2.505**, which corresponds to the optimal value. The **1-th percentile** of the response variable distribution is **0.033** (**minimization of** Y_4).

- *Molecular Weight*, Y_5: the minimum value of Y_5 is **291.3**, which corresponds to the optimal value. The **1-th percentile** of the response variable distribution is **339.3** (**minimization of** Y_5).

The goal of the multi-objective optimization is to discover the three molecules that satisfy the constraints of the problem and reach their best response values. These molecules are represented (in red) in the following Pareto front representation of the molecule *Solubility* and *Safety* after having selected the molecules with an *Activity* greater than 6, a *ClogP* less than 3 and a *Molecular Weight* less than 450.

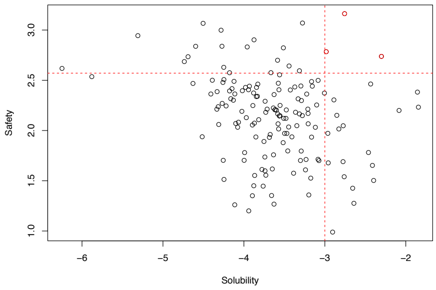


Fig. 1: Pareto front representation of the molecule *Solubility* and *Safety*, respecting the defined constraints for *Activity*, *ClogP* and *Molecular Weight*.

3 The best molecules

We built the EDO approach to optimize the five response variables for the lead optimization process, under the hypothesis to conduct a number of experimental tests less than 140 (on the total of 2500 candidate compositions). At first, to evaluate the performance of the approach, we developed the procedure for each single response variable for a single objective optimization. The evolution in EDO has been driven by the information achieved with the Lasso model, Stepwise regression, the Boosting model, and finally with a mixture of these models. Moreover, in order to evaluate the robustness of EDO we also repeated the procedure 1000 runs.
The results achieved for the single optimization process are represented in the following table:

Notice that EDO procedure is able to achieve the best response values in a very high proportion of 1000 runs, showing also a better performance of the Mixture of models with respect to the single model optimization. Concerning the response values in the region of optimality (1% best values of the distribution)

Objective		Lasso	Stepwise	Boosting	Mixture of Models	NN
Activity	Optimum	844	782	665	916	660
	Region of opt.	1000	995	998	1000	990
Solubility	Optimum	875	745	872	912	556
	Region of opt.	995	998	1000	1000	996
Safety	Optimum	387	358	278	467	228
	Region of opt.	1000	1000	1000	1000	999
ClogP	Optimum	848	821	917	918	760
	Region of opt.	950	946	981	1000	945
Molecular Weight	Optimum	738	822	751	887	346
	Region of opt.	905	966	956	1000	780

Table 1: Single objective optimization: number of runs (out of 1000 runs) in which EDO uncovers the optimum value and values in the region of optimality.

we observe that the Mixture of Models is able to achieve these values in all the 1000 runs and for all the variables.

We then developed the multi-objective optimization by using different approaches for combining the achieved response values and, in comparing the results, we noticed that the simple linear combination of the best values has a very good performance. In the following table we present the results achieved with the Lasso model, the Neural Networks model (hereafter NN) and the Mixture of Models. The three ways to optimize give similar results in discovering the best values, and the difference may lie in discovering just one, or at least one, or all three molecules. Mixture of Models again outperforms the alternatives in discovering at least one molecule of the three in more than 90% of 1000 runs.

Number of best molecules	Lasso	NN	Mixture of Models
0	130	161	92
1	43	59	51
2	320	288	384
3	506	491	472
At least one	869	838	907

Table 2: Multi-objective optimization: number of runs (out of 1000 runs) in which m-EDO uncovers the best molecules.

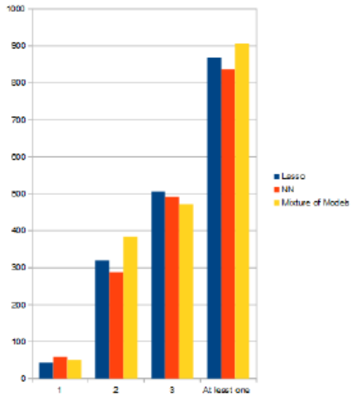


Fig. 2: Multi-objective optimization: best molecules found in 1000 runs.

From these results one can also see the value of the evolution principle in the search process: from the first generation there is a clear tendency for the procedure to converge towards the optimal values.

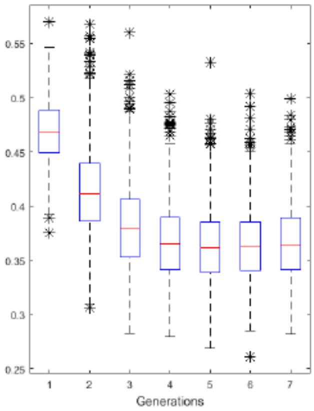


Fig. 3: Evolution through generations: box-plot of the molecule values achieved in 1000 runs at each generation with the Mixture of Models.

Acknowledgements

The authors would like to acknowledge the GlaxoSmithKline Medicines Research Centre (UK) for the very fruitful collaboration in developing this research.

References

1. Pickett, S.D., Green, D.V.S., Hunt, D.L., Pardoe, D.A., Hughes, I.: Automated lead optimization of MMP-12 inhibitors using a genetic algorithm. *A*CS Med. Chem. Lett. **2(1)**, 28–33 (2011).
2. Mameli, Valentina, Lunardon, Nicola, Khoroshiltseva, Marina, Slanzi, Debora, Poli, Irene: Reducing Dimensionality in Molecular Systems: A Bayesian Non-parametric Approach in Federico Rossi; Stefano Piotto; Simona Concilio, Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry, Springer International Publishing, **708**, 114–125 (2017).

Control of signal transduction and communication through model membranes in networks of coupled chemical oscillators

Federico Rossi[1]*, Kristian Torbensen[2], Sandra Ristori[3] and Ali Abou-Hassan[2]

[1] Department of Chemistry and Biology, University of Salerno,
Via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy
frossi@unisa.it
[2] Laboratoire Physico-chimie des Electrolytes, Nanosystemes InterfaciauX (PHENIX),
Sorbonne Universites,UPMC Univ Paris 06,
4 Place Jussieu-Case 51, 75252 Cedex 05 Paris, France
[3] Department of Chemistry & CSGI, University of Florence,
Via della Lastruccia 3, 50019 Sesto Fiorentino (FI), Italy

Synchronization of dynamic elements via chemical communication is a widespread phenomenon in nature and, hence, in many scientific fields, such as in biology, physics, and chemistry, where systems capable of giving and receiving information are commonly found. [1] In these systems, coupling and synchronisation between elements is achieved by messenger molecules diffusing from one element to others that trigger and spread a chemical reaction. In nature, an important example of chemical communication and synchronicity can be found in cell populations, where the plasma membrane governs the traffic of ions or molecules into and out of the cells, thus dictating their collective dynamic. Herein, in a biomimetic approach, we used a microfluidic system to confine a *chemical information generator*, consisting of the far-from-equilibrium Belousov-Zhabotinsky (BZ) reaction, in the aqueous core of monodisperse simple emulsion microdroplets. These microdroplets were surrounded by an oil phase containing the phospholipid 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine, doped with other amphiphilic molecules. Stabilised by the lipids layer contained in the oil phase, the drops could be brought in closest contact to be arranged in a 1D array, using a microfluidic device (see Figure 1). [2–5] The as-formed lipid membrane, at the contact surface, provided a diffusion path between the drops for the chemical species governing the dynamical behaviour of the BZ oscillating reaction, such as the activator $HBrO_2$ or the inhibitors Br_2 or $BrO_2^{\cdot}$. We showed that the coupling mediated by the membranes had mostly an inhibitory character and that the communication could be controlled by the insertion of sodium tetradecyl sulfate and cholesterol as membrane dopants. Numerical simulations suggested that the hydrophobic properties and the lipid packing at the interface were of paramount importance for the transmembrane crossing of the pertinent chemical species.

* Corresponding author

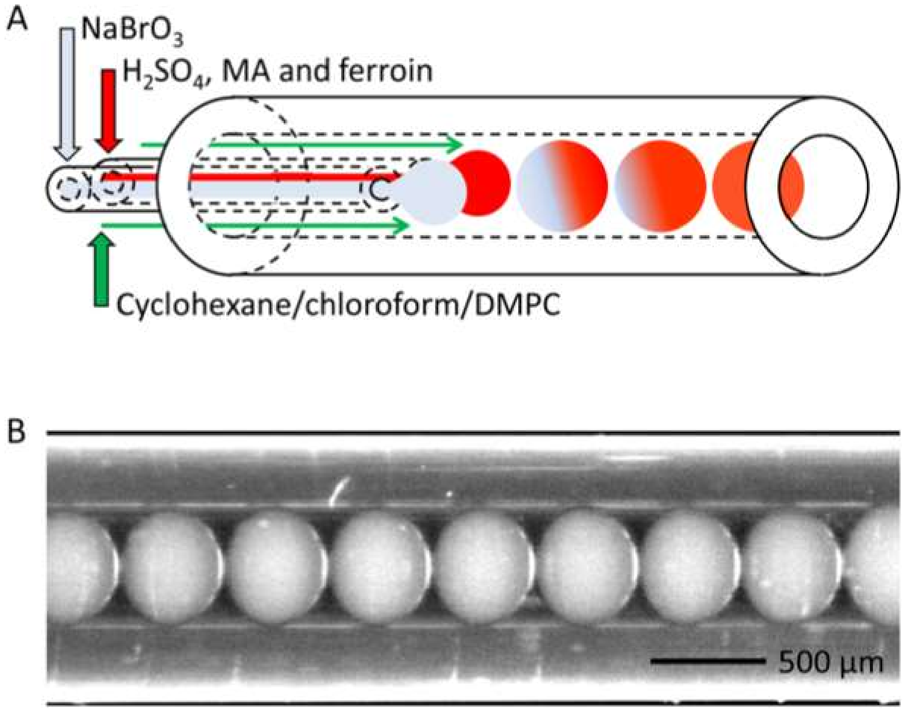


Fig. 1. (A) Sketch of the microfluidic device used to generate the droplet arrays, showing the principal of the coaxial flow of the BZ reactants prior to the drop formation. (B) 1D array of BZ containing droplets as collected in a PTFE tube for monitoring.

References

1. Gentili, P.L.: A Strategy to Face Complexity: The Development of Chemical Artificial Intelligence. In Rossi, F., Piotto, S., Concilio, S., eds.: Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry. Volume 708. Springer International Publishing, Cham (2017) 151–160 DOI: 10.1007/978-3-319-57711-1_13.
2. Tomasi, R., Noel, J.M., Zenati, A., Ristori, S., Rossi, F., Cabuil, V., Kanoufi, F., Abou-Hassan, A.: Chemical communication between liposomes encapsulating a chemical oscillatory reaction. Chemical Science **5**(5) (2014) 1854–1859
3. Rossi, F., Zenati, A., Ristori, S., Noel, J.M., Cabuil, V., Kanoufi, F., Abou-Hassan, A.: Activatory Coupling Among Oscillating Droplets Produced in Microfluidic Based Devices. International Journal of Unconventional Computing **11**(1) (2015) 23–36
4. Torbensen, K., Rossi, F., Ristori, S., Abou-Hassan, A.: Chemical communication and dynamics of droplet emulsions in networks of BelousovâĂŞZhabotinsky micro-oscillators produced by microfluidics. Lab on a Chip **17**(7) (2017) 1179–1189
5. Torbensen, K., Ristori, S., Rossi, F., Abou-Hassan, A.: Tuning the Chemical Communication of Oscillating Microdroplets by Means of Membrane Composition. The Journal of Physical Chemistry C **in press** (2017)

Controlling chemical chaos in the Belousov-Zhabotinsky oscillator

M.A. Budroni[1] M. Rustici[1], N. Marchettini[2] and F. Rossi[3]

[1] Department of Chemistry and Pharmacy, University of Sassari, Via Vienna 2, 07100 Sassari, Italy
[2] Department of Physical Sciences, Earth and Environment, University of Siena, Pian dei Mantellini 44, 53100 Siena, Italy
[3] Department of Chemistry and Biology, University of Salerno, via Giovanni Paolo II 132, 84084 Fisciano, Italy

1 Introduction

Chaos is ubiquitous in Nature and represents one of the most fascinating expressions of real world complexity. Depending on the specific context, the onset of chaotic behaviors can be undesirable and controlling possible mechanisms at the basis of chaotic dynamics is of fundamental interest in many areas, including cardiology [1], information processing [2], hydrodynamics [3] and optics [4], to name a few.

In this work we review our recent results showing how the active interplay between nonlinear a kinetics and hydrodynamic instabilities can be exploited to induce and control chemical chaos.

To this end, we consider as a model system the prototype of chemical oscillators, the Belousov -Zhabotinsky (BZ) reaction. Thanks to a chemo-hydrodynamic coupling, the reaction can undergo chaotic oscillations when carried out in batch conditions (Fig. 1).

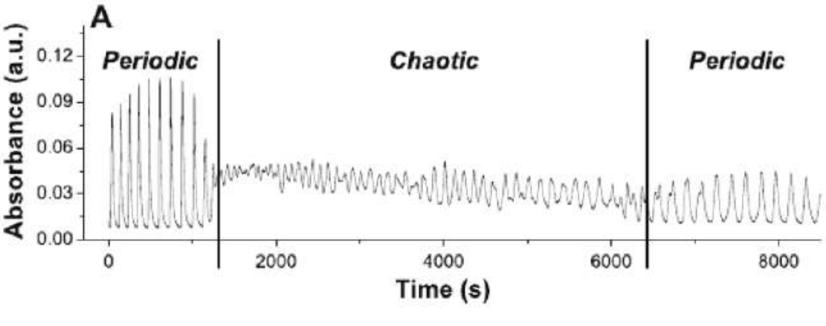


Fig. 1. Transition to chemical chaos of a cerium-catalyzed BZ reaction in a batch and unstirred reactor.

Chaos appears and disappears by following a Ruelle-Takens-Newhouse scenario both in the cerium- and ferroin-catalyzed BZ system [5]. We present experimental evidence that the transition to chemical chaos can be directly controlled by tuning either kinetic or hydrodynamic parameters of the system, namely the reactant concentrations [6], the temperature [7], the reactor geometry [8] and the medium viscosity [9].

We benchmark experimental results by using a theoretical reaction-diffusion-convection (RDC) model in which the nonlinear kinetic functions describing the BZ oscillator are coupled to Navier-Stokes equations, governing the hydrodynamic field in the reactor [10, 11]. Numerical solutions of this RDC problem clearly indicate that natural convection can feedback on the spatio-temporal evolution of the concentration fields and can, in turn, change bulk oscillation patterns. Distinct bifurcations in the oscillation patterns are found when the Grashof numbers (governing the entity of convective flows into the system) and the diffusion coefficients of the chemical species are varied. The consumption of the initial reagents is also found to be a critical phenomenon able to modulate the strength of the RDC coupling and drive order-disorder transitions.

References

1. A. Garfinkel, M.L. Spano, W.L. Ditto, J.N. Weiss, Science, 1992, 257, 1230-1235 (1992).
2. P.L. Gentili, M.S. Giubila, B.M. Heron, PhysChemPhys, 2017, 18, 1831-1841.
3. J. Singer, Y-Z. Wang, H.H. Bau, Phys. Rev. Lett., 1991, 66, 1123.
4. R. Roy, T.W. Murphy, Jr., T.D. Maier, Z. Gills, E.R. Hunt, Phys. Rev. Lett., 1992, 68, 1259.
5. M. Rustici, C. Caravati, E. Petretto, M. Branca, N. Marchettini, J. Phys. Chem. A, 1999, 103, 6564–6570.
6. G. Biosa, M. Masia, N. Marchettini, M. Rustici, Chem. Phys., 2005, 308, 7-12.
7. M. Masia, N. Marchettini, V. Zambrano, M. Rustici, Chem. Phys. Lett., 2001, 341, 285-291.
8. M. L. Turco-Liveri, R. Lombardo, M. Masia, G. Calvaruso, M. Rustici, J. Phys. Chem. A, 2003, 107, 4834-4837.
9. N. Marchettini and M. Rustici, Chem. Phys. Lett., 2000, 317, 647-651.
10. M. A. Budroni, M. Masia, M. Rustici, N. Marchettini, V. Volpert, P. C. Cresto, J. Chem. Phys., 2008, 128, 111102-4.
11. M. A. Budroni, M. Masia, M. Rustici, N. Marchettini, V. Volpert, J. Chem. Phys., 2009, 130, 024902-8.

Fragment based molecular dynamics for drug design

L. Sessa[1], L. Di Biasi[1], P. Iannelli[1] and S. Concilio[2], S. Piotto[1],

[1]Department of Pharmacy, University of Salerno, Via Giovanni Paolo II, 132 - 84084 - Fisciano (SA), Italy
[2]Department of Industrial Engineering, University of Salerno, Via Giovanni Paolo II, 132 - 84084 Fisciano (SA), Italy

Molecular docking is a computational efficient method used to predict the conformations adopted by the ligand within the target binding site. When the binding pocket is known and a molecule is similar to known ligands, the prediction of the complex geometry is quite precise. A positive aspect of conventional docking is the possibility to easily distribute the calculation on dedicated grid or cluster. The main drawback of this procedure is that the receptor is kept rigid, involving only one structure to represent the receptor and overlooking the changes in the binding pocket geometry induced by ligand binding.

Here we present an approach for drug design which achieves fully flexible molecular docking, an accurate energy calculation, and maintaining high computational efficiency.

The workflow starts with a receptor structure optimization bound to a ligand. The ligand is let vary using a combinatorial approach in which the sequential moves consisting of swap, deletion, addition, or molecule cyclization. After each move, the complex ligand-receptor undergoes a Metropolis Monte Carlo run. The step can be then accepted, if the energy of binding ligand-receptor is higher than the previous one, or rejected and the ligand is further modified.

The binding energy of the ligand is calculated using a new binding energy function (YaEnergy) [1]. The whole procedure ensures to steadily increase the binding energy of the designed ligand and permits the exploration of hidden cavity of the receptor. The high computational cost required by this fully flexible procedure is alleviated by a specialized grid (GRIMD) [2] that guarantees a close-to-linear scaling performance.

1. Piotto, S., Di Biasi, L., Fino, R., Parisi, R., Sessa, L., Concilio, S.: Yada: a novel tool for molecular docking calculations. Journal of Computer-Aided Molecular Design **30**(9), 753-759 (2016)

2. Piotto, S., Di Biasi, L., Concilio, S., Castiglione, A., Cattaneo, G.: GRIMD: distributed computing for chemists and biologists. Bioinformation **10**(1), 43 (2014)

A study on complexity measures for the automatic design of robot swarms

Andrea Roli[1]*, Antoine Ligot[2], and Mauro Birattari[2]

[1] Dept. of Computer Science and Engineering
Campus of Cesena
Alma Mater Studiorum Università di Bologna, Cesena, Italy
[2] IRIDIA-CoDE
Université libre de Bruxelles, Brussels, Belgium

The design of control software for robot swarms is a challenging endeavour as swarm behaviour is the outcome of the entangled interplay between the dynamics of the individual robots and the interactions among them. Automatic design techniques are a promising alternative to classic *ad-hoc*, code-and-fix design procedures and are especially suited to deal with the inherent complexity of swarm behaviours. In an automatic method, the design problem is cast into an optimisation problem: the solution space comprises instances of control software and an optimisation algorithm is applied to tune the free parameters of the architecture. Recently, some information theory and complexity theory measures have been proposed for the analysis of the behaviour of single autonomous agents; we believe that a similar approach can be fruitfully applied also to swarms of robots. Indeed, complex systems science may provide a corpus of theories and methods that enable the designer to formally and quantitatively analyse the dynamics of a robot swarm and its internal information processes. Complexity measures may be applied to the automatic design of robot swarms with the following objectives:

1. Understanding individual and swarm behaviour from observations of measurable quantities (e.g. sensor readings, actuation, controller state)
2. Provide task-agnostic merit factors for the automatic design procedures
3. Classify swarm tasks in terms of their intrinsic complexity so as to optimally tune the complexity of individual robot and robot interactions

In this work, we present a preliminary study on the applicability of complexity measures to robot swarm dynamics. The aim of this investigation is to compare and analyse prominent complexity measures when applied to data collected during the time evolution of a robot swarm, performing a simple stationary task. In the long term we plan to address the following questions: *(i)* Are the intuition behind the measures in accordance with the observed robot swarm behaviour? And is the observed behaviour coherent with the complexity values measured? *(ii)* What are the most informative measures? *(iii)* What are the complexity measures most suited for such an application? *(iv)* Are there

* Corresponding author, `andrea.roli@unibo.it`

phenomena in the swarm behaviour that can be detected just by observing the complexity values measured? The outcome of this study is expected to provide guidelines for the choice of the most informative indicators for more complex tasks. In the following, we summarise the measures considered in this study and we detail the robot swarm task. Subsequently, we provide a succinct view of the main results, emphasising the ones that enable us contributing to answer the questions raised above.

Measures of complexity

In the scientific literature the word *complexity* appears with different meanings, each related to a specific interpretation of the term. As a consequence, there is no unique measure of complexity, but in fact many measures have been proposed in the literature.[3] In this research we selected complexity measures according to some important features for swarm robotics. *In primis*, complexity is interpreted as difficulty in predictability and related to the presence of patterns in the dynamics of the swarm. In addition, we focus primarily on the complexity of the dynamics of the system observed in its environment, rather than the individual complexity of a robot controller. Finally, as a consequence of the fact that we deal with data collected during experiments, the measures used should be applied to time series of finite length. We selected and implemented complexity measures based on Shannon entropies [5], Lempel-Ziv [3] and similar measures, as well as set-based complexity [2]. Due to excessive computational resources required, for this preliminary step we did not applied measures of complexity based on model construction, such as the ones by Crutchfield et al. [1].

Case study: Random walk with collision avoidance

We defined a case study that requires a simple software controller for the robots and few parameters to be tuned; moreover, the mission the swarm has to accomplish should be modelled as a stationary process and its level of complexity should be sufficiently high to be measured and produce non-trivial results, but at the same time rather limited so as to allow an easy interpretation of the results. We performed our experiments in a simulated environment by the means of ARGoS [4], one of the most widespread swarm robotics simulators. The robot chosen to be simulated is an *e-puck*, equipped with 8 infra-red transceivers positioned around the circular body and two wheels.

Behavior: random walk with collision avoidance The random walk behaviour is a strategy for space exploration commonly used in swarm robotics. We implemented this strategy as the alternate execution of straight movements

[3] A survey on the literature on complexity measures is out of the scope of this report. An introduction to the subject may be found in Lindgren's lecture notes in Information Theory for Complex Systems available at `http://studycas.com/c/courses/it`.

and static rotations: at each time step of an experiment, the e-puck robots can either move forward for a given distance or rotate at a given angle. In our implementation of the random walk behaviour the robots walk straight for a maximal distance W_s. After this maximal distance is travelled, or if an obstacle is perceived in front of the robot, the static rotation phase is triggered. During the rotation, a robot turns left or right with same probability, with a rotation angle R_a given by a multiple of 10° taken uniformly between 0 and R_a. Once the robot has completed the rotation, it can once again move forward under the condition that no obstacles are on the way. Conversely, if the path is not clear in front of the robot, another static rotation phase is immediately started.

Experimental settings The state of a robot performing this kind of random walk can be simplified and expressed by means of three possible states: Straight, Left, Right. Hence, at each instant, the state of the whole swarm of N robots can be represented by a vector of symbols, each from the alphabet {S,L,R}. For each run, we recorded the state vector of the swarm every 10 milliseconds. The complexity measures where applied to this symbolic sequence depending on the definition of the measure, i.e., either to the whole vector state (e.g. for set-based complexity) or by averaging the values computed across all the robots (e.g. for entropies). We executed multiple runs of the random walk behaviour with different parameters: $N \in \{1, 10, 20, 40\}$, $W_s \in \{10, 20, 30\}$, $R_a \in \{4, 9, 18\}$; all possible combinations of parameter values have been tested. Additionally to the parameters of random walk behaviour, we ran two types of experiments. The first one involves a single robot evolving in an infinite space with no obstacles nor boundaries. The second one involves a swarm evolving in an enclosed environment. The walls form a dodecagonal shape with an area equal to 4.91 m^2. The swarm is composed of robots all controlled by the same random walk behaviour. At the beginning of each experiment, the robots were uniformly distributed in the dodecagonal arena. Every possible combinations of parameter were used in the two environments. Each experiment was repeated 10 times. Therefore, a total of 300 experiments were ran.

Results

Although preliminary, these results enable us to state that the complexity measures we used are able to capture the relevant features of robot swarm dynamics: the higher the number of forward steps, the more regular the robot trajectory, and the larger the maximum turning angle, the less regular the robot trajectory. Indeed, entropy measures—as well as the measures related to compression—are higher in the dodecagonal arena than in the infinite one, and increase with the turning angle. Moreover, in the dodecagonal arena we observe the effect of varying the number of robots: the higher the number of robots, the less regular the robot trajectory, and the higher the number of robots, the stronger the interaction among robots. Also these phenomena are reflected into the measures applied. Block entropies and their derivatives such as correlation length seem to be particularly informative and deserve further detailed investigations. For lack of space, in Fig.1 we just show a representative example, which was obtained with correlation information[4]. An in-depth analysis of the results is subject of ongoing work. The next steps of this work will be focused on experiments concerning stationary cases, such as flocking and memoryless foraging with random walk. In addition, we will move also to non-stationary cases, for example an aggregation mission.

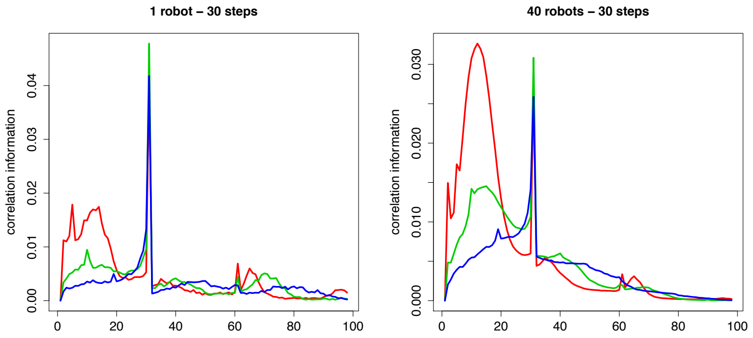


Fig. 1. A representative example involving the correlation information from length n in the dodecagonal arena. Colours are used to differentiate among the three possible turning angle values: 40° in red, 90° in green and 180° in blue. Note the peak corresponding to the length step denoting periodicity in the behaviour and an irregular peak at a lower number of steps which corresponds to the collision avoidance behaviour.

References

1. Crutchfield, J.: The calculi of emergence: Computation, dynamics, and induction. Physica D 75, 11–54 (1994)
2. Galas, D., Nykter, M., Carter, G., Price, N.: Biological information as set-based complexity. IEEE Tran. on information theory 56, 667–677 (2010)
3. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Trans. on Information Theory 22(1), 75–81 (1976)
4. Pinciroli, C., Trianni, V., O'Grady, R., Pini, G., Brutschy, A., Brambilla, M., Mathews, N., Ferrante, E., Di Caro, G., Ducatelle, F., Birattari, M., Gambardella, L., Dorigo, M.: ARGoS: a modular, multi-engine simulator for heterogeneous swarm robotics. Swarm Intelligence 6(4), 271–295 (2012)
5. Shannon, C.: A mathematical theory of communication. The Bell System Technical Journal 27(1,2), 379–423,623–656 (1948)

[4] The *correlation information from length n* is defined as $k_n = -\Delta^2 H_n = -H_n + 2H_{n-1} - H_{n-2}$, where H_n is the block entropy [5] for length n.

Inferring Global Properties of Biological Networks with a Relevance Index Method

Marco Villani[3], Laura Sani[1], Michele Amoretti[1],
Emilio Vicari[1], Riccardo Pecori[1,4], Monica Mordonini[1],
Stefano Cagnoni[1], and Roberto Serra[3]
Contact: marco.villani@unimore.it

[1]Dip. di Ingegneria e Architettura, Università di Parma
[2]Dip. di Informatica - Scienza e Ingegneria - Università di Bologna - Sede di Cesena
[3]Dip. Scienze Fisiche, Informatiche e Matematiche
Università di Modena e Reggio Emilia
[4]SMARTEST Research Centre, Università eCAMPUS, Novedrate (CO)

Keywords: Complex Systems, Biological Networks, Dynamical Behavior, Relevance Index, T-Helper Cells

1 Motivation

Nowadays a plethora of molecular data results in a vast amount of pathways, networks of interactions and molecular scenarios. A large quantity of information is available on many biological systems, and researchers use it to infer global properties of biological networks [4, 7]. In spite of the strong representational power and flexibility of networks, there are, however, two major limitations affecting most studies in the field [5, 9]:

- the information about the underlying true interactions is often incomplete, so the inferred networks do not provide a complete picture of the interactions in the system under study;
- network studies are often concerned with "static" topological information, like connectivity and betweenness, whereas, in order to understand the functionality of a system, it is important to study its *dynamical properties.*

Modeling the dynamic behavior of such systems is difficult, due to the lack of kinetic data and to computational limitations. Among the methods for facing this problem, those based on steady-state approximation are widely used [3, 10]. Nevertheless, these kinds of analysis do not provide enough constraints to find a unique solution to the problem: thus researchers support these techniques by means of suitable hypotheses as, for example, minimization or maximization issues [10].

In order to overcome the limitations of steady-state methods, it is worthwhile to resort to methods able to directly deal with the dynamical repertoire of the system. In this paper, we use a recently proposed approach, the *Relevance Index* (RI for short) method [11, 12], which has the following features:

1. it is based on the observation of the dynamical states of the system, without requiring any *a priori* knowledge of the interactions among variables;
2. it does not impose any limitation on the type of dynamical behavior;
3. it provides information about the organization of the system itself; indeed, complex systems often display complex organizational features that cannot be captured by a simple tree-like structure;
4. it is robust against noisy or incomplete data.

2 Method

We focus on identifying subsets of nodes that are good candidates as relevant subsets (*Candidate Relevant Subsets*, CRSs in the following) for describing the organization of a dynamical system. We suppose that

- the values of the system nodes, or variables, express the observed states of the system;
- there exist one or more subsets where these variables are acting in a coordinated way;
- the variables of each subset interact with the other system variables more weakly than among one another internally.

This methodology is based on the *Relevance Index* (RI for short) [11, 12]. The computation of the RI, which is based on the Shannon's Entropy, is usually performed through observational data, and probabilities are estimated as the relative frequencies of the values observed for each variable. The RI scales with the size of the CRS, so it has to be normalized with respect to a reference system where no dynamical structures are present, i.e., a homogeneous system. For this reason, a statistical significance index T_c was introduced, to measure the deviation of the normalized RI of a group of variables with respect to the statistics of a corresponding homogeneous system.

In the generation of the homogeneous system, homogeneity is maintained by forcing all off-diagonal elements of the correlation matrix to have the same constant value ρ. Such a value is computed as the average of all pairwise correlations of the observed variables. In this way we preserve both homogeneity and dependence among the different variables. In order to generate a homogeneous system with the aforementioned features, we use the NORTA method [1], a mathematical procedure that creates the issue of creating random vectors of correlated samples, given the set of their marginal distributions and a measure of the dependence among them.

Finally, a further *sieving algorithm* [2] can be used to isolate the most representative CRSs, i.e., those having the highest T_c. This procedure is based on the following criterion: if CRS $C1$ is a proper subset of $C2$ and ranks higher than CRS $C2$, then $C1$ is considered to be more relevant than $C2$. Thus it is possible to keep only those CRSs not included in or not including any other CRS with higher T_c. The sieving activity stops when no more eliminations are possible and the remaining sets of variables are the true relevant sets.

3 Experimental Results

As a use case of the RI method, we illustrate the analysis of the T-helper network. In the vertebrate immune system there are two main kinds of T lymphocytes: the T cytotoxic cells (Tc) and the T-helper cells (Th), differently distributed in the two main T-helper cell sub-types Th1 and Th2. Both sub-types derive from a common precursor Th0 through a rather complex differentiation path, modeled in [6, 8]. In this work, we use the discretization of an updated version of these paths described in [6].

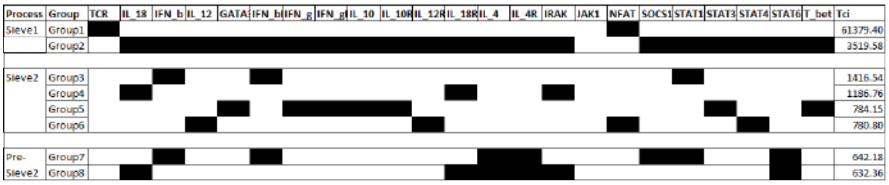
We simulated a gene regulatory network by means of a synchronous Boolean system (19 nodes + 4 "sensor" nodes, which receive their input from outside the Th differentiation system and constitute the way the system is aware of its context). There are 2^{19} different initial conditions for each of the 2^4 different scenarios identified by the "sensor" nodes. However, we found only 33 different asymptotic behaviors (all fixed points). Three of these attractors coincide with the gene expression of Th0, Th1 and Th2 cells. These attractors are presented in [6] as the only really stable states.

However, the gene regulatory network can express 33 different asymptotic behaviors. Indeed, this fact should give us some information about the dynamical organization of the system. Therefore, to extract this information, we applied the RI methodology (i) to the mere juxtaposition of these attractors or (ii) by weighting their presence proportionally to their basins of attraction.

In both cases the RI analysis induces an interesting interpretation of the dynamical data, which provides an expressive explanation of the system functioning. The same knowledge is not derivable from the static analysis alone.

Furthermore, the usual dynamical analyses are mainly focused on the detailed reproduction or prediction of the systems behaviors [6] and therefore are not suitable for a highly abstracted and "global" vision of the system functioning.

It is worth mentioning, however, that the RI method can be applied directly to experimental data, if available. In this case, it can provide an effective idea of the dynamical organization of the observed system without requiring any knowledge of topology, dynamical rules, or parameters [11].



Process	Group	TCR	IL_18	IFN_b	IL_12	GATA3	IFN_bR	IFN_g	IFN_gR	IL_10	IL_10R	IL_12R	IL_18R	IL_4	IL_4R	IRAK	JAK1	NFAT	SOCS1	STAT1	STAT3	STAT4	STAT6	T_bet	Tci
Sieve1	Group1	■																■							61379.40
	Group2		■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■	■	■	■	■	3519.58
Sieve2	Group3			■			■													■					1416.54
	Group4		■										■			■									1186.76
	Group5					■		■	■	■	■										■			■	784.15
	Group6				■							■						■				■			780.80
Pre-Sieve2	Group7			■			■							■	■				■	■			■		642.18
	Group8		■										■	■	■	■							■		632.36

Fig. 1. The table shows the groups detected by the application of the RI methodology followed by the sieving algorithm (groups 1-6): each group is represented as a row where black boxes denote the variables belonging to it.

Acknowledgments

The work of Michele Amoretti was supported by the University of Parma Research Fund - FIL 2016 - Project “NEXTALGO: Efficient Algorithms for Next-Generation Distributed Systems”.

This work greatly benefited from discussions with Andrea Roli, to whom the authors are warmly thankful.

References

1. Cario, M.C., Nelson, B.L.: Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Tech. rep. (1997)
2. Filisetti, A., Villani, M., Roli, A., Fiorucci, M., Serra, R.: Exploring the organisation of complex systems through the dynamical interactions among their relevant subsets. In: Proc. of the European Conference on Artificial Life. pp. 286–293 (2015)
3. Herrgård, M.J., Covert, M.W., Palsson, B.Ø.: Reconstruction of microbial transcriptional regulatory networks. Current opinion in biotechnology 15(1), 70–77 (2004)
4. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L.: The large-scale organization of metabolic networks. Nature 407(6804), 651–654 (2000)
5. Johnson, J.: Hypernetworks in the science of complex systems, vol. 3. World Scientific (2013)
6. Mendoza, L., Xenarios, I.: A method for the generation of standardized qualitative dynamical systems of regulatory networks. Theoretical Biology and Medical Modelling 3(1), 13 (2006)
7. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.L.: Hierarchical organization of modularity in metabolic networks. Science 297(5586), 1551–1555 (2002)
8. Remy, E., Ruet, P., Mendoza, L., Thieffry, D., Chaouiya, C.: From logical regulatory graphs to standard Petri nets: dynamical roles and functionality of feedback circuits. In: Transactions on Computational Systems Biology VII, pp. 56–72. Springer (2006)
9. Roberto Serra, R., Villani, M.: Modelling Protocells. Springer Netherlands Science+Business Media Dordrecht, (2017)
10. Ruppin, E., Papin, J.A., De Figueiredo, L.F., Schuster, S.: Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. Current opinion in biotechnology 21(4), 502–510 (2010)
11. Villani, M., Filisetti, A., Benedettini, S., Roli, A., Lane, D., Serra, R.: The detection of intermediate-level emergent structures and patterns. In: Proceedings of the European Conference on Artificial Life. pp. 372–378 (2013)
12. Villani, M., Roli, A., Filisetti, A., Fiorucci, M., Poli, I., Serra, R.: The search for candidate relevant subsets of variables in complex systems. Artificial life (2015)

K-means PSO for searching relevant variable subsets in complex systems

Gianluigi Silvestri[1], Laura Sani[1], Michele Amoretti[1], Riccardo Pecori[1,2],
Emilio Vicari[1], Monica Mordonini[1], and Stefano Cagnoni[1]
Contact: stefano.cagnoni@unipr.it

[1]Dip. di Ingegneria e Architettura, Università di Parma
[2]SMARTEST Research Centre, Università eCAMPUS, Novedrate (CO)

Keywords: Complex Systems, Relevant Sets, Particle Swarm Optimization, K-means clustering

Motivation The study of a complex system is often related to the identification of its emergent dynamical structures. Complex systems can be described by analyzing the collective behaviors and the emerging properties of their components, which are usually well-known and defined in terms of the system state variables. In several cases, however, the specific interactions among the elements of a system are not known; therefore, it becomes necessary to deduce some information about the organization of the system by observing the behavior of its relevant dynamic components.

In a previous work, Villani et al. proposed to use a method, previously introduced by Tononi et al. [6] for analyzing the coordinated behavior of sets of neurons in the brain cortex, with the more general aim of identifying candidate dynamical structures in complex systems [8]. Such a method detects subsets of the system variables which behave in a coordinated and coherent way while loosely interacting with the remainder of the system, by associating to each of them an information theoretical measure, called *Relevance Index* (*RI*). This measure can be normalized with respect to a reference system (termed *homogeneous system*) wherein the variables have the same marginal distribution as in the data set but are homogeneously correlated with each other. The normalized measure that can thus be computed, termed T_c index, quantifies how much a subset of variables of the system under investigation deviates from such a *neutral* condition.

The subsets can thus be ranked according to their T_c: the higher the T_c, the higher the correlation degree between the variables in a subset and the lower the interaction with the variables outside the subset. The most relevant sets, characterized by the highest T_c values, are referred to as *Candidate Relevant Sets* (CRSs). In fact, the properly called *Relevant Subsets* (RSs) are CRSs that do not include (or are not included in) other CRSs with higher T_c values.

This means that a full description of a dynamical system requires that the T_c index be computed for each possible subset of the system variables. An exhaustive analysis becomes unfeasible as the dimension of the system increases, because the number of CRSs increases exponentially with the system size. This *curse of dimensionality* makes it impossible to analyze large systems exhaustively, even using massively parallel hardware such as GPUs, which fit the computational needs of this problem particularly well [7].

Method In this paper, we use a niching Particle Swarm Optimization (PSO) [4] variant, the so-called *K-means PSO* [3], for searching relevant dynamical structures and extract its CRSs when the dimension of the variable space makes an exhaustive search unfeasible.

In our method, each particle of the swarm represents a subset of variables as a binary string of size N, in which each bit set to 1 denotes the inclusion in the CRS of the corresponding variable. Since PSO operates on floating-point vectors, such a binary encoding is obtained by setting to 1 the particle's element having a positive value, and to 0 the negative ones.

The fitness function to be maximized corresponds to the T_c index of the CRS associated to the particle and is implemented through a CUDA C [1] kernel that can compute in parallel the fitness values of large blocks of particles.

With respect to the basic PSO algorithm, which is aimed at finding the global optimum of the fitness function, in the K-means PSO the search process is enhanced by a niching technique that maintains the diversity among the particles of the swarm and allows the swarm to explore and converge onto many peaks in parallel. In particular, in K-means PSO, at regular intervals, the K-means clustering algorithm [2] is applied to the swarm to reorganize it into sub-swarms characterized by their elements' proximity in the search space. The standard PSO algorithm is then independently applied to each sub-swarm thus identified.

The K-means PSO pseudo-code is reported in Algorithm 1.

Algorithm 1 K-means PSO pseudo-code.

```
procedure KPSO
    randomly initialize the particles' positions and velocities
    compute each particle's fitness
    for t = 1 to T do                                   ▷ T = number of iterations
        if t mod c = 0 then                             ▷ every c steps
            run the K-means algorithm to identify niches
        end if
        update each particle's velocity                 ▷ as in standard PSO
        update each particle's position
        compute each particle's fitness
        update each particle's and each niche's best position
    end for
end procedure
```

Experimental Results The K-means PSO has been evaluated on a set of meaningful systems described by Boolean variables. The results have been compared with those achieved by an exhaustive search, when computationally feasible, and by another hybrid meta-heuristic, based on genetic algorithms and local search, which we had previously developed [5].

The first case study is a simulation of a chemical system called Catalytic Reaction System (CatRS), featuring 26 variables. The second one is a stochastic artificial system reproducing a Leaders & Followers (LF) behavior, described by 28 variables. In the third example, denoted as Green Community Network (GCN), the data come from a real situation (participation of partners in project meetings) and are described by 56 variables, a size for which an exhaustive search is not feasible on a standard computer, even using GPU parallelization. Therefore, it was analyzed only by the two meta-heuristics.

The results were evaluated both in terms of quality and of speed-up with respect to an exhaustive search. Quality was evaluated, when feasible, counting the number of CRSs, among the highest-T_c CRSs, that were detected by the exhaustive search but not by K-means PSO. Regarding the lager-sized system, for which the results of the exhaustive search were not available, we relied on the opinion of an expert to assess their quality. Given the stochastic nature of PSO, we repeated each experiment for 30 times to also assess the repeatability of the results.

The results obtained with the smaller-size systems for which the comparison is possible, are virtually the same provided by the exhaustive search based on the same code used to compute the T_c index. Only occasionally at most one out of the first 50 CRSs, which is usually more than enough to understand the main dynamics of the systems we took into consideration, was not detected by K-means PSO. The results obtained on the GCN were judged as reasonable by the expert. In this case, too, the same results were obtained in the different runs, with marginal occasional differences confined to the least significant CRSs.

K-means PSO exhibited both good exploration capabilities, thanks to its niching behavior, and fast convergence. The latter is probably justified by the nature of the problem that is such that *dominant* variables exist, that are repeatedly present in the relevant sets. This means that PSO, which converges extremely fast when solving separable optimization problems, but struggles when dealing with strongly-dependent variables, can easily find relevant variables even searching independently along each dimension as it actually does.

This made it possible for K-means PSO to achieve a significant speed-up with respect to both the exhaustive search and the hybrid meta-heuristic. Of course, basically, the lower speed-up of the latter is partially due to the presence of the local search, which ensures a better exploration of the neighborhoods of the local minima. In the tests we made, however, K-means PSO did not suffer from the lack of such a feature.

It is to be noticed that all methods relied on the same GPU implementation of the fitness function, which means that the differences observed depend only on the algorithms' efficiency and complexity and not on their implementation.

Table 1. Results obtained by K-means PSO (K) on three different systems and comparison with [5] in terms of time and speed-up with respect to an exhaustive search (E).

System	Size	N. data	Time[s](E)	Time[s] [5]	Time[s](K)	Speed-up [5]	Speed-up (K)
CatRS	26	751	53	24	6	2.2	8.8
LF	28	150	196	19	2	10.3	98
GCN	56	124	n.a.	71	18	n.a.	n.a.

Acknowledgments

The work of Michele Amoretti was supported by the University of Parma Research Fund - FIL 2016 - Project “NEXTALGO: Efficient Algorithms for Next-Generation Distributed Systems”.

The authors would like to thank Andrea Roli, Roberto Serra, and Marco Villani for the enlightening discussions and comments on this work.

References

1. CUDA Toolkit. http://developer.nvidia.com/cuda-toolkit, [Online; accessed July 10, 2017]
2. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297 (1967)
3. Passaro, A., Starita, A.: Particle swarm optimization for multimodal functions: A clustering approach. Journal of Artificial Evolution and Applications (2008)
4. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intelligence 1(1), 33–57 (2007)
5. Sani, L., Amoretti, M., Vicari, E., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: Efficient search of relevant structures in complex systems. In: Proceedings of 15th International Conference of the Italian Association for Artificial Intelligence, (AI*IA 2016), Genoa. (2016)
6. Tononi, G., McIntosh, A., Russel, D., Edelman, G.: Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. Neuroimage 7, 133–149 (1998)
7. Vicari, E., Amoretti, M., Sani, L., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: GPU-based parallel search of relevant variable sets in complex systems. In: Rossi, F., Piotto, S., Concilio, S. (eds.) Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry: 11th Italian Workshop, WIVACE 2016, Fisciano, Italy, October 4-6, 2016, Revised Selected Papers. pp. 14–25 (2017)
8. Villani, M., Roli, A., Filisetti, A., Fiorucci, M., Poli, I., Serra, R.: The search for candidate relevant subsets of variables in complex systems. Artificial Life 21(4) (2015)

Functional interactions in socio-economic complex networks: detection of subsets of agents through the application of the Relevance Index (RI)

Riccardo Righi

European Commission, Joint Research Centre (JRC), Unit B6 - Digital Economy
Riccardo.RIGHI@ec.europa.eu.it

Since the strong arise of the use network analysis in middle 90's [1], social sciences have mostly focused their efforts in the comprehension of the structure (and its evolution) of durable relationships (e.g. friendship). However, in some contexts the concept of connections has to be intended distinctly, since it deals with something that is more similar to a series of flickering interactions, than to a structure of stable relationships. There are cases in which, even if any new relationship is established, new interactions occur: people daily exchange messages with long-time friends, and enterprises repeatedly collaborate with partners they already know. The awareness of the distinction between the conceptualization of connections intended as relationship and the conceptualization of connections intended as interactions, has a crucial relevance especially in the detection of subsets of agents in complex networks. When interactions are observed, the presence and the evolution of meso-structures can be hardly investigated through the application of methods that looks for a process of stepwise creation/dissolution of connections [2]. The continuous activation/inactivation of agents' links makes necessary the use of methodologies that, instead of considering the statistical significance of the formation of relational architectures, shall focus into the physical order contained in the phenomena occurred [3].

Following the pioneering contribution of Tononi *et al.* [4,5,6,7], concerning the efficacy of the use of an entropy based indicator (called Cluster Index, henceforth CI) to detect functional groups of neurons in the brain region, Villani *et al.* [8,9,10] developed a statistical approach to evaluate the significance of the integration of agents' joint behaviors. The developed algorithm, called Relevance Index (henceforth, RI), was tested in Boolean artificial network models, in catalytic reaction networks and in biological gene regulatory systems [8,9,10], producing consistent results in the identification of emergent meso-level structures [8]. The capacity shown by the RI was to uncover existing relationships without using as inputs any information concerning the topology of the network, but only focusing on agents' states of activation of agents over time. The creation and the development of the RI algorithm have opened new paths[1] to take on a new relevant dimension of the analysis: the level of integration that is embodied in the

[1] Up to now, the detection and the analysis of communities typically have been performed through the consideration of similar characteristics of agents, or through the analysis of network structures [11,12]

reduction of the entropy of the joint behaviors of group of agents having similar functions with regard to the functioning of the whole considered system. Thanks to a specific case study, concerning network innovation policies implemented by Region Tuscany (Italy) in the programming period 2000-2006 [13,14,15,16], an application of RI algorithm in a socio-economic context of analysis is under development. After describing the available data, with particular attention to the definition of variables concerning agents' activity profiles, a heuristic is proposed to overcome the resulting redundancy in RI output[2]. The suggested approach follows the issues that arose after the implementation of the methodology into large datasets and, acknowledging other researches [17,18], attempts to explore ways to develop the process of analysis.

The number R of rounds in which to run the RI analysis over progressively reducing sets of agents (before each round of analysis, the best mask detected at the round before is skimmed), is the first parameter according to which the entire process of analysis is developed. The output of each round is processed with an agglomerative hierarchical cluster algorithm[3] aimed at grouping together the most similar masks that have been detected. A second parameter, the percentage threshold v_{OV}, is used to select an appropriate number of leaves in which to cut the dendrogram of each round of analysis. The cut is performed using the largest number of leaves among those producing a partition in which the percentage of agents that belong to more than one detected mask (up to the number of agents that are included in at least one mask) is lower than, or equal to, the percentage threshold v_{OV}, so acting as an upper limit to the degree of overlaps among masks. For each resulting cluster of masks, the group of agents with the joint behaviors the significantly farthest from randomness (the statistical significance of the CI is measured by the statistic called t_{CI}) is taken as representative for the entire cluster. This proposed refinement of the results of each round of the RI analysis is performed towards the identification of masks that:

- are the most different (in terms of agents belonging to them), thus making possible to attempt a spread exploration of the system under analysis;
- are the most significant (in terms of t_{CI});
- produce a limited degree of overlaps, thus guaranteeing simplicity in the structure of the emergent partition.

Finally, a third parameter, the percentage threshold v_{SM}, is used as reference for the similarity (in terms of Jaccard Index) between all the masks detected in all the R rounds, in order to avoid redundancies. Independently from the round

[2] The problem of redundancy in RI output is caused by the specific proportion, in the considered data, between variables and observations. The number of agents is six times higher than the number of instants in time over which the system is observed. To better evaluate agents' behavior over time, this proportion should be the opposite. The initial test of RI was developed on a system of 12 agents observed over 30 instants in time [10]

[3] Simple matching to compute the distances between masks of agents, and complete linkage as aggregation method.

of the RI analysis in which the masks have been detected, if a couple of masks producing a level of similarity higher than the threshold v_{SM} is found, only the mask with the highest t_{CI} is kept.

Since at the time being no specific values of the parameters v_{OV} and v_{SM} have been identified, two discrete sets of percentage thresholds, namely V_{OV} and V_{SM}, are used to develop an explorative estimation of the two aforementioned parameters. Since all possible combinations of the two are considered, a final set of $|V_{OV}| \cdot |V_{SM}|$ partitions is obtained. Considerations are made on some partitions selected in according with (i) a principle of maximization of the percentage of agents overall involved in the partition (not necessarily all agents belong to at least one mask) and (ii) a principle of minimization of the percentage of agents that belong to more than one subset[4].

Keywords - complex networks, interactions, physical order, functional subsets, socio-economic system

References

1. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press (1994)
2. Righi, R., Roli, A., Russo, M., Serra, R., Villani, M.: New paths for the application of dci in social sciences: Theoretical issues regarding an empirical analysis. In Rossi, F., Piotto, S., Concilio, S., eds.: Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry: 11th Italian Workshop, WIVACE 2016, Fisciano, Italy, October 4-6, 2016, Revised Selected Papers, Cham, Springer International Publishing (2017) 42–52
3. Hidalgo, C.: Why Information Grows: The Evolution of Order, from Atoms to Economies. Basic Books (2015)
4. Tononi, G., Sporns, O., Edelman, G.M.: A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proceedings of the National Academy of Sciences of the United States of America **91**(11) (1994) 5033–5037 ISSN: 0027-8424.
5. Tononi, G., Sporns, O., Edelman, G.M.: A complexity measure for selective matching of signals by the brain. Proceedings of the National Academy of Sciences of the United States of America **93**(8) (1996) 3422–3427 ISSN: 0027-8424.
6. Tononi, G., McIntosh, A.R., Russell, D.P., Edelman, G.M.: Functional clustering: identifying strongly interactive brain regions in neuroimaging data. NeuroImage **7**(2) (1998) 133–149 DOI: 10.1006/nimg.1997.0313.
7. Tononi, G., Sporns, O., Edelman, G.M.: Measures of degeneracy and redundancy in biological networks. Proceedings of the National Academy of Sciences **96**(6) (1999) 3257–3262 ISSN: 0027-8424, 1091-6490.
8. Villani, M., Filisetti, A., Benedettini, S., Roli, A., Lane, D., Serra, R.: The detection of intermediate-level emergent structures and patterns. In: ECAL. (2013) 372–378

[4] This selection tries to address, respectively, the two following objectives: i) to consider partitions in which a large part of the initial system is considered and to (ii) avoid considering partitions in which an extreme overlapping of the detected subsets does not allow a clear interpretation of the results.

9. Villani, M., Benedettini, S., Roli, A., Lane, D., Poli, I., Serra, R.: Identifying emergent dynamical structures in network models. In Bassis, S., Esposito, A., Morabito, F.C., eds.: Recent Advances of Neural Network Models and Applications. Number 26 in Smart Innovation, Systems and Technologies. Springer International Publishing (2014) 3–13 DOI: 10.1007/978-3-319-04129-2_1.
10. Filisetti, A., Villani, M., Roli, A., Fiorucci, M., Serra, R.: Exploring the organisation of complex systems through the dynamical interactions among their relevant subsets. In: Proceedings of the European Conference on Artificial Life 2015 (ECAL 2015). Volume 13. (2016) 286–293
11. Fortunato, S.: Community detection in graphs. Physics reports **486**(3) (2010) 75–174
12. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports **659** (2016) 1–44
13. Caloffi, A., Rossi, F., Russo, M.: The emergence of intermediary organizations: A network-based approach to the design of innovation policies. In: Handbook On Complexity And Public Policy. Edward Elgar Publishing Cheltenham GBR (2015) 314–331 ISBN: 978-1-78254-951-2.
14. Caloffi, A., Rossi, F., Russo, M.: What makes SMEs more likely to collaborate? analysing the role of regional innovation policy. European Planning Studies **23**(7) (2015) 1245–1264
15. Rossi, F., Caloffi, A., Russo, M.: Networked by design: Can policy requirements influence organisations' networking behaviour? Technological Forecasting and Social Change **105** (2016) 203–214 DOI: 10.1016/j.techfore.2016.01.004.
16. Russo, M., Caloffi, A., Rossi, F.: Evaluating the performance of innovation intermediaries: insights from the experience of tuscany's innovation poles. In: Plattform Forschungs- Und Technologieevaluierung. Volume 41. (2015) 15–25 ISSN: 1726-6629.
17. Sani, L., Amoretti, M., Vicari, E., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: Efficient search of relevant structures in complex systems, Cham, Springer International Publishing (2016) 35–48
18. Vicari, E., Amoretti, M., Sani, L., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: Gpu-based parallel search of relevant variable sets in complex systems. In Rossi, F., Piotto, S., Concilio, S., eds.: Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry: 11th Italian Workshop, WIVACE 2016, Fisciano, Italy, October 4-6, 2016, Revised Selected Papers, Cham, Springer International Publishing (2017) 14–25

Modelling Emerging Topics in a Techno-Economic Segment (TES) Network Extended Abstract

Sofia Samoili[1], Riccardo Righi[1], Montserrat Lopez-Cobo[1], and Giuditta De Prato[1]

European Commission, Joint Research Centre (JRC), Directorate B - Growth and Innovation, Unit B6 - Digital Economy
sofia.samoili@ec.europa.eu, riccardo.righi@ec.europa.eu,
montserrat.lopez-cobo@ec.europa.eu, giuditta.de-prato@ec.europa.eu,
Edificio EXPO, Calle Inca Garcilaso, 3,
E-41092, Seville, Spain

1 Research Description

The impact of innovation is multidimensional and covers technological, economical, geographical and social aspects. The fuzzy and incessantly changing boundaries of evolutionary and revolutionary or radical technological innovations challenge the current economic, social, environmental and institutional foundations, describing the unravelling of the fourth industrial revolution.

As stated by David Lane in his theory of innovation [1–4], innovations flourish from a process of interactions occurring between agents and artefacts [1]. Far from trying to predict the object of innovation [2], a schematic representation of this process is possible and some crucial elements can be identified. In particular, the existence of a cascade process that follows the introduction of an innovation can be recognized: agents interacting with the new artefacts lead to a development and an amelioration of the same [3]. Following this approach, with the aims to

[1] Artefacts are those means that are produced and used by agents in order to wrest more usable matter and energy from their environments, allow more of them to live in a given environment, and sometimes allow them to live longer and dedicate more of their time and energy to reproduction [2].

[2] The theory proposed by David Lane does not have as objective the prediction of the results of innovation processes. This could be conceded as an oxymoron. Quoting Lane (2011): ”From one point of view, a ”theory of innovation” is an oxymoron. If, as many scientists believe, a theory is supposed to lead to verifiable predictions of the phenomenon under study, then a theory of innovation should predict innovations which would mean the process leading to innovations the theory was meant to explicate is just an historical dead-end that could be replaced as innovation-generator by the theory itself! Of course, this is silly: the theory could illuminate aspects of the process without ”predicting” the new artefacts that were the process outcomes of primary economic and social interest”.

[3] Two kinds of inventions have to be distinguished: those that develop some existing functionalities (better-faster-cheaper) and those that develop new functionalities. Here, we intend the first type.

analyse the presence of processes of specification of an existing technology (and not the disruptive generation of a new one) and to investigate the patterns of the observed cascades, the development of a tech mining approach is developed in the current study that will capture such patterns and will suggest the emerging technological innovative trends that could be potentially anticipated.

This multi-level approach combines a network analysis and text clustering with a topic model in the field of photonics, as a representative technologically promising area with numerous applications. The multi-layer analysis is invoked to define and analyse the communities among agents found in publications, EU-funded projects (FP7, H2020) and patents, carrying a combination of technological and economical properties. Each of the ensued communities is characterised through text topic modelling, by several noun-phrases, and new emerging topics are eventually predicted in the early stages of co-development. Two of the most effective non-parametric models are employed, the dynamic topic modelling and the n-gram Markov chain model (MCM). The qualitative analysis of the resulted relational data that convey both technical and economic information, aims to provide such insights that will enable the identification of an original topic. This will be semantically analysed in the last level of the approach to map the affected aspects through the relational data. In the following sections the brief literature review reveals the need for the presented approach and subsequently the proposed framework of the on-going study is described.

2 Multi-level Methodological Approach

The dataset consists of EU-funded projects (FP7 and H2020), publications and patents in the field of photonics with structured and unstructured data. The data source includes agent's related variables: name and type of the agent (universities, research institutions and companies), financial indicators when available (e.g. companies ́turnover), location; and document-related variables: allocated funds, abstract, title, key phrases, IPC/CPC codes for patents, etc.

2.1 Level 1: Network Analysis

In order to consider into the analysis the process of interaction that occurs among agents in the development of functionalities [2], initially a community detection analysis is performed to evaluate which agents exchange intensively information. In this scope, a multi-layer network is built in accordance with the different observed activities (publications, patents, participations in FP7 projects, and participations in H2020 projects) and the Infomap multi-layer community detection algorithm [5] is run on it. Since Infomap is an information based algorithm that minimizes the length of the description of a random flow (circulating over the network) using a two-level binary description [6–8], the resulting communities can be interpreted as groups of agents within which communications are fluent [9]: flows between communities happen, but the majority of flows occur within each of them. This fact is crucial from the point of view of the adopted approach concerning innovation: groups of agents exchanging information through interactions are identified.

2.2 Level 2: Topic Modelling

In this level the focus is set on the text data for each community detected in the previous level of the approach. The relations that each term holds will be addressed in the following level, following the reduction of the terms in the semantic space to the most significant for the current and/or the emerging technological trends. Natural language processing methods capture information from human-stored text formats. To ensure robust results two of the most effective according to the recent research state [10–16], non-parametric models are employed to describe the available text dataset in the field of photonics; the n-gram Markov chain model (MCM) and the dynamic topic model (DTM). The results of each model will be evaluated in the next level.

The n-gram MCM of natural language is used to retrieve information regarding the occurrence of a term and the prediction of another according the last few (n-gram) detected words, so as to assess eventually the likelihood of certain established hypotheses as dictated by the latest concept (n-gram) that is useful in the scope of the research, such as the zones of expansion and influence of the specific techno-economic segment (TES). In the stochastic process, the word sequence that is most likely to occur is defined from the following joint probability of a sequence of words $w_1^n = \{w_1, ..., w_n\}$, which is computed as the product of conditional probabilities of a word given previous words $P(w_n|w_1^{n-1})$ (Equation 1) [15]:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)...P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \tag{1}$$

then in the n-gram MCM the probability of a certain word given its entire history is approximated by the last few $n-1$ words (Markov assumption), since in a natural language the occurrence of a word sequence does not exclude its reappearance according to the number of times it occurred in its entire history. Thence by Eq. 1, the general equation for n-gram approximation of the conditional probability of the next word in a sequence becomes Eq. 2:

$$P(w_n|w_1^{n-1}) \approx \prod_{k=1}^{n} P(w_n|w_{n-N+1}^{n-1}) \tag{2}$$

The n-gram probabilities are estimated through the relative frequency (Equation 3 [15]) that computes the ratio of the frequency of a specific sequence and the frequency of a prefix.

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})} \tag{3}$$

The topic models approach is especially designed for categorical data, as the text data. The statistical assumptions that are applied follow the static latent Dirichlet allocation model. In a K-topic model, each corresponds to a distribution over a fixed vocabulary, with k the distribution of the k^{th} topic. In the static models the draw of each document is effectuated by the following generative process: (1) for each topic k, $1 \leq k \leq K$ the topic's proportions θ are defined from a distribution as the Dirichlet [16], then (2) for each word a topic assignment z is drawn (with $Z \sim Mult(\theta)$) and a word w (with $W \sim Mult(\beta)$). Hence, this assumptions for the static model indicate that the draws are exchangeable from a common set of topics. In order to incorporate the time dimension, a dynamic topic approach is studied, with the topics evolving for each time slice t. Let V denote the terms in a $K-$component topic model, then the $V-$vector of natural parameters for a k topic in time slice t is denoted by $\beta_{t,k}$ and for sequential modeling the randomness of the draws are ensured by Gaussian distributions (Eq.4) chains with the ensued values mapped in the simplex (Eq. 5).

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I) \tag{4}$$

$$\pi(\beta_{t,k})_w = \frac{exp(\beta_{t,k,w})}{\Sigma_w(exp(\beta_{t,k,w}))} \tag{5}$$

2.3 Level 3: Semantic Analysis

In this level the final aim to provide information on the characteristics of both the detected and the predicted segments by relating the entirety of the acquired results in the previous levels, is achieved through semantic analysis.

The ensued technical terms from the text modelling level are addressed in relation to their attributes from the community in which they are classified, in order to evaluate the techno-economic impact. The models will be validated by testing a subset (the FP7 projects), in order to assess if the occurred terms confirm the topics that emerged in the subsequent H2020 projects, publications or patents. The results of each model are analysed separately with the Pearson's correlation coefficient, as a similarity criterion between estimated and observed terms. The highest value of the criterion determines the most efficient model, and thence the relevant terms.

Keywords: techno-economic segment, network analysis, community detection, text modelling, n-gram markov-chain model, topic model, semantic analysis

References

1. D. A. Lane, "Innovation cascades: artefacts, organization and attributions," *Phil. Trans. R. Soc. B*, vol. 371, no. 1690, p. 20150194, 2016.

2. D. A. Lane and C. Antonelli, "Complexity and innovation dynamics," *Handbook on the economic complexity of technological change*, vol. 63, 2011.
3. D. A. Lane and R. R. Maxfield, "Ontological uncertainty and innovation," *Journal of evolutionary economics*, vol. 15, no. 1, pp. 3–50, 2005.
4. D. Lane, R. Maxfield *et al.*, "Foresight, complexity, and strategy." Santa Fe Institute New Mexico, 1995.
5. M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall, "Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems," *Physical Review X*, vol. 5, no. 1, p. 011027, 2015.
6. M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.
7. A. Esquivel and M. Rosvall, "Compression of flow can reveal overlapping modular organization in networks 2011," *arXiv preprint arXiv: 1105.0812.*
8. R. Lambiotte and M. Rosvall, "Ranking and clustering of nodes in networks with smart teleportation," *Physical Review E*, vol. 85, no. 5, p. 056107, 2012.
9. S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.
10. P. Bakhtin and O. Saritas, "Tech mining for emerging sti trends through dynamic term clustering and semantic analysis: The case of photonics," in *Anticipating Future Innovation Pathways Through Large Data Analysis.* Springer, 2016, pp. 341–360.
11. V. Prabhakaran, W. L. Hamilton, D. A. McFarland, and D. Jurafsky, "Predicting the rise and fall of scientific topics from trends in their rhetorical framing." in *ACL (1)*, 2016.
12. C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," *arXiv preprint arXiv:1206.3298*, 2012.
13. C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2011, pp. 448–456.
14. Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: How can citations help?" in *Proceedings of the 18th ACM conference on Information and knowledge management.* ACM, 2009, pp. 957–966.
15. M. J. Jurafsky D., "Speech and language processing - chapter 4: N-grams," *Lectures*, 2014.
16. D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning.* ACM, 2006, pp. 113–120.

Modeling emerging topics on sustainable urban development perception: the case of Hangzhou Future Sci-Tech City.

Debora Slanzi[1,2], Valentina Anzoise[1], and Irene Poli[1,2]

[1] European Centre for Living Technology, S. Marco 2940, 30124 Venice, Italy
[2] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari
University of Venice, Cannaregio 873, 30121 Venice, Italy
{debora.slanzi, valentina.anzoise, irenpoli}@unive.it

Abstract

Recent social and economical literature has been particularly concerned in the investigation of urban development [1,2,3]. Especially in China, where the process of accelerated urbanization is a result of rapid economic growth and dedicated policies, cities and countrysides are changing at an unprecedented scale and pace. As a consequence, landscape and lifestyle are radically transformed raising social, economic, and environmental sustainability issues and stability problems [4]. Due to the complexity of the study of urban systems, it emerges that there is an increasing need of adopting appropriate methods for analyzing and modeling social data, both from a quantitative and a qualitative perspective [5]. One particular approach for the analysis of social systems is the textual data analysis. Textual documents in fact provide a valuable source of data for the identification and the measurement of latent variables, and statistics and machine learning researchers have developed several approaches to study these structures of data [6,7,8]. Among them, Topic Modeling approaches (TM) aim to automatically inferred from textual data the rich latent topics of a set of documents or texts [6]. TMs have been successfully used across a variety of fields as they can discover complex latent structure in the data [9,10]. Topics are estimated with probabilistic distributions over a vocabulary of words and according to the co-occurence of words within each analyzed text according with a probabilistic generative process. This process considers a collection of D documents (or texts), each containing $N_d \subseteq V$ words, $d = 1, \ldots, D$, and V represents the set of distinct elements of the vocabulary used in the analysis. Moreover a set of K latent topics is defined and assumed to be representative of the documents. The probabilistic generative process consists then of the following steps:

- a V-dimensional Dirichlet probability distribution, $\beta_k \sim Dir(\eta)$, is determined for each topic k, $k = 1, \ldots, K$, assessing the probability according to which words are generated from the k-th topic;
- a K-dimensional Dirichlet probability distribution, $\theta_d \sim Dir(\alpha)$ is determined for each document d, $d = 1, \ldots, D$, assessing the expected proportion of words that can be attributed to each topic;

- for each word in the document
  - a value $z_{d,n}$ for a multinomial distribution $Z_{d,n} \sim Mult_K(\theta_d)$, $n = 1, \ldots, N_d$, is sampled denoting which topic is associated with such word, and
  - a word value $w_{d,n}$ from a multinomial distribution $W_{d,n} \sim Mult_V(\mathbf{B}z_{d,n})$, is sampled where the matrix $\mathbf{B} = [\beta_1 \cdots \beta_K]$ encodes the distributions over words in the vocabulary associated with the K topics.

When additional information regarding the documents is available, it can be included in the model as a set of covariates $\mathbf{X}$. The Structural Topic Model (STM) proposed by Roberts et al. [11,12,13] represents a particular class of TMs where the inclusion of covariates of interest can affect the topical prevalence (i.e., the frequency with which a topic is discussed) and topical content (i.e., the words used to discuss a topic) of the model. The covariates are introduced in the TM approach by means of different prior distributions for document-topic proportions and topic-word distributions. For the specific procedure on how these prior distributions are defined and how the Topic Model estimation process is modified, we refer to [13].

In this work we analyze the citizens' perception on the urban development of a recently established high-tech zone in China, i.e. Hangzhou Future Sci-Tech City, collected through several face-to-face interviews which have been conducted in spring-summer 2016. This area, 113 km^2 large, was previously covered by farmlands and recently is benefitting of dedicated-national policies to implement talents strategies, improve scientific and technology innovation and foster new entrepreneurship. The planning of this new territorial entity is producing different effects on the economy, on the environment and on the landscape, both positive and negative, and this generates different perception and understanding in the social system. The interviews were conducted using a composition of photos of the area. Images are inherently polysemic, but each of them poses the focus on a different -even controversial- aspects of the urban development of the area. This can be seen in Figure 1, in which the wordcloud shows the relationship between the most frequent words (stemmed with standard pre-processing textual analysis techniques) and the content of the photos used in the interviews. We then develop a Structural Topic Model using the photos of the interview as covariates in the model to extract the key topics of collected textual data. The introduction of covariates in the model is able to highlight if particular visual stimuli bring out specific perceptions or latent issues. The emerging categories of perception are presented in Figure 2, where each category represents an estimated topic described by its most frequent and exclusive words. From the results of this analysis we notice that the perception of the people interviewed is mostly of great appreciation for the great economic development, with some, but minor, concerns on the negative effects of this development on the society and on the environment. Future developments of this research will concern the estimation of the network of relationships among these categories with Probabilistic Graphical Models (PGMs). Next step of the research will the estimation of the network of relationships among these emerging categories by means of

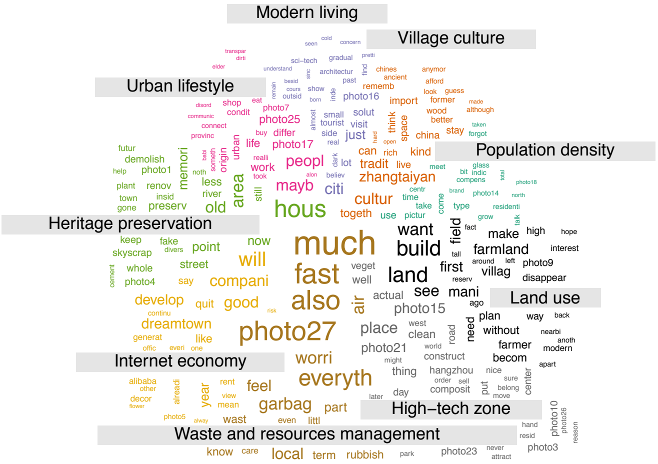
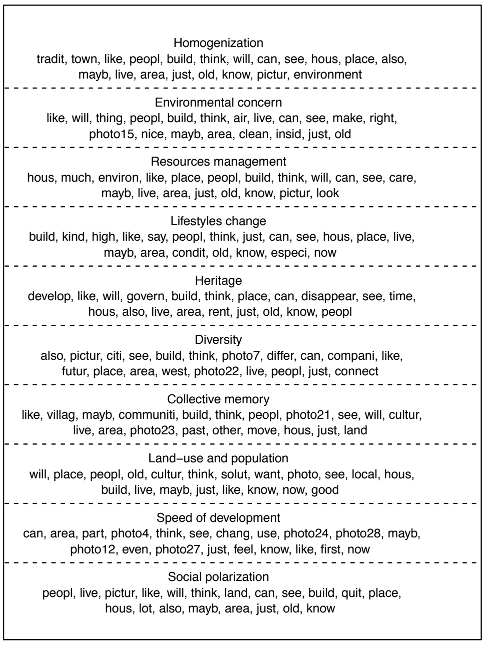


Fig. 1: Wordcloud of the most frequent words (vocabulary) associated with the content of the photos used in the interviews.

Probabilistic Graphical Models (PGMs). PGMs are in fact very efficient and effective statistical models to estimate the complex structures of probabilistic dependences and independencies which characterized complex social systems.

References

1. Zhao, P., Li, P.: Rethinking the relationship between urban development, local health and global sustainability. Current Opinion in Environmental Sustainability. 25, 14–19 (2017)
2. Yang, B., Xu, Y., Shi, L.: Analysis on sustainable urban development levels and trends in China' cities. Journal of Cleaner Production. 141, 868–880 (2017)
3. Riffat, S., Powell, R., Aydin D.: Future cities and environmental sustainability. Future Cities and Environment. 2:1 (2016)
4. Wu, F.: Emerging Chinese Cities: Implications for Global Urban Studies. The Professional Geographer. 68(2), 338–348 (2016)
5. Dolfin, M., Leonida, L., Outada. N.: Modeling human behavior in economics and social science. Physics of Life Reviews. In press (2017)
6. Blei, D.M.: Probabilistic topic models. Communications of the ACM. 55(4), 77–84 (2012)
7. Grimmer, J., Stewart, B.: Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents. Political Analysis. 21(3), 267–297 (2013)
8. Li, G., Feng, S., Jun, T.: Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science. 2(3): 153–170 (2016)



Homogenization
tradit, town, like, peopl, build, think, will, can, see, hous, place, also, mayb, live, area, just, old, know, pictur, environment

Environmental concern
like, will, thing, peopl, build, think, air, live, can, see, make, right, photo15, nice, mayb, area, clean, insid, just, old

Resources management
hous, much, environ, like, place, peopl, build, think, will, can, see, care, mayb, live, area, just, old, know, pictur, look

Lifestyles change
build, kind, high, like, say, peopl, think, just, can, see, hous, place, live, mayb, area, condit, old, know, especi, now

Heritage
develop, like, will, govern, build, think, place, can, disappear, see, time, hous, also, live, area, rent, just, old, know, peopl

Diversity
also, pictur, citi, see, build, think, photo7, differ, can, compani, like, futur, place, area, west, photo22, live, peopl, just, connect

Collective memory
like, villag, mayb, communiti, build, think, peopl, photo21, see, will, cultur, live, area, photo23, past, other, move, hous, just, land

Land–use and population
will, place, peopl, old, cultur, think, solut, want, photo, see, local, hous, build, live, mayb, just, like, know, now, good

Speed of development
can, area, part, photo4, think, see, chang, use, photo24, photo28, mayb, photo12, even, photo27, just, feel, know, like, first, now

Social polarization
peopl, live, pictur, like, will, think, land, can, see, build, quit, place, hous, lot, also, mayb, area, just, old, know

Fig. 2: Estimated topics with the emergent categories of perception.

9. Tvinnereim, E., Liu, X., Jamelske, E.M.: Public perceptions of air pollution and climate change: different manifestations, similar causes, and concerns. Climatic Change 2016.1–14 (2016)
10. Reich, J., Tingley, D., Leder-Luis, J., Roberts, M.E., Stewart, B.M.: Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. Journal of Learning Analytics. 2(1), 156–184 (2015)
11. Roberts, M.E., Stewart, B.M., Tingley, D., Airoldi, E.M.: The Structural Topic Model and Applied Social Science. Neural Information Processing Society (2013)
12. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science. 58(4),1064-1082 (2014)
13. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. Journal of the American Statistical Association. 111(515), 988-1003 (2016)

A genetic approach to the calibration of selected dynamic factor models for macroeconomic forecasting

Fabio Della Marra[1]

Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University of Venice,
Cannaregio 873, 30121
Venice, Italy
fabio.dellamarra@unive.it

Abstract

In this work, a comparative analysis of the forecasting performance of three Large-Dimensional Dynamic Factor Models is presented. As a key feature, Dynamic Factor Models represent each variable in a dataset as the sum of two orthogonal terms: a *common component* χ_t , driven by a reduced (as compared to the number of series in the dataset) number of common factors, and an *idiosyncratic component* ξ_t , which represents measurement errors or local features. Among the different versions of the Dynamic Factor Models we selected:

(i) *SW model*: this time-domain method was introduced in [11,12]. The factors are estimated by computing static principal components of the variables in the dataset. Let y_{it} be the variable of the dataset to be forecasted at time t, its h-step-ahead prediction equation (also called *Diffusion Forecast Index*) is obtained by regressing y_{it+h} on the factors and on y_{it} itself. Lags of the factors and of y_{it} may be added.

(ii) *FHLR model*: this frequency-domain method was proposed in [7,8] and requires the computation of two steps. In a first step, the common component χ_t, the idiosyncratic component ξ_t and their covariances are estimated using a frequency-domain method introduced in [7] named Dynamic Principal Component. In the second step, the factors are estimated by computing Generalized Principal Components.

(iii) *FHLZ model*: this frequency-domain method was proposed in [9,10] Here, the underlying assumption in (i) and (ii) that the common components span a finite-dimensional space as n tends to infinity is relaxed.

There exists some literature comparing the forecasting performances of SW and FHLR, but universal consensus still does not seem to have been reached. Theoretically, time-domain methods consider only relations among the variables at the same time, whereas frequency-domain methods exploit leaded and lagged relations among the variables. However time-domain methods require less parameters to be calibrated. Hence they are more robust to misspecification than

frequency-domain methods. Instead, a systematic comparison of the forecasting performances of SW, FHLR and FHLZ can be found only in [2,3]. In this work, the EU dataset is the same employed in [2]. This dataset is split into two subsamples. To guarantee balancedness, time series with missing data are discarded. The former, from February 1986 to December 2000, is used to calibrate the models, i.e. to produce in-sample forecasts of the variables of the EU dataset for several specifications of SW, FHLR and FHLZ. Then, for each class of models, we select the specification which shows the minimum mean square forecast error (MSFE). These models are then run and compared in the remaining sample, from January 2001 to November 2015. Instead, the US dataset employed in our exercise is accurately described in [3]. This dataset is split into two subsamples. The former, from March 1960 to December 1984, is used to calibrate the models. Then, for each class of models, we select the specification which shows the minimum mean square forecast error (MSFE). These models are then run and compared in the remaining sample, from January 1985 to November 2014. Both dataset contain real variables (import/export price indexes, employment, Industrial Production) and nominal variables (money aggregates, consumer price indexes, wages), asset prices (stock prices and exchange rates) and surveys. To achieve stationarity, several series are deseasonalized and transformed. No treatment for outliers is applied. In addition to SW, FHLR, FHLZ, the forecasts of an autoregressive process (AR) are computed. The order p of the AR process is determined in the calibration process. As in [1,12], to assess the forecasting performances, the variables which are taken into account are the level of the logarithm of the Industrial Production (IP) and the yearly change of the logarithm of the Consumer Price Index (CPI). Forecasts are computed h-months ahead, with $h \in \{1, 3, 6, 12, 24\}$. For each methods, we employ a rolling-window scheme $[t-l, t]$, whose size l is determined in the calibration sample. To assess the forecasting performance of each model, the mean-square forecast error (MSFE) is employed as a metric. Since each method is characterized by several parameters, an exhaustive exploration of the parameter space would be computationally infeasbile. Hence, we employ genetic algorithms to explore more efficiently the parameter space in the calibration sample of each dataset. At each epoque, the population of the genetic algorithm is a subset of the strings containing all the possibile configurations of the parameters. We set the fitness as the inverse for its MSFE. For each method, we iterate the genetic algorithm ten times on the calibration sample of the two datasets. The fitness of each individual is stored in a data structure. Eventually, for each method, we select as the most performing configuration the one endowed with the greatest fitness, regardless of the final solutions of the ten run of the genetic algorithms (as in). The convergence of each iteration of the genetic algorithms is graphically shown by plotting the boxplot of the results. These plots are not reported here. The forecasting performance of the three dynamic factor models over the IP and CPI are compared on the proper sample of each dataset. As in [2,3], to assess the forecasting performance of each couple of methods locally, each time series of the dataset is smoothed by a centered moving average of length $m = 61$ (with coefficients equal to $1/m$)

and then the Fluctuation test ([4]) is run, at 5% significance level. The results for the IP are not reported here. As to the EU dataset, all methods outperform AR significatively from the crisis on (which, according to CEPR, starts in April 2008). Globally, FHLR and FHLZ outperforms SW from the crisis on. As to the IP, FHLR tends to outperform FHLZ from the crisis on. Instead, as to the CPI, FHLZ tends to outperform FHLR from the crisis on, but evidencies are less significative. As in [2,3], this exercise has been extended to the other variables in the dataset. The results achieved are omitted here, but it can be seen that FHLR tends globally to outperform the other methods on the real variables and that FHLZ tends globally to outperform the other methods on the nominal variables. On the US dataset, all methods tend, instead, to lose ground against AR significatively during the Great Recession. FHLR tends globally to outperform the other methods on the real variables and that FHLZ tends globally to outperform the other methods on the nominal variables. In this paper, we have shown that FHLR globally outperforms the other methods on the real variables and that FHLZ globally outperforms the other methods on the nominal variables. As to EU dataset, [2] found similar results for the CPI, but mixed evidencies appeared for the IP. As to the US dataset, [3] found similar but less significative results. Hence, we have empirically shown that the calibration process plays a crucial role in these applications, since a more efficient exploration of the parameter space allowed us to empirically prove the superiority of frequence-domain dynamic factor models against time-domain factor models in a macroeconomic forecasting setting.

References

1. D'Agostino, A., Giannone, D.: Comparing alternative predictors based on large-panel factor models. Oxford Bulletin of economics and statistics. 74(2), 306 - 326 (2012).
2. Della Marra, F.: A forecasting performance comparison of dynamic factor models based on static and dynamic methods. Communications in Applied and Industrial Mathematics. 8(1), 44-66 (2017).
3. Forni, M., Giovannelli, A., Lippi, M., Soccorsi, S.: Dynamic factor model with infinite dimensional factor space: forecasting. CEPR discussion paper.
4. Giacomini, R., Rossi, B.: Forecast comparisons in unstable environments. Journal of Applied Econometrics. 25(4), 595-620 (2010).
5. Kapetanios, G.: Variable selection in regression models using nonstandard optimisation of information criteria. Computational Statistics and Data Analysis. 52(1). 4 - 15 (2007).
6. Kapetanios, G., Marcellino, G. M., Papailias F.: Variable selection for large unbalanced datasets using non-standard optimisation of information criteria and variable reduction methods. Quantf working paper (2014).
7. Forni, M., Hallin, M., Lippi, M., Reichlin, L.: The generalized dynamic factor model: Identification and estimation. Review of Economics and Statistics. 82(4), 540 - 554 (2000).
8. Forni, M., Hallin, M., Lippi, M., Reichlin, L.: The generalized dynamic factor model: One-sided estimation and forecasting. Journal of the American Statistical Association. 100, 830 - 840 (2005).

9. Forni, M., Hallin, M., Lippi, M., Zaffaroni, P.: Dynamic factor model with infinite dimensional factor space: representation. Journal of Econometrics. 185, 359 - 371 (2015).
10. Forni, M., Hallin, M., Lippi, M., Zaffaroni, P.: Dynamic factor model with infinite dimensional factor space: asymptotic analysis. EIEF working paper (2016).
11. Stock, J.H., Watson, M.W.: Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association. 97(460), 1167 - 1179 (2002).
12. Stock, J.H., Watson, M.W.: Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics. 20(2), 147 - 162 (2002).

Urban Evacuation Plan: a Simulation Study with Cognitive Agents in a Cellular Automata Context

Salvatore Di Gregorio

Dept. of Mathematics and Computer Science, University of Calabria, 87040 Rende, Italy
salvatore.digregorio@unical.it

Abstract. Here, a simple study is reproposed for modelling the evacuation of a population in a urban environment, considering a previous planning and the effects of information and messages during the emergency. A cellular space, representing the urban tissue, is inhabited by cognitive agents, which change both their position and their mental state, after a given interval of time. The agent changes his mental attitude activating rules of revision (particularly of believes) after having listened to messages and/or having observed events, where he is set. A range of agent personalities (confident, anxious, sceptic) related to the plan indications effects their behavior. Simulations of evacuation with different mixture of personalities point out the conditions for favorable outcomes.

Keywords: Pedestrian Evacuation, Cognitive Agents, Cellular Automata, Complex Systems, Modelling and Simulation

1 Introduction

Many Cellular Automata (CA) models of Evacuation of Pedestrian Agents were produced, but they regard mainly areas with dimensions of rooms, buildings, ships and so on, a large literature can be found in sessions of conferences, e.g., ACRI. Rare models for evacuation in larger spaces (Wang Li et ali., 2013) don't consider cognitive agents. This paper recuperates previous investigations (Di Gregorio et al., 2001) in order to specify a simple CA social model for evacuation with cognitive agents.

In a social simulation, agents have to be located in an explicitly represented space-time; a two-dimensional CA, which accounts for local interactions, is selected. There are different types of agent interactions. Message is a basic mechanism of interaction: a point-to-point communication, where two agents exchange messages in a direct manner, and a so called "broadcast communication" where an agent community receives (in the same instant) one message, e.g., by radio, or in terms of the effect of a physical phenomenon (smoke, earthquake) potentially perceptible by everybody. Such "external influences" are considered in a CA extension, MCA, the Macroscopic CA (Di Gregorio & Serra, 1999), that is here adopted and where the emergent macroscopic properties and the local microscopic ones coexist at the same level.

The considerable features of the model are specified in the next chapter, simulation results follow, some comments appear at the end.

2 Main characteristics of the Model

The CA is a two dimensions MCA with a Moore's neighborhood, sub-states account for the features of an agent (just one agent for cell); possible signals for prescribed walking direction, information from a broadcasting mechanism, are specified in terms of external influences. The elementary processes account for the agent movements with random priority rule in case of conflict according his cognitive state.

The classical cycle of an agent (Shoham, 1993) is considered, both an activity of reception and interpretation of messages on macroscopic events, and an activity of individuation of macroscopic emergent situations (crowd, danger for this model); the agent is apt to recognize similar situations, that are identified and recognized by an agent as macroscopic events ME_1, ME_2, … ME_f. With these new kinds of believes as a starting point, an agent takes decisions about the process of revision and behaviour (actions to perform). The evacuation plan P is specified as $P::=E_1 \wedge E_2 \wedge ... \wedge E_x \wedge ... E_n$, where E_f represents the final/exit event.

Here, the emergent macroscopic properties and the local microscopic ones coexist at the same level of description.

The principal elements that characterize the cycle in exam are the following ones:

The cognitive status S(t) of an agent A at the time t is given by the set of attitudes (**comm**i**t**ments, **bel**ieves, **int**entions and **goal**s) on facts/events from the world he belongs to: $S(t)::=[\cup_i cmt(A,E_i),\cup_j bel(A,E_j),\cup_k int(A,E_k),\cup_s goal(A,E_s)]_t$

The Time Grain ΔT is the interval necessary to an agent to perform his cycle of revision and corresponds to the step of MCA. A cycle generates the Ag world perception from his Cognitive State and Actions at step i,

R function (part of the CA transition function) represents the change of state of a cell (an agent) expressed as a set of revision rules responsible for the generation of agent cognitive states at time t+ΔT: $S(t+\Delta T)=R(S(t),\cup_f M_f(S(t)))$.

In order to have the possibility of distinguishing different typologies of behaviour in the simulation, we have individuated three classes of agents with a different personality: Confident agents Anxious agents Sceptic agents.

When an agent is generated in the simulation ambient, he is labelled with one of these personalities. A different personality label (particular values of α_1 and α_2) determinates a different behaviour in the agent in a situation of state change as described in the former paragraph. For example, an agent with a calm personality when a state of trust is generated, always follows the plan indications.

Table 1 specifies some rules concerning cognitive states, Table 2 the revision:

Table 1: Examples of cognitive states (trust, mistrust, indecision)

$trust(A,E_x,E_f,P) \leftarrow$	$mistrust(A,E_x,E_f,P) \leftarrow$	$indecision(A,E_x,E_f,P) \leftarrow$
$bel(A,E_x \in P,\alpha_1) \wedge$	$bel(A,E_x \in P,\alpha_1) \wedge$	$bel(A,E_x \in P,\alpha_1) \wedge$
$bel(A,E_f \leftarrow P,\alpha_2) \wedge$	$bel(A,E_f \leftarrow P,\alpha_2) \wedge$	$bel(A,E_f \leftarrow P,\alpha_2) \wedge$
$goal(A,E_f) \wedge$	$goal(A,E_f) \wedge$	$goal(A,E_f) \wedge$
$\alpha_1\ \alpha_2 >$ trust_theshold	$\alpha_1\ \alpha_2 <$ mistrust_theshold	$\alpha_1\ \alpha_2 <$ trust_theshold $\wedge$
		$\alpha_1\ \alpha_2 >$ mistrust_theshold

Table2: Example of revision rule

If the agent A is following the north direction and meets a street signal (or a radio communication), that indicates to follow the east direction on his way, he will choose the new direction on the basis of the information carried by the signal (radio).	rev(A,goal(A,dir(north))) ← goal(A,dir(north)) ∧ bel(A,exists(x_1,y_1,A)) ∧ bel(A,sign(street,x_1,y_1+1,east))) ∧ remove(A,goal(A,dir(north))) ∧ add(A,goal(A,dir(east)).

3 Simulation Results

The simulation scenario is an urban area, where pedestrian agents, in a situation of danger, move on the basis of an evacuation plan towards the exit gates starting from the portion of urban areas, where they have been generated (Dijkstra et alii, 2000). The following classes of elements are considered in a manner similar to Jiang (1999):

- an agent is represented by one cell, with a different color depending on the current cognitive state associated to the agent (green for the confident state; yellow for the indecision state; red for the mistrust one);
- walls are represented by sets of black cells aligned along the ambient borders;
- buildings, represented by sets of orange cells assembled in quadrangular blocks;
- agent generators, represented by blue cells localised on one side of the buildings;
- street signals, represented by either light green cells, light pink cells or dark pink depending on the direction shown;
- exit gates, represented by light green;
- buildings delimiting a set of streets, represented by block of grey cells.

There are two kinds of communication: a point-to-point communication, between agents moving in the territory, and a broadcast one between agents and the territory.

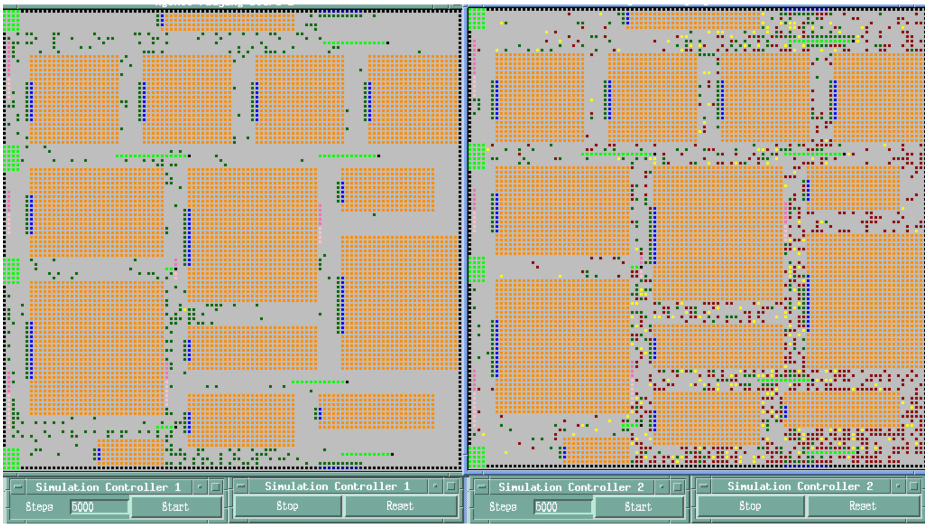


Fig.1: Step 5000 of the simulations

Two simulations were developed, where conditions differ only for the generations of the type of agents: their number is the same, but anxious and sceptic agents are generated more numerous in the second simulation.

The step 5000 of both the simulations is reported in Fig.1 for a quick comparison of the population distribution; on the left, evolution for people statistically more observant of the rules; on the right, people more anxious or sceptic. There is a difference of 887 agents between the former case (only 332) and the latter one (1219). More agents were able to reach the gates (safety) because of the greater observance of instructions, while dramatic jam situation are visible in the simulation on the right.

The number of agents was almost the same (a difference of 3 agents) in the two cases at the step 60, because of initial conditions with few agents produce rarely conditions of mistrust or indecision also in the latter case.

4 Comments

A large part of this work is recovered by ideas and investigations of Francesco Mele and Giovanni Minei. The model was applied for an ideal and simple case in order to detect various emergence situations in an emergency case A first evaluation of the simulation is positive because those characteristics, foreseen from an intuitive viewpoint, emerge clearly. The next step would have been to simulate a simple and real evacuation exercitation and tuning opportunely the parameters of the model in order to verify how simulations could approach the real event, but security regulations didn't permit to receive sufficient data. A further evolution in this research will mean an increase of complexity, introducing more rules in order to refine the representation of both the cognitive (emotive) states of agents and their interaction with other agents and the urban territory, furthermore physical conditions of agents, effecting, by example pedestrian speed, have to be related with possible actions.

References

Di Gregorio S., Mele F., Minei G. (2001): "Automi Cellulari Cognitivi. Simulazione di evacuazione" Atti di "INPUT 2001" (in CDROM), Conferenza nazionale informatica e pianificazione urbana e territoriale, democrazia e tecnologia, 27-29 giugno 2001, Isole Tremiti.

Di Gregorio, S., Serra, R. (1999).: "An empirical method for modelling and simulating some complex macroscopic phenomena by cellular automata" Future Generation Computer Systems, Vol.16 2/3 pp.259-271.

Dijkstra, J., H.J.P. Timmermans and A.J. Jessurun (2000): "A Multi-Agent Cellular Automata System for Visualising Simulated pedestrian Activity" Theoretheical and Practical Issues on Cellular Automata - Proceedings on the 4th International Conference ACRI 2000, pp.29-36, Springer 2000, ISBN 1-85233-388-X.

Jiang, B. (1999): "SimPed: Simulating Pedestrian Flows in a Virtual Urban Environment" Journal of Geographic Information and Decision Analysis, Vol.3 1, pp.21-30.

Shoham Y. (1993): "Agent Oriented Programming", Artificial Intelligence, 60: pp.51-92.

Wang Li, Liu Mao, Meng Bo (2013) "Incorporating topography in a cellular automata model to simulate residents evacuation in a mountain area in China" Physica A 392 pp.520–528.

The impact of self-loops in random boolean network dynamics

Sara Montagna, Michele Braccini, and Andrea Roli

Dept. of Computer Science and Engineering
Campus of Cesena
Alma Mater Studiorum Università di Bologna, Cesena, Italy

Random Boolean Networks (RBN) have been successfully used as genetic regulatory network (GRN) models both for identifying generic properties of cell dynamics and for reproducing a specific (partial) genetic network reconstructed from biological data [5,1,2]. When generic properties are sought, the typical approach consists in studying ensembles of boolean networks generated according to a given, biologically plausible, model, such as the one proposed by Kauffman [3]. The most used model defines a RBN of n nodes according to the following rules: each node has exactly k inputs, randomly chosen among the other nodes; boolean functions are assigned to each node on the basis of the 2^k truth table entries, in which the entry is set to 1 with probability p (the *bias*). Variants of this model have also been considered, for example by restricting the set of boolean functions to canalising ones, or by imposing a scale-free topology. These variants are inspired by biological plausibility and are often suggested by the identification of crucial properties and mechanisms observed in GRNs reconstructed from biological data.

To the best of our knowledge, the impact of self-loops, which abounds in biological genetic networks, has not been studied in RBNs. Indeed, within a GRN, a self-loop models the property of a gene[1] producing some chemical substances that contribute maintaining the activation state of that gene. This mechanism is particularly evident in the differentiation process, where cells, from a stem state, choose a fate towards specific specialised cells. For example, from the Drosophila GRN shown in Figure 8 of [4], all the four main genes responsible for the patterning of gap genes expression during embryo development, are involved in autocatalytic reactions.

In this work we study the impact of self-loops in RBNs dynamics. The long term goal of this research is to identify basic mechanisms and crucial motifs of GRNs underlying fundamental cellular processes that can be modelled as structural and functional elementary bricks in BNs, thus making it possible to study generic properties of cell dynamics by means of ensembles of more realistic BNs models. Furthermore, this repertoire of bricks may be used inside algorithms for the automatic generation of BNs endowed with specific dynamical properties.

[1] For simplicity, here we suppose that one node in the GRN corresponds to one gene.

To explore the impact of the introduction of self-loops into a RBN, we modified a randomly generated boolean network in different ways:

1. add a self-loop and extend the truth table randomly (with the same bias used for generating the original RBN);
2. add a self-loop and change the node boolean function into an OR between the node value and the previous function;[2]
3. remove an incoming link and replace the input with a self-loop, without changing the node boolean function;
4. remove an incoming link and replace the input with a self-loop, and change the node boolean function into an OR between the node value and the previous function.

Preliminary results show that self-loops affect the number of attractors, their robustness and the overall attractor landscape in terms of Threshold Ergodic Sets [5] (TESs), which is defined by the transitions among attractors as a consequence of single-node temporary perturbations. An excerpt of the results (concerning variant no. 1, for RBNs with $n = 20$ and 10 nodes with self-loops) is shown in Figures 1, 2 and 3. We can observe that the number of attractors is higher in the networks with self-loops and that their robustness tends to be smaller than in classical RBNs. A notable result is that the TESs structure is richer in the RBNs with self-loops. In these experiments we have found that the results are qualitatively the same for all the variants considered for introducing self-loops. Results are not shown here for lack of space and only an excerpt is shown in Figure 4. In general, the effects observed are gradually more evident with increasing number of self-loops. Moreover, the impact is even more striking when the boolean functions are changed into an OR between the previous function and the node value involved in the self-loop. These results suggest to study the evolutionary advantage of having self-loops in genetic networks and to investigate what mechanisms counterbalance the effect of self-loops on attractor robustness, which is a fundamental property for modelling cell dynamics.

References

1. Balleza, E., Alvarez-Buylla, E.R., Chaos, A., Kauffman, S., Shmulevich, I., Aldana, M.: Critical dynamics in genetic regulatory networks: examples from four kingdoms. PloS one 3(6), e2456 (Jan 2008)
2. Chaos, Á., Aldana, M., Espinosa-Soto, C., de León, B.G.P., Arroyo, A.G., Alvarez-Buylla, E.R.: From genes to flower patterns and evolution: Dynamic models of gene regulatory networks. Journal of Plant Growth Regulation 25(4), 278–289 (Dec 2006)
3. Kauffman, S.A.: The origins of order. Oxford University Press (1993)
4. Montagna, S., Viroli, M., Roli, A.: A framework supporting multi-compartment stochastic simulation and parameter optimisation for investigating biological system development. SIMULATION (June 2015)
5. Villani, M., Barbieri, A., Serra, R.: A dynamical model of genetic networks for cell differentiation. PloS one 6(3), e17703 (Jan 2011)

[2] This choice is motivated by observing that most self-loops in GRNs have a self-activating effect.

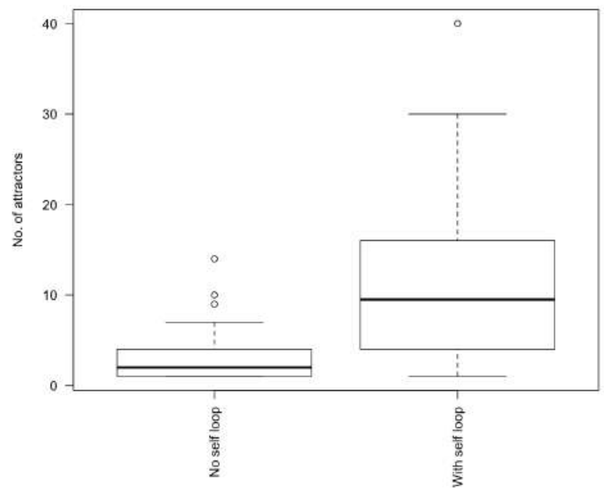


Fig. 1: Comparison of the number of attractors. Statistics are taken across 30 different RBNs.

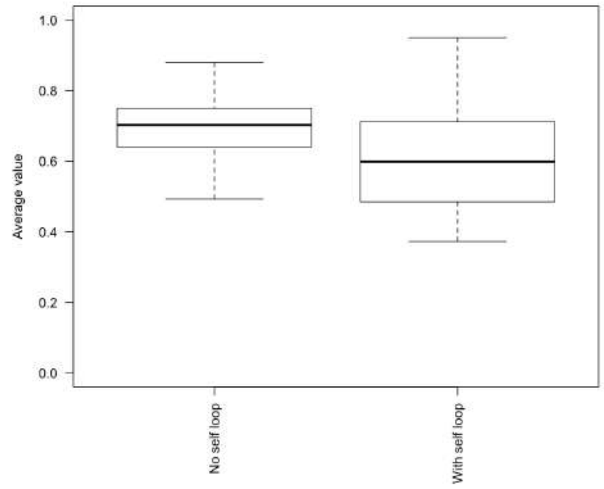


Fig. 2: Comparison of the robustness of attractors computed as the average fraction of times in which after a temporary flip on one node the trajectory returns to the same attractor. Statistics are taken across 30 different RBNs.

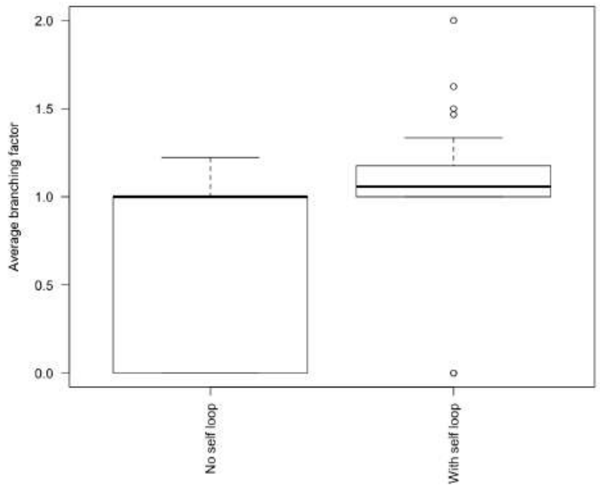


Fig. 3: Comparison of the average branching factor of TES differentiation trees [5]. The highest the branching factor, the richer the differentiation potential of the BN. Statistics are taken across 30 different RBNs.

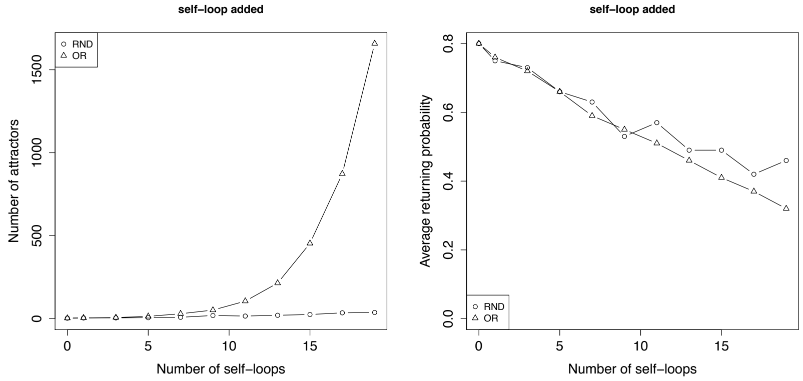


Fig. 4: Average number of attractors (left) and average probability of returning to an attractor after a node flip (right) as a function of the number of self-loops added in RBNs originally with $k = 2$ ($n = 20$).

On the dynamical properties of a gene-protein model

Davide Sapienza[1], Marco Villani[1,2] and Roberto Serra[1,2]

[1] Department of Physics, Informatics and Mathematics, Modena and Reggio Emilia University
[2] European Centre for Living Technology, Ca' Foscari University, Venezia, Italy
{marco.villani,rserra}@unimore.it

Keywords: gene-protein model, generic properties, memory effect, dynamical regimes.

1 Introduction

Random Boolean models of genetic regulatory networks (RBNs) are very well-known [10][11] and, in spite of their long age [9], they still provide useful descriptions of important observational and experimental results [12][13][14][5]. A major limitation of the classical RBN model is its synchronous updating: from a physical viewpoint, this amounts at assuming that all the proteins are synthesized at the same rate, and that all the proteins decay at equal rates; this unrealistic assumption allows one to write the gene activation pattern at time t+1 as a function of that pattern at time t, forgetting the previous history. Asynchronous updating has been taken into account, but also this would lead to difficult interpretations. Other interesting "intermediate" update strategies have also been proposed [2].

Some properties of RBNs are robust with respect to the updating strategy, but in general there is no guarantee that this is the case. We have been particularly interested in the response of genetic networks to perturbations like gene knock-out and we have shown that, if the RBN model is chosen, the distribution of avalanches in gene expression levels in S. Cerevisiae that follows a single knock-out provide information about the dynamical regime of the biological network [13][5]. This result is particularly relevant, given the importance of the "criticality hypothesis", which states that biological systems should preferentially be found in dynamically critical states. However, issues of dynamical criticality should be addressed in a way that does not depend upon not quite realistic assumptions.

In order to do so, while retaining the simplification related to the use of Boolean variables and to the random spirit of RBNs, we introduced the GPBN model (Gene-Protein Boolean Network), where the network is composed by two different sets of nodes, labelled G and P with reference to genes and proteins. There is a one-to-one correspondence between genes and proteins. Both kinds of nodes take Boolean values: the state at time t+1 of a G node depends upon the state of a fixed set of P nodes at same time, while the state at time t+1 of a P node depends upon the state of its corresponding G node at time t. Once a P node is set active (its state is 1), it remains active for at least a fixed number of steps. If a new activation signal comes in before decaying, the counter is reset. If no activation signal arrives, the P node is set to 0 at the end of its "lifespan". If we look at the genes only, it is like adding some memory terms, so the new state of the network is no longer "Markovian", i.e. it does no longer depend upon the previous state only.

This model has been thoroughly studied in a PhD project and its properties have been described elsewhere [6][7]. In those papers the usual definition of dynamical criticality, based on the value of the so-called Derrida parameter, had been used. We have recently shown some limitations related to the use of that single measure to characterize critical states in RBNs [15]. This prompted a more thorough analysis of the dynamics of non Markovian systems, whose main features are presented in this paper together with some results concerning its application to the GPBN model.

2 The GPBN model

A GPBN model [7][6][8] is a bipartite oriented graph containing two types of Boolean nodes: the G nodes, which represent the genes set, and the P nodes, which represent the set of proteins set (or, in general, gene products). A G node can be active or inactive (producing or not its protein), whereas a P node describes the presence (or absence) of a protein within the system. There are two types of links: a *synthesis link*, which goes from a G node to only one P node, and a *transcriptional regulation link*, from a P node to one or more G nodes.

Each G node, say the j-th, produces its protein when active (synthesis link) and a G node is driven by the action of 2^k inputs (k being the number of its transcriptional regulation links, coming from P nodes), according to a fixed Boolean function f_j of associated to it.

Each P node, say the i-th, has a *decay time* dt_i, representing its discrete-length lifespan, once activated. In the implementation of the model we consider a counter, called *decay phase* h_i, that starts from dt_i and ends in 0. When the incoming G node is active, then the corresponding P node resets its counter to the decay time. Until the decay phase vanishes, the P node have a regulation role on its outgoing links. Each decay time is taken with uniform probability between 1 and a parameter defined as *maximum decay time* (MDT); note that when MDT is equal to 1 the GPBN framework is identical to that of RBN. If the value of a G node is 1 at time t then the value of corresponding P node will be 1 at time t+1 and its decay phase is set to dt_i, otherwise the decay phase of the P node is decremented by one unit (in case of dt_i=0, the activation of P is set to 0). On the other hand, the value of the G node at time t is immediately determined by its function f_j, which depends on the states of its incoming P nodes at time t.

3 Dynamical regimes

The RBN framework can support different dynamical regimes [10] [11] [1], usually identified by means of the so-called Derrida parameter λ. Essentially, this index measures the tendency of a temporary perturbation to vanish, to persist or to spread through the entire system: so, ordered, critical and (pseudo)chaotic dynamical regimes correspond respectively to $\lambda<1$, $\lambda\approx1$ and $\lambda>1$ [3] [4].

In GPBN systems the initial perturbation could affect G nodes, P nodes, or both. In our approach a perturbation of a P node can correspond either (i) to an activity change from 0 to 1, with a decay phase h_i uniformly chosen within the range [1, dt_i] or (ii) to an activity change from 1 to 0, with h_i=0. A perturbation of a G node can correspond

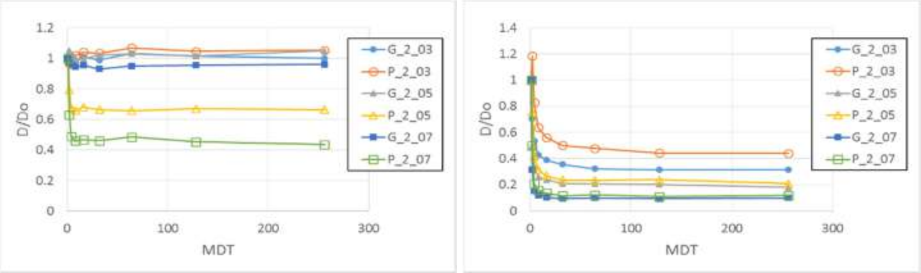
(i) to an activity change from 0 to 1, followed by the appropriate effect on the protein or (ii) to an activity change from 1 to 0 – in this case, the G node is not producing its protein, and the P node reduces its decay phase by one

The Derrida approach is based upon perturbations of some initial states: in the full paper we will describe the choice we made in order to tailor this approach (called the "extended" Derrida approach) to GPBN systems.

4 Results

We choose the GPBN ensembles in such a way that – in case of MDT=1 – the corresponding RBNs show ordered, critical or chaotic behavior. We are interested in observing the influence of the memory of the GPBN system (measured by the MDT parameter) on its dynamical regimes – whose presence is monitored by the extended Derrida approach.

In Fig.1 the effect of perturbing a P node or a G node on the set of genes or on the set of proteins are shown. So, by varying the MDT parameter, it is possible to observe that the memory has different effects on different kinds of nodes. Indeed, if we perturb a P node (Fig.1a) the G nodes are not affected by the MDT parameter, while the P nodes rapidly shows ordered behaviors (with a saturation effect). On the other hand, if we perturb a G node (Fig.1b) both G and P nodes rapidly show ordered behaviors, with a saturation effect. So, the memory has different effects in case of information transmission from G to P nodes or from P to G nodes. But, in order to measure the "global" effect of memory on the GP systems we measure the robustness to perturbations of the attractors of the system, finding that an increase of memory length leads to an increase in order – an effect that however rapidly saturates.



(a) (b)

Fig. 1. The effect of perturbing a protein node (a) or a gene node (b) on genes (G) or protein (P) nodes. All Derrida measures are normalized on this case (whence the symbol D/Do). The network ensemble in this figure has k=2 and bias={0.3, 0.5, 0.7}, corresponding to ordered, critical and ordered systems in case of MDT=1.

5 Conclusion

The GPBN model is a model of genetic regulatory systems that maintains the abstraction level of the RBN framework and that at the same time allows an explicit modelling of time delay effects. This kind of memory has different effects in case of information transmission from G to P nodes or from P to G nodes, and pose some interesting questions about the correct way of measuring of the system dynamical regimes through Derrida-like procedures. Anyway, the robustness of the system's attractors can constitute a sort of global measure related to its general "degree of order".

References

1. Aldana M., Coppersmith S., Kadanoff L.P. (2003) Boolean Dynamics with Random Couplings in Perspectives and Problems in Nolinear Science Kaplan E., Marsden J.E., Sreenivasan K.R. Eds. Springer-Verlag New York pp. 23-89
2. Darabos C., Giacobini M., Tomassini M. (2010) Generalized Boolean networks: how spatial and temporal choices influence their dynamics. Handbook of Research on Computational Methodologies in Gene Regulatory Networks (pp. 429-449). IGI Global, Hershey, PA 17033, USA.
3. Derrida B., Pomeau Y. (1986) Random networks of automata: a simple annealed approximation. Europhys. Lett. 1, 1(2):45-49.
4. Derrida B., Weisbuch G. (1986) Evolution of overlaps between configurations in random Boolean networks. J. Physique, 47:1297-1303
5. Di Stefano, M.L., Villani, M., La Rocca, L., Kauffman, S.A., Serra, R.: Dynamically Critical Systems and Power-Law Distributions: Avalanches Revisited. In: Rossi, F., Mavell, F., Stano, P., and Caivano, D. (eds.) Advances in Artificial Life, Evolutionary Computation and Systems Chemistry. pp. 29–39. Springer International Publishing (2016).
6. Graudenzi A., Serra R., Villani M., Damiani C., Colacci A., Kauffman S.A. (2011a) "Dynamical properties of a Boolean model of gene regulatory network with memory" Journal of Computational Biology v.18 Mary Ann Liebert, Inc., publishers, NY
7. Graudenzi, A., Serra, R. (2009): A new model of genetic network: the gene-protein network. In R. Serra, I. Poli & M. Villani (eds): Artificial Life and Evolutionary Computation: 283-291
8. Graudenzi, R. Serra, M. Villani, A. Colacci, S.A. Kauffman (2011b) "Robustness analysis of a Boolean model of gene regulatory network with memory" Journal of Computational Biology v.18, n.4 Mary Ann Liebert, Inc., publishers, NY
9. Kauffman, S.A.: Homeostasis and Differentiation in Random Genetic Control Networks. Nature. 224, 177–178 (1969).
10. Kauffman S. A., (1993). The Origins of Order: Self Organization and Selection in Evolution. Oxford University Press.
11. Kauffman S. A., 1995. At Home in the Universe. New York, Oxford University Press
12. Serra, R., Villani, M., Semeria, A.: Genetic network models and statistical properties of gene expression data in knock-out experiments. J. Theor. Biol. 227, 149–157 (2004)
13. Serra, R., Villani, M., Graudenzi, A., Kauffman, S.A.: Why a simple model of genetic regulatory networks describes the distribution of avalanches in gene expression data. J. Theor. Biol. 246, 449–460 (2007).
14. Shmulevich, I., Kauffman, S.A., Aldana, M.: Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc. Natl. Acad. Sci. U. S. A. 102, 13439–44 (2005).
15. Villani M., Campioli D., Damiani C., Roli A., Filisetti A., Serra R. (2017) Dynamical regimes in non-ergodic random Boolean networks Natural Computing, June 2017, Volume 16, Issue 2, pp 353–363

Threshold ergodic sets vs. stochastic simulation of noisy boolean networks: comparison of two approaches for modelling cell differentiation

Michele Braccini[1], Andrea Roli[1], Marco Villani[2], and Roberto Serra[2]

[1] Dept. of Computer Science and Engineering
Alma Mater Studiorum Università di Bologna
[2] Dept. of Physics, Informatics and Mathematics
Università di Modena e Reggio Emilia &
European Centre for Living Technology, Venice

Cell differentiation is the process whereby a cell undergoes a *cell type* change, from less specialised to more specialised types. This process is at the roots of several crucial biological phenomena, such as morphogenesis, and models of cell differentiation may also help understand the dynamics of severe diseases, like e.g. cancer. Cell differentiation processes are characterised by highly complex dynamics.

Recently, a cell differentiation model based on boolean networks [2,3] subject to noise has been proposed [5,6]. The model is based on boolean networks (BNs), which are a prominent example of complex dynamical systems and have been introduced by Kauffman [2,3] as a genetic regulatory network (GRN) model. The noisy version of this model reproduces the main abstract properties of cell differentiation, such as the attainment of different degrees of differentiation, deterministic and stochastic differentiation, reversibility, induced pluripotency and cell type change [6]. This differentiation model considers BNs subject to noise, such as the transient flip of a node value. Attractors of BNs are unstable with respect to noise even at low levels. In fact, even if the flips last for a single time step one sometimes observes transitions from that attractor to another one. Here we are interested in the asymptotic behaviours of a noisy BN. In such networks, in general the higher this noise, the higher the probability to move across attractors. High levels of noise correspond to pluripotent cell states, where the BN trajectory can wander freely among the attractors; conversely, low levels of noise induce low probabilities to jump between attractors, thus representing the case of specialised cells. A fundamental role in the model is played by *threshold ergodic sets* (TES$_\theta$) which are sets of attractors in which the dynamics of the network remains trapped, under the hypothesis that attractor transitions with probability less than threshold θ are not feasible.[3] The transitions between attractors and their probabilities are summarised in the *attractor transition matrix* (ATM), which is computed by counting the number of transitions from an

[3] This hypothesis is supported by the observation that cells has a finite lifetime, which enables their dynamics to explore only a portion of the possible attractor transitions.

attractor to another one by applying a transient flip to each node in each state of every attractor. Starting from the ATM, a differentiation tree is built by iteratively removing the entries of the ATM with values less than a threshold θ, which is progressively increased from 0 to 1.[4]

The generic abstract properties of the model have been already shown to match those of the real biological phenomenon under these specific hypotheses on the noise affecting the cell. However, this process producing a TES-based differentiation tree is based on operations on a transition matrix and captures a static global picture summarising all the possible outcomes of the BN dynamics that may happen under this specific noise mechanism. As these results might differ from those that can be obtained by simulating a BN subject to random external perturbations, we compared the properties of the ATM/TES model with the actual dynamical simulation of the BN with the aim of assessing to what extent the two approaches show comparable results and their respective strengths and weaknesses.

Experimental setting

The boolean networks used in the experiments have $n = 100$ nodes and $k = 2$ distinct inputs per node assigned randomly (self-loops are not allowed). Boolean functions have been set by assigning a 1 in the node truth table so as to attain exactly a frequency of 0.5 across all the truth tables (for $k = 2$, this corresponds to the critical value [1]). The BN is subject to a synchronous dynamics, i.e. all nodes update their state in parallel and functions are applied deterministically. On the basis of plausible hypotheses on the time needed to create and destroy proteins, we set to 5×10^4 the number of steps for a BN run. The only stochastic component reside in the noise, which has been simulated as a temporary flip of the value of a node applied with probability ν; hence, at each step of the temporal evolution of the network, νn nodes are flipped on average. We ran experiments with ν so as to have on average one flip every τ steps, with $\tau \in \{1, 5, 10, 15, 20, 50, 100, 200, 500, 10^3, 5 \times 10^3, 10^4, 2 \times 10^4, 5 \times 10^4\}$. In the following, we will denote the corresponding noise probabilities as ν_τ. This noise mechanism emulates possible temporary fluctuations in the expression level of genes and may occur both during stationary phases (i.e. along attractors of the BN) and transients. We run experiments with 30 random BNs; for each of them the ATM was computed according to the procedure by Paroni et al. [4] and the associated differentiation tree (TES-tree) was built by progressively removing entries from the ATM. The time evolution of each BN was also simulated 100 times, starting from a random initial state (hereinafter, we will call a *story* any of these runs). We collected the trajectories of the BNs and computed statistics on the compatibility between the stories and the TES-tree, besides other ancillary statistics on the overall dynamics of the BNs.

[4] See [5,6].

Results

In this section we provide a sketch of the main results obtained. A typical TES-tree is depicted in Figure 1. The comparison between TES-trees and simulations with stochastic noise is mainly based on counting the transitions between attractors that are observed in the stochastic simulation but that are not allowed by the ATM, given a probability threshold θ. For each value of ν_τ, we counted the incompatibilities observed in all the 100 stories w.r.t. the lowest non-zero value of θ (level 1 of the TES-tree) and the highest one, where all TESs are single attractors (level n of the TES-tree). As expected, the higher ν_τ, the higher the number of these incompatibilities. Moreover, this increases with θ. Despite the discrepancy which is apparent at high noise levels, we observe that already for medium noise levels, i.e. not higher than ν_{200}, the incompatibilities are limited and tend to be negligible towards low noise levels.[5]

Complimentary detailed analyses on the dynamics of the BNs (not shown in this abstract) support the statement that there exists a significant noise level under which the two models are in agreement. Therefore, *(i)* under this threshold they can be both used to model differentiation phenomena—and their observations can be combined—and *(ii)* the new dynamic simulations may add interesting pieces of information on the heterogeneities of the possible individual configurations.

References

1. Bastolla, U., Parisi, G.: A numerical study of the critical line of kauffman networks. Journal of Theoretical Biology 187(1), 117 – 133 (1997)
2. Kauffman, S.A.: Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology 22(3), 437–467 (March 1969)
3. Kauffman, S.A.: The origins of order. Oxford University Press (1993)
4. Paroni, A., Graudenzi, A., Caravagna, G., Damiani, C., Mauri, G., Antoniotti, M.: CABeRNET: a cytoscape app for augmented boolean models of gene regulatory NETworks. BMC Bioinformatics 17, 64–75 (2016)
5. Serra, R., Villani, M., Barbieri, A., Kauffman, S., Colacci, A.: On the dynamics of random Boolean networks subject to noise: attractors, ergodic sets and cell types. Journal of theoretical biology 265(2), 185–93 (Jul 2010)
6. Villani, M., Barbieri, A., Serra, R.: A dynamical model of genetic networks for cell differentiation. PloS one 6(3), e17703 (Jan 2011)

[5] Plots are omitted for lack of space.

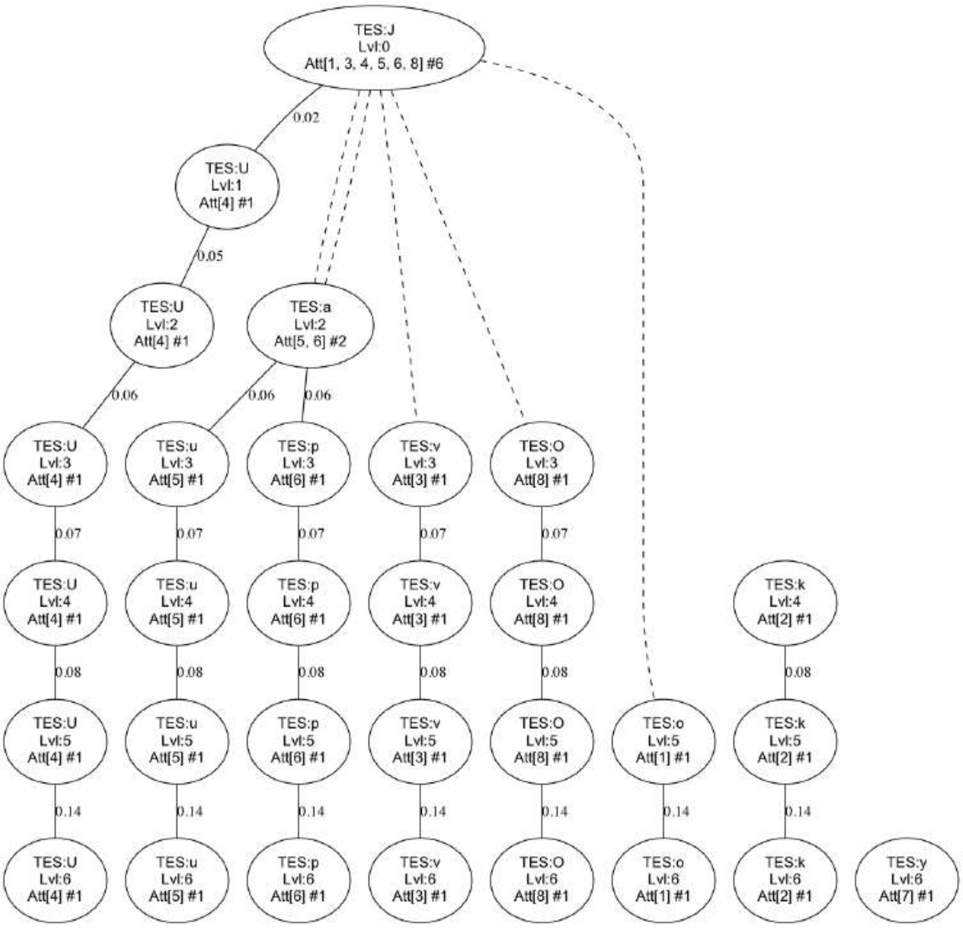


Fig. 1. An example of a TES-tree. Levels are numbered from 0, the topmost, to n, the lowermost; $n = 6$ in this example. Labels on the edges indicate the minimum threshold value at which any TESs of the previous level splits or reduces. Continuous lines denote paths along the differentiation tree that can be followed by increasing the threshold at minimum steps (these values are directly obtained by the ATM). Dashed lines denote the paths that can instead be followed if the threshold was increased by larger steps.

Simulating a population of protocells with uneven division

Martina Musa[1], Marco Villani[1,2] and Roberto Serra[1,2]

[1] Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia
[2] European Centre for Living Technology, Ca' Foscari University, Venezia, Italy
{marco.villani,rserra}@unimore.it

Abstract. Protocells should be similar to present-day biological cells, but much simpler. They are believed to have played a key role in the origin of life, and they may also be the basis of a new technology with tremendous opportunities. In this work we study the effect of uneven division processes on the synchronization of the duplication rates of protocells' membrane and internal materials.

Keywords: protocell, protocell populations, models, synchronization, replicators.

Protocells should be similar to present-day biological cells, but much simpler. They are believed to have played a key role in the origin of life, and they may also be the basis of a new technology with tremendous opportunities (see e.g. the books [1, 8] and further references quoted therein). Among various candidate protocell architectures, those based on lipid vesicles are particularly promising since they can spontaneously undergo fission, giving rise to two daughter protocells. Protocells should also contain a self-replicating set of molecules (the "replicators"): their composition should also affect the growth and replication rate of the container, so that some kind of competition can take place between cells with different chemical compositions

The simplest case is that of even division, where a vesicle splits into two identical daughter cells. In order to assure sustainable growth of a population of protocells, it is necessary that the duplication rate of the replicators be equal to that of the lipid container. It has been shown in a series of papers [1-7] that this synchronization takes spontaneously place, generation after generation, under a wide set of hypotheses concerning the protocell architecture and the kinetic equations for the replicators, and that it is robust with respect to random fluctuations. This is indeed a beautiful example of dynamical self-organization.

It has however also been observed that there are other ways in which lipid vesicles can divide. In this paper we consider a case (sometimes referred to as "budding") where the vesicle splits in two daughter vesicles of different sizes, a "large" one and a "small" one. Other types of division, like e.g. those due to extrusion processes, might also be investigated. Here we will suppose that, when a critical size has been reached, the protocell splits into two: the large daughter will inherit a constant fraction of both the lipid container and the replicators. Among various possible choices, we will assume that the key reactions, i.e. generation of new replicators and of new lipid membrane, take place inside the protocell aqueous phase (i.e. we will consider an "internal reaction model" according to the definition given in [5]). We will also assume that the shape of the mother protocell, as well as those of her daughters, are all spherical. As it has been discussed elsewhere, this implies that some part of the mother's internal aqueous environment is lost in fission, alongside with a fraction of replicators; the lost fraction can be calculated by simple geometric reasoning, if the concentration of the replicators is uniform in the internal water phase.

By adopting simplifications and hypotheses like those of [7], the simplified equations for the total quantities of lipid container C and replicator X during the continuous growth phase from an initial condition to the critical size θ are

$$\begin{cases} \dfrac{dC}{dt} = \alpha X \\ \dfrac{dX}{dt} = \eta X \end{cases} \qquad (1)$$

where it has also been assumed that the vesicle is thin, so the volume of the lipid membrane is approximately proportional to its surface (by renormalizing time, it has already been verified in the case of even division that the onset of synchronization is not affected by this simplification, although the kinetics is different [7]).

We will assume that a protocell splits into two daughters when its membrane reaches a certain threshold θ. After splitting, one of the daughter cells inherits a fraction ω of the lipid container, while the other one inherits 1-ω. Part of the replicators is lost in fission, as mentioned above, while the remaining part is also shared between the two daughters, which inherit Lω and L(1-ω) respectively, L being the fraction of x that is not lost in the bulk.

These rules determine the initial conditions of the two daughter cells at the next generation. The small one will need a longer time to reach the critical size and to undergo fission, while the larger one will be faster.

The continuous growth described by Eq. 1, starting from an initial condition where $C=C_i$ and $x=x_i$ up to the time $T=T_{div}$ when $C=\theta$ (i.e. when splitting takes place) can be analytically determined to be

$$\begin{cases} T_{div} = \dfrac{\ln\left(\dfrac{\eta(\theta - C_i)}{\alpha x_i} + 1\right)}{\eta} \\ x_f = \dfrac{\eta(\theta - C_i)}{\alpha} + x_i \end{cases} \qquad (2)$$

where x_f is the quantity of replicator at splitting time.

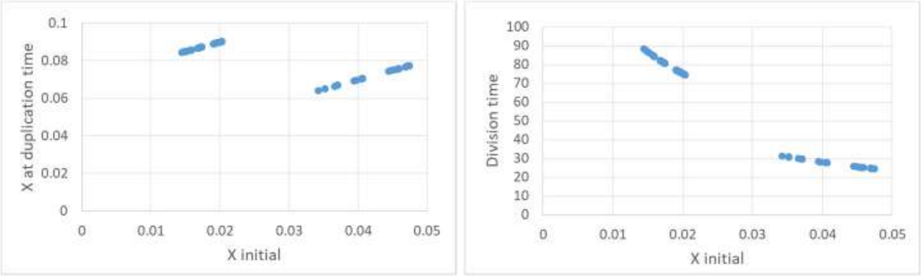
Of course, a large daughter cell will also give rise to another large one, etc., and this “pure lineage” of large cells will tend to synchronize, in a way similar to the case of even division. The same will happen for the pure lineage of small cells, although with a different division frequency. But as generations increase, the fraction of cells belonging to the pure lineages will decline, and most cells will have both large and small cells among their ancestors. Indeed, after k generations the large cells will be k, while the total number of cells will be 2^k, so the fraction of pure lineage cells will vanish in the long k limit.

An interesting question then concerns the distribution of fission times after several generations: will there be, on average, a uniform distribution of fission events in time, or will there be some pace, at population level, in the fission processes?

We have therefore simulated the growth of such populations of monoclonal protocells, originated from a single protocell. Because of computational limits (which anyway mimic real physical constraints) when the population size reach its maximum value each division implies the substitution of an already “born” protocell (stable population).

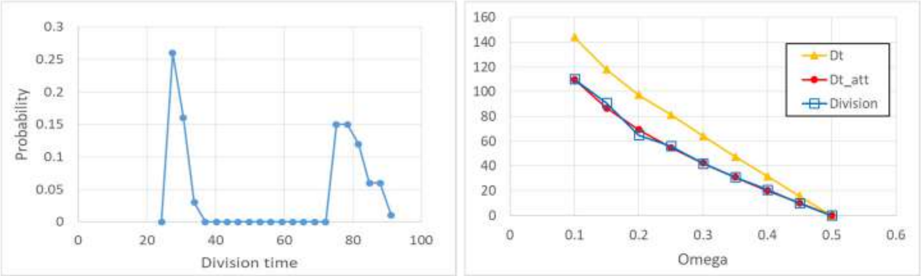
The data are noisy, of course. However, an interesting observation is that the data concerning both the fission intervals and the values of the replicators before fission are divided in two groups and they do not become homogeneous (Fig.1). As it can be expected, the division time is smaller in the case of the larger initial vesicles with a larger initial quantity of replicators (Fig. 1b). On the other hand, the smaller vesicles, with longer division times, synthesize a larger final quantity of replicators (Fig. 1a). The bimodality of the distribution of division times can also be directly observed in Fig.2, which shows their probability distribution.

It is also possible to analyze the difference between division times as a function of ω (note that, given the geometrical hypotheses, ω determines also the fraction L of replicators that are not lost). Let us define the “theoretical distance” between division times as the difference between the division times of the two pure lineages, which can be determined analytically. Surprisingly enough, in the case of uneven division one sometimes observes that the actual difference between division times sometimes exceeds this value, as shown in fig.3. Indeed, the theoretical distance closely approximates the zero-level plateau of fig.2, while some data show larger differences.



(a) (b)

Fig. 1. (a) Quantity of replicators before division and (b) division time (i.e. the interval between two successive divisions) vs initial quantity of replicators.



(a) (b)

Fig. 2. (a) Probability distribution of division times observed in simulations. (b) Difference between the observed and the "theoretical" values for the maximum difference between division times, vs ω. Dt is the measured maximum distance in division times, Dt_att is the "theoretical distance" defined in the text, Division is the width of the zero-level plateau of part (a).

Therefore, simulations and analytical computations show that the synchronization processes maintain their effectiveness also in case of uneven division. On the other hand, the budding process induces within the protocell population a stable bimodality on division times and quantity of replicators before division. Further works will explore the effects of splitting at not constant threshold θ.

References

1. Serra, R. & Villani, M. (2017, in press). Modelling protocells. Heidelberg: Springer
2. Serra, R. (2014): The complex systems approach to protocells. In C.Pizzuti & G.Spezzano Eds. Advances in Artificial Life and Evolutionary Computation, Communications in Computer and Information Sciences 445, Springer
3. Villani, M., Filisetti, A., Graudenzi, A., Damiani, C., Carletti, T., Serra, R. (2014): Growth and division in a dynamic protocell model. Life 4, 837-864
4. Filisetti A., Serra R., Carletti T., Villani M. & Poli I., (2010): Non-linear protocell models: synchronization and chaos. Europhys. J. B 77, 249-256
5. Carletti, T., Serra, R., Poli, I., Villani, M. & Filisetti, A. (2008): Sufficient conditions for emergent synchronization in protocell models. J. Theor. Biol. 254: 741-751
6. Filisetti, A., Serra, R., Carletti, T., Poli, I. & Villani, M. (2008): Synchronization phenomena in protocell models. BRL: Biophysical Reviews and Letters 3 (1/2), 325-342
7. Serra, R., Carletti, T. & Poli, I. (2007): Synchronization phenomena in surface reaction models of protocells. Artificial Life 13, 1-16
8. Rasmussen, S., Bedau, M.A., Chen, L., Deamer, D., Krakauer, D.C., Packard, N.H., Stadler, P.F. Eds. Protocells. The MIT Press, Cambridge MA (2008).

Computing attractors of asynchronous genetic regulatory networks

Marco Pedicini[1], Maria Concetta Palumbo[2], and Filippo Castiglione[2]

[1] Department of Mathematics and Physics, Roma Tre University
marco.pedicini@uniroma3.it
[2] CNR - Institute for Applied Computing “M. Picone”, Rome, Italy

Abstract. It is useful, in the field of theoretical biology, to study the dynamics of gene regulatory networks. Most of the models use the synchronous update schedule while reality is far from being so. The asynchronous update carries the computational burden of computing all possible updates at each single instant. In the present work, we describe a method that tries to solve the problem with minimal computational effort.

keywords: SAT Solver, discrete dynamical systems, Tarjan’s algorithm, gene regulatory networks

A Gene Regulatory Network (GRN) can be regarded as a discrete dynamical system with a transition function $T : S \to S$ from state to state which is determined from activation/inhibition dependencies between genes, transcription factors or RNA molecules. In its simplest form, where the number of states is $q = 2$, representing genes that are either activated or silent, this function is represented as a Boolean expression. In this and other more complicated cases (i.e., $q > 2$) T is a special form of a polynomial and the system can be considered Polynomial Dynamical Systems on Finite Fields $\mathbb{F}_q$ [VCL12].

In any case, the state space size of the generated dynamics is exponential ($N = q^n$) in the number of genes n and any exhaustive method for the investigation of its structure rapidly becomes intractable even for small n. In the special case of asynchronous use of the updating rule, things are more difficult, in the sense that the structure of the state space is even more complex.

In this work, we consider “biologically relevant” conditions such that the state space structure of attractors can be reconstructed when it is possible to generate transition paths that avoid to cross previously determined limit cycles. Despite the exponential size of the state space in the number of genes, we achieve a complexity bound which corresponds to Tarjan’s algorithm for determining *strongly connected components* (SCC) in a directed graph, in time $o(N + M)$ where N is the number of states (i.e., nodes in the graph) and M is the number of transitions between states (i.e., edges in the graph), [Tar72]. This can be attained because we never explicitly compute the transition graph. Instead, by following the approach in [DT11], we generate transition paths which avoid nodes belonging to previously-discovered cycles. In this way, if conditions on the structure of SCCs are respected, we can describe a variant of Tarjan's algorithm which has a polynomial SAT-complexity bound on the number of genes n. By SAT-complexity we mean that we count any call to the SAT-solver at unitary cost: this is the same situation assumed by Dubrova in the case of synchronous dynamics, where the main-loop iterates on the limit cycles of the network. Nevertheless, at each iteration a call to the SAT-solver is performed, and from the complexity point of view, this call could generate an exponential cost. Therefore, strictly speaking the procedure cannot be said to have polynomial complexity but we can measure the complexity of the procedure in terms of calls to the SAT-solver and also in terms of the fraction of the phase space that is visited in order to determine the structure of the SCCs.

The algorithm is inspired to those of Tarjan for finding the SCCs and combines a SAT-solver in the same way as Dubrova [DT11]. We describe how the resulting algorithm proved to perform well on biologically-relevant networks.

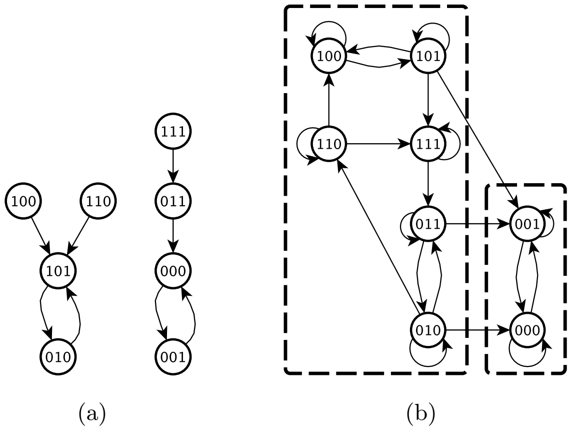


Fig. 1. Example of synchronous (a) and asynchronous (b) dynamics phase spaces of the same Boolean Network with $q = 2$, $|V| = 3$ and $T(\cdot)$ specified as follows: $T_1(s_1, s_2, s_3) = \neg s_3 \wedge (s_1 \vee s_2)$; $T_2(s_1, s_2, s_3) = s_1 \wedge s_3$; $T_3(s_1, s_2, s_3) = \neg s_3 \vee (s_1 \wedge s_2)$.

References

[DT11] E. Dubrova and M. Teslenko. A SAT-based algorithm for finding attractors in synchronous boolean networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(5):1393–1399, Sept 2011.

[Tar72] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1(2):146–160, 1972.

[VCL12] Alan Veliz-Cuba and Reinhard Laubenbacher. On the computation of fixed points in Boolean networks. *J. Appl. Math. Comput.*, 39(1-2):145–153, 2012.

Identification of “Die Hard” Nodes in Complex Networks

Angela Lombardi[1], Sabina Tangaro[2], Roberto Bellotti[2,3], Angelo Cardellicchio[1], and Cataldo Guaragnella[1]

[1] Dipartimento di Ingegneria Elettrica e dell’Informazione, Politecnico di Bari, Bari, Italy,
angela.lombardi@poliba.it,
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy
[3] Dipartimento Interateneo di Fisica “M. Merlin”, Universitá degli studi di Bari “A. Moro”, Bari, Italy.

Abstract A novel metric is proposed to quantify the importance of nodes in weighted networks in relation to its resilience. The resilience index takes into account the complete connectivity patterns of each node with all the other nodes in the network and is not correlated with other centrality metrics in heterogeneous weight distributions.

Keywords: Complex Networks, Resilience, Percolation, Centrality

1 Introduction

Complex networks have become a powerful tool for analyzing interactions in a great variety of contexts[2]. By using the complex framework, a system can be modeled in terms of nodes connected by binary or weighted edges whose magnitude quantify respectively the presence or the strength of the links between them. Weighted networks are particularly interesting for assessing topological properties in systems whereby is critical the importance of connections, e.g., social networks, ecological and biological systems, well-known for their hierarchical organization. As an example, there is clear evidence that many real-world networks exhibit a power law degree distribution which confers properties of structural robustness amongst attacks or failures[1].

Such properties are strictly related to the concept of percolation, i.e. the existence of a critical probability at which a single connected giant component exists and below which the network is composed of isolated clusters[2]. Different approaches have been proposed to assess the degree of the fragmentation of a network when a finite number of links are removed. Typically, metrics that quantify the importance or centrality of nodes are evaluated as edges are gradually removed from the network [3,4]. Several studies have tested the vulnerability of some synthetic networks for both binary and weighted cases, finding some network topologies more prone to attacks than others.

However, most of the real networks are characterized by a great heterogeneity of topological properties that are only partially taken into account in synthetic simulations. For instance, in weighted networks even weak connections can be statistically significant for a particular structural topology [5].

In this work, we present a new resilience index able to capture the node centrality for each percolation level. This metric quantifies the importance of the node in relation to its survival rate for progressive removal of links in the network. We simulate weighted undirected networks with scale-free structural topology and two levels of correlations between link weights and topology in order to reflect the heterogeneity of node behavior in real-world networks. We also test other graph metrics known in the literature to show their capability to detect the most important nodes in the two situations.

2 Methods

2.1 Synthetic networks

We used the B-A algorithm [2] to generate synthetic scale-free networks. Accordingly, the resulting networks have a power law degree distribution $P(k) \sim k^{\gamma}$ with $\gamma = 2$. The mechanisms of growth and preferential attachment are the main features of the scale-free networks: new nodes tend to connect to the more connected nodes that become hubs of the network. In order to represent different kind of weight-topology correlations, we assigned the weight of the link $w_{i,j}$ between nodes i and j as:

1. $w_{i,j} = k_i k_j$. In this way, the weight distribution follow a power law, too.
2. $w_{i,j} \sim \mathcal{U}(0, 1)$, to remove the correlation between weights and topology structure.

Figure 1 shows the same topology of a scale-free network composed by $N = 100$ nodes with the aforementioned weight distributions. As it can be seen, the most significant link in the power law case is established between the the two hubs of the network.

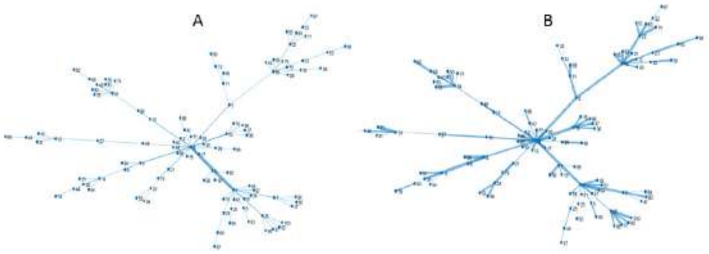


Figure 1. A scale-free network topology composed by $N = 100$ nodes with A) power law weight distribution; B) uniform weight distribution. Line width of the links are proportional to their weight.

2.2 Centrality metrics

In order to assess the importance of a node, we used:

- degree, i.e., the number of links connected to the node;
- strength, i.e., the sum of weights of links connected to the node;
- betweenness centrality, i.e., the fraction of all shortest paths in the network that contain a given node;
- eigenvector centrality, i.e., a self-referential measure of centrality: nodes have high eigenvector centrality if they are linked to other nodes that have high eigenvector centrality.

2.3 Resilience Index

Given the adjacency matrix W of the network in which the entry (i, j) indicates the weight of the connection $w_{i,j}$ between the nodes i and j, we divided the range of the weights into L levels and incrementally percolated M by removing all the links whose weight is below the threshold value at each level of percolation. We defined a tensor T in which the entry (i, j, l) is the weight of the link $w_{i,j}$ at the l^{th} level of percolation. In order to evaluate a resilience index for each node of the networks, we computed:

1. the connectivity pattern of the node i for the l^{th} level of percolation as:
$$P_{i,l} = T_{i,j,l} \qquad j = 1, \cdots, N. \tag{1}$$
2. the cosine similarity between connectivity patterns for each node and each level of percolation;
3. the cosine similarity returns a percolation curve for each node, which is equal to 1 for all the percolation levels in which at least a link of the node survives and assumes 0 value when the node becomes completely isolated from the rest of the network. So we defined the area under the percolation curve as the resilience index of the node.

3 Results and Discussion

The centrality metrics were evaluated for each node of the two networks and are shown in Figure 2. Obviously, the degree function is the same for both networks because they have the same structural topology. Strength and eigenvector values are emphasized for both hub nodes of the network (1 and 4 nodes) and for few other nodes with degree greater than that of the “leaf” nodes, i.e., those nodes with only one link. The values of betweenness are slight different only for the two hubs, so this metric does not take almost into account the distribution of the weights. On the other hand, some nodes with few connections exhibit high values of the resilience index for the random weight distribution, while its trend is closely correlated with that of the eigenvector centrality for the weight power law. The resilience index considers complete connectivity patterns of each node

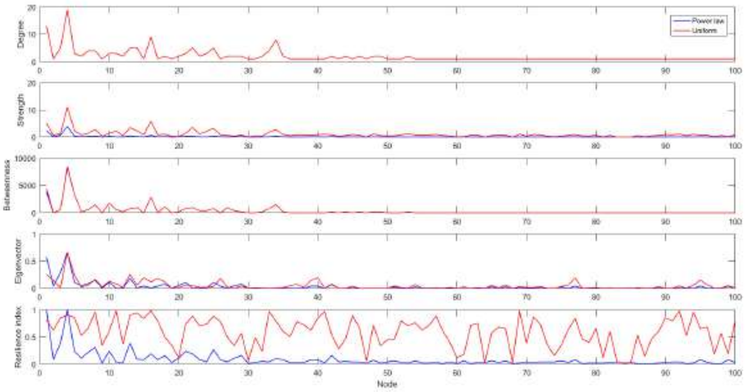


Figure 2. Values of the centrality metrics (degree, strength, betweenness, eigenvector and resilience index) for each of the 100 nodes for the two networks reported in Figure 1.

with the rest of the network at varying degrees of percolation. For this reason, even leaf nodes strongly connected to a particularly resilient node, can also be resilient nodes. The results of the correlation analysis between the proposed metric and the other centrality measures are listed in Table 1.

Table 1. Correlation between the resilience index and the other centrality metrics ($p < 0.0001$).

Network	Degree	Strength	Betweenness	Eigenvector
Power law	0.88	0.93	0.85	0.98
Uniform	0.32	0.50	0.29	0.38

References

1. Barabási, B. A. L., Bonabeau, E.: Scale-free. Scientific American, 288(5), 50–59 (2003).
2. Albert,R., Barabási, B. A. L.: Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47 (2002).
3. Holme, P., et al.: Attack vulnerability of complex networks. Physical review E 65(5), 056109 (2002).
4. Mishkovski, I., Biey, M., Kocarev, L.: Vulnerability of complex networks. Communications in Nonlinear Science and Numerical Simulation 16(1), 341–349 (2011).
5. Granovetter, M. S.: The strength of weak ties. American journal of sociology 78(6), 1360–1380 (1973).

Stochastic numerical modeling of selected oscillatory phenomena

R. D'Ambrosio[1], M. Moccaldi[1], B. Paternoster[1], and F. Rossi[2]

[1] Department of Mathematics, University of Salerno, Italy
[2] Department of Chemistry and Biology, University of Salerno, Italy

Keywords: BZ reaction, stochastic differential equations, exponential fitting

We present a stochastic numerical model for what can be considered, at the present state-of-the-art, a prototype chemical oscillator, i.e. the Belousov-Zhabotinsky reaction. We mainly focus our attention on properly modifying an usual model based on systems of deterministic differential equations, commonly known as Oregonator. We aim to provide a mathematical description of the Belousov-Zhabotinsky reaction by means of systems of stochastic differential equations of Ito type, together with the corresponding numerical discretization. Together with standard numerical schemes that approximate the whole stochastic system, such as the Euler-Maruyama method and its improvements, we also aim to separately treat the deterministic term and the stochastic one, by coupling the so-called exponential fitting technique [2] and Montecarlo simulations. This coupling looks particularly suitable to provide numerical approximations of oscillatory problems, since classical methods could require a very small stepsize to accurately reproduce oscillatory behaviours. We rather propose a method that is constructed in order to be exact on functions other than polynomials for the approximation of the deterministic part. The coefficients of the resulting adapted method are no longer constant as in the classical case, but rely on a parameter linked to oscillatory character, whose value is clearly unknown. The proposed estimation strategy is based on exploiting the informations belonging to known time series of experimental data which are available in the literature [1, 3]. Numerical experiments will be provided to show the effectiveness of the presented approach.

References

1. D'Ambrosio, R., Moccaldi, M., Paternoster, B., Rossi, F.: On the employ of time series in the numerical treatment of differential equations modelling oscillatory phenomena. In: Advances in Artificial Life, Evolutionary Computation, and Systems Chemistry - 11th Workshop, WIVACE 2016, Fisciano, Italy, ed. by F. Rossi, S. Piotto, S. Concilio, Comm. Comput. Inf. Sci., Springer (2017).
2. Ixaru, L.Gr., Vanden Berghe, G.: Exponential Fitting. Kluwer. Boston-Dordrecht-London (2004)
3. Rossi, F., Budroni, M. A., Marchettini, N., Cutietta L., Rustici, M. and Turco Liveri, M. L.: Chaotic dynamics in an unstirred ferroin catalyzed Belousov-Zhabotinsky reaction. Chem. Phys. Lett. 480, 322–326 (2009)

An integrated metabolism, growth and cell cycle model quantitatively describing budding yeast growth

Pasquale Palumbo[1-2] and Marco Vanoni[1-3] and Federico Papa[1-2] and Stefano Busti[1-2] and Meike Wortel[4-5] and Bas Teusink[4] and Lilia Alberghina[1-2]

[1] SYSBIO Centre for Systems Biology, Italy
[2] IASI-CNR Via dei Taurini 19, Rome, Italy
[3] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, Milan, Italy
[4] Systems Bioinformatics, VU University, Amsterdam, De Boelelaan 1087, 1081 HV, The Netherlands
[5] Centre for Ecological and Evolutionary Synthesis (CEES), the Department of Biosciences, University of Oslo, Blindernveien 31, 0371 Oslo, Norway
lilia.alberghina@unimib.it; marco.vanoni@unimib.it

Abstract. Computational models are expected to increase the understanding of how complex biological functions arise from the interactions of large numbers of gene products and biologically active low molecular weight molecules. Recent studies underline the need to develop quantitative models of the whole cell in order to tackle this challenge and to accelerate biological discoveries.

In this work we describe *iMeGroCy*, an *i*ntegrated model of the three major functions of a yeast cell: *Me*tabolism, *Gro*wth and *Cy*cle. The MeGro (Metabolism and Growth) and GroCy (Growth and Cycle) modules are linked together in a unified, low granularity model where MeGro acts as a *parameter generator* for GroCy. The model can be used as a scaffold for molecularly detailed models of yeast functions

Keywords: Computational Models, Systems Biology, Whole cell models.

1 Introduction

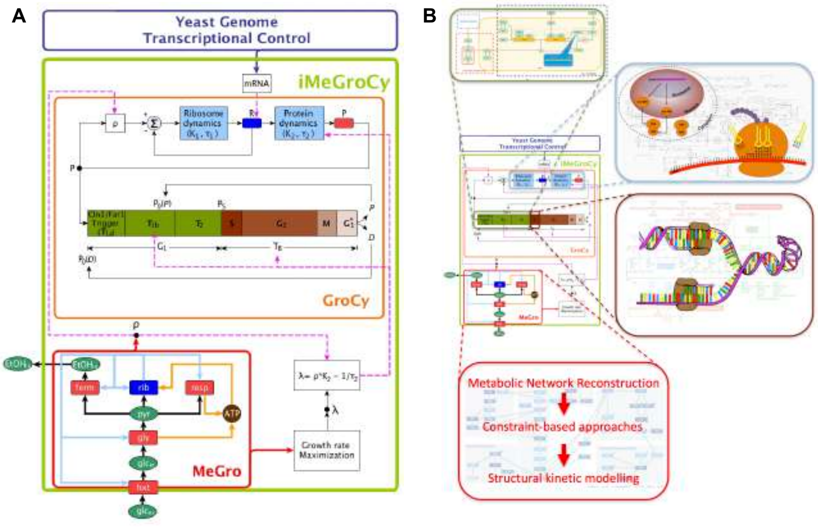
Saccharomyces cerevisiae is a major eukaryotic model organism in both fundamental and applied research. Computational approaches are required to analyze, structure and integrate the ever increasing data sets available for yeast. Ultimately, a dynamic, comprehensive computational model of *S. cerevisiae* should be the ambition: it would, in part, allow further improvement of industrial bioprocesses by extending the understanding presently possible by genome-scale metabolic model [1]. It also would allow translation of the methodologies to human cells, as it previously happened for genome sequencing, functional analysis and interactomics, just to name a few fields in which yeast research has recently led the way [2].

The design rules followed in the construction of the pioneering *Mycoplasma* whole cell model [3] were to divide the functionality of the cell into modules, each modeled for short enough periods of time to assume module independence. To translate this approach to a much more complex eukaryote such as yeast, however, is not straightforward and may produce models that are difficult to manage and from which is difficult to extract knowledge. In contrast to *Mycoplasma*, yeast has a compartmentalized cellular organization, a ten-fold larger genome [4] and sophisticated nutritionally modulated sensing and differentiation pathways [5] and an asymmetric cell division that results in population heterogeneity in terms of size, age and cellular content of individual cells. Several significant challenges facing the successful building of models of cells more complex than *Mycoplasma* have been described [6].

To deal with yeast complexity we developed a different approach. In the following, we present an integrated coarse-grained model of the basic functions of a yeast cell (metabolism, growth and cycle), investigate how they respond to availability of a major yeast nutrient, glucose, and discuss how the model can be used as a scaffold for molecularly detailed models of yeast functions.

2 Integrated Metabolism, Growth and Cycle model (*iMeGroCy*)

The *i*ntegrated *Me*tabolism, *Gro*wth and *Cy*cle (iMeGroCy) model is modular and hierarchical. The growth activity combined to the other two main cellular activities of metabolism (MeGro) and cycle (GroCy) define the modular building blocks constituting the coarse-grained backbone of iMeGroCy (Fig. 1A).



A
Yeast Genome
Transcriptional Control
mRNA
iMeGroCy
ρ
Σ
Ribosome dynamics (K_1, τ_1)
R
Protein dynamics (K_2, τ_2)
P
P
$P_s(P)$
P_c
Cln3/Far1 Trigger (T_1)
T_{1b}
T_2
S
G_2
M
G_1^*
P
D
G_1
T_B
$P_s(D)$
GroCy
ρ
EtOH
$EtOH_i$
ferm
rib
resp
pyr
ATP
gly
glc_i
MeGro
tsxt
glc_o
$\lambda = \rho^*K_2 - 1/\tau_2$
λ
Growth rate Maximization
B
Yeast Genome Transcriptional Control
iMeGroCy
GroCy
MeGro
Metabolic Network Reconstruction
Constraint-based approaches
Structural kinetic modelling

Fig. 1. A: iMeGroCy model (green block) composed by two main sub-blocks: the MeGro (red block) and the GroCy (orange block) interconnected (by dashed magenta arrows). **B:** finer modules of the G1/S transition, of RNA and ribosomes synthesis, of DNA synthesis and of the metabolism sub-module as blow-ups of the iMeGroCy.

2.1 The Metabolism and Growth model (MeGro)

The Metabolism and Growth Model (MeGro) connects growth and metabolism in *S. cerevisiae*. It is a coarse-grain representation of a yeast cell that maximizes its specific growth rate by allocating total protein synthesis capacity to different protein pools. MeGro - derived from the generic "self-replicator" model proposed in [7] for unicellular microorganisms - is conceived to highlight the common patterns connecting growth rate-dependent regulation of cell size, ribosomal content and metabolic efficiency in a cell. All the metabolic reaction rates, the kinetic parameters and the stoichiometry of the flux balance constraints in MeGro are suitably tuned for *S. cerevisiae* and only the relevant classes of enzymes and metabolites are considered.

MeGro accounts for five classes of proteins and five kinds of metabolites. The proteins with enzymatic activity (square blocks in the MeGro scheme of Fig. 1A) are (i) the hexose transporters, '*hxt*', (ii) the glycolytic enzymes, '*gly*', (iii) the ribosomes, '*rib*', (iv) the respiration and (v) fermentation pathways enzymes, '*resp*' and '*ferm*' respectively. Three kinds of metabolites are involved in metabolic conversions (green ovals in the MeGro scheme of Fig. 1A): the (a) extracellular and (b) intracellular glucose, '*glc,ex*', and '*glc,in*' respectively, and (c) pyruvate, '*pyr*'; other two kinds of metabolites are involved in energy production/consumption: (d) *ATP* and (e) *ADP*. In the following we indicate with c_x, $x \in \{Prot, Met\}$, $Prot = \{hxt, gly, rib, resp, ferm\}$, Met = $\{glc,ex, glc,in, pyr, ATP, ADP\}$, the protein/metabolite concentrations, (mM), and with v_x, $x \in Prot$, the metabolite fluxes, (mM/h) catalyzed by a specific protein x.

MeGro captures resource allocation strategies: too much investment into ribosomes would result in too little metabolic protein and hence too little substrate for the ribosomes. Proper tuning of the protein production rates will result in an optimal rate of the production of all cell components. Protein turnover is ignored in this study. All enzymatic activities are catalysed with Michaelis-Menten enzyme kinetics.

Protein synthesis is modeled as a proper fraction α_x of the ribosomal flux v_{rib}, thus, in exponential growth conditions we have the following steady-state constraints:

$$\lambda c_x - \alpha_x v_{rib} = 0, \qquad x \in Prot \tag{1}$$

where λ (h^{-1}) is the specific growth rate and $\sum_{x \in Prot} \alpha_x = 1$, with $\alpha_x \geq 0$.

The dynamics of $c_{glc,in}$, c_{pyr}, c_{ATP} are determined by the combination of the fluxes of synthesis and degradation:

$$dc_{glc,in}/dt = v_{hxt} - v_{gly}, \tag{2}$$

$$dc_{pyr}/dt = 2v_{gly} - v_{ferm} - v_{resp} - 600\,v_{rib}, \tag{3}$$

$$dc_{ATP}/dt = 2v_{gly} + 10\,v_{resp} - 2000\,v_{rib}, \tag{4}$$

providing steady-state constraints by imposing the derivatives equal to zero.

The total amount of ADP + ATP is constant, according to the following relationship

$$c_{ATP} + \mathrm{c}_{ADP} = 1. \tag{5}$$

All protein and metabolite concentrations c_x are such that $c_x \geq 0$. The fluxes of metabolites through the reactions are modeled using the Michaelis-Menten formalism:

$$v_{hxt} = \frac{k_{cat,hxt}\, c_{glc,ex}\, c_{hxt}}{(c_{glc,ex} + k_{m,hxt})\,(1 + c_{glc,in}/k_{i,glui})}\,, \tag{6}$$

$$v_{gly} = \frac{k_{cat,gly}\, c_{ADP}\, c_{glc,in}\, c_{gly}}{(c_{ADP}\, c_{glc,in} + k_{m,gly}\, c_{ADP} + k_{m,ADPgly}\, c_{glc,in} + k_{gly} k_{m,ADPgly})\,(1 + c_{pyr}/k_{i,pyr})}, \tag{7}$$

$$v_{rib} = \frac{k_{cat,rib}\, c_{ATP}\, c_{pyr}\, c_{rib}}{c_{ATP}\, c_{pyr} + k_{m,rib}\, c_{ATP} + k_{m,ATPrib}\, c_{pyr} + k_{m,rib}\, k_{m,ATPrib}}\,, \tag{8}$$

$$v_{resp} = \frac{k_{cat,resp}\, c_{ADP}\, c_{pyr}\, c_{resp}}{c_{ADP}\, c_{pyr} + k_{m,resp}\, c_{ADP} + k_{m,ADPresp}\, c_{pyr} + k_{m,resp}\, k_{m,ADPresp}}\,, \tag{9}$$

$$v_{ferm} = \frac{k_{cat,ferm}\, c_{pyr}\, c_{ferm}}{c_{pyr} + k_{m,ferm}}\,. \tag{10}$$

(1) to (10) define the set of algebraic-differential equations of MeGro. The exponential growth rate λ is maximized as a function of the external glucose concentration $c_{glc,ex}$ (model input), with the fractions α_x as optimization variables, and subject to exponential growth constraints (1), flux balance constraints (derived from steady-state eqs.(2-4)), feasible constraints (5) and Michaelis-Menten flux equations (6-10).

The optimal set of λ and α_{hxt}, α_{gly}, α_{rib}, α_{resp}, α_{ferm} together with the proteins/metabolites concentrations and the protein fluxes provide a first level of MeGro outputs. A second level of cellular outcomes are computed by properly exploiting concentrations and fluxes. These are (i) the fermentative ratio F,

$$F = v_{ferm} \,/\, (v_{ferm} + v_{resp}), \tag{11}$$

(ii) the ribosome-over-protein ratio ρ,

$$\rho = c_{rib} \,/\, (600\, (c_{hxt} + c_{gly} + c_{rib} + c_{resp} + c_{ferm})), \tag{12}$$

with proteins expressed in terms of number of polymerized amino acids, which explains the division by 600, the average number of polymerized amino acids per protein [8, 9] and (iii) the yield of ethanol $Y_{EtOH/glc}$,

$$Y_{EtOH/glc} = v_{ferm} \,/\, v_{hxt}. \tag{13}$$

MeGro can treat the fermentative ratio F as an input rather than an output, thus allowing the modeler to compute the optimal growth rate (as well as all the other model outputs) according to different values of F. Indeed, by properly exploiting the flux balance constraints (derived from steady state eqs.(2-4)) and the fermentative ratio definition (11), we can write v_{ferm} and v_{hxt} in terms of F and of the ribosomal flux v_{rib}:

$$v_{ferm} = (1.4{\cdot}10^3\, F) \,/\, (1 + 10(1 - F))\; v_{rib}, \tag{14}$$

$$v_{hxt} = 10^3{\cdot}(1 + 3(1 - F)) \,/\, (1 + 10(1 - F))\; v_{rib}, \tag{15}$$

so that, according to the ethanol yield definition (13) the fermentative ratio F is provided as a function of a given ethanol yield:

$$F = 20\ Y_{EtOH/glc} / (7 + 15Y_{EtOH/glc}). \tag{16}$$

This last equation will be exploited to feed MeGro with the fermentative ratio associated to experimental yield, Fig. 2A.

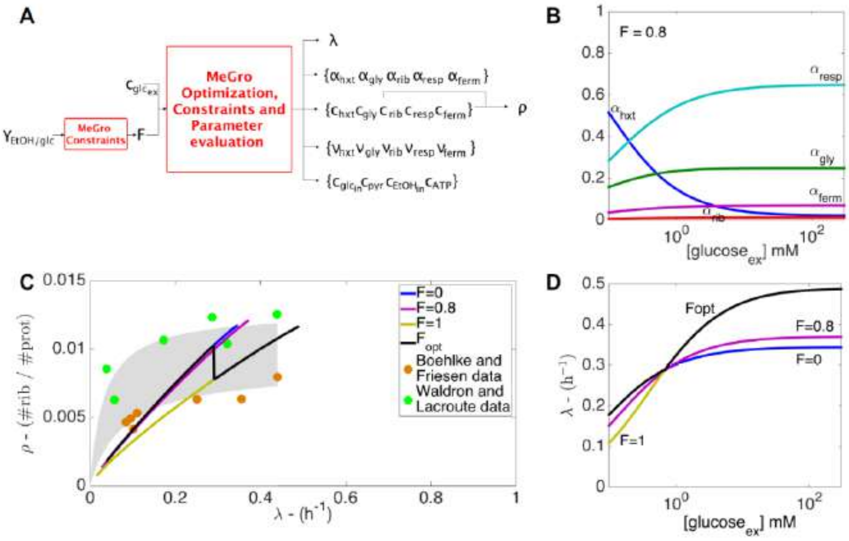


Fig. 2. A: MeGro outcomes, when both the external glucose concentration and the fermentative ratio (i.e., the yield of ethanol) are exploited as model inputs. **B:** optimal fractions of ribosomal activity (α_j) engaged in the synthesis of the corresponding protein modules as functions of the external glucose concentration. **C:** MeGro optimal ribosome-over-protein ratio ρ as a function of the MeGro optimal growth rate λ, with fixed fermentative ratio F (colored curves, F ranging in [0,1]) and not fixed F (bold black line). MeGro simulations are compared to the experimental data redrawn from [10, 11]: in grey we highlight the region between Michaelis-Menten experimental data best fitting. **D:** MeGro optimal growth rate λ as a function of the external glucose concentration $c_{glc,ex}$, with fixed fermentative ratio F (colored curves, F ranging in [0,1]) and not fixed F (bold black line). If we leave F as an optimization variable, the model predicts that the cell behavior is fully respiratory for values of external glucose smaller than a critical value and then switches to purely fermentative for values of external glucose greater than the threshold. MeGro parameters have been fixed as follows for any glucose input: $k_{cat,hxt}$ = 37492h^{-1}, $k_{cat,gly}$ = 4166h^{-1}, $k_{cat,rib}$ = 670h^{-1}, $k_{cat,resp}$ = 99h^{-1}, $k_{cat,ferm}$ = 6427h^{-1}, $k_{m,hxt}$ = 20mM, $k_{m,gly}$ = 0.2mM, $k_{m,rib}$ = $k_{i,pyr}$ = $k_{i,glui}$ = 1mM, $k_{m,resp}$ = $k_{m,ADPgly}$ = $k_{m,ADPresp}$ = $k_{m,ATPrib}$ = 0.5mM, $k_{m,ferm}$ = 5mM.

Fig. 2B reports the steady-state protein fluxes for the different protein pools as a function of glucose concentration for a fermentative ratio F = 0.8. λ increases as a function of external glucose concentration following a saturation kinetics, whose parameters depend on the fermentative ratio F (Fig. 2D). Fig 2C shows the behavior of the ribosome-over-protein ratio ρ as a function of the external glucose concentration, at different fixed values of the fermentative ratio F. Model predictions (solid lines) are compared with two sets of experimental data (green and orange circles) from different yeast strains [10, 11], showing overall agreement between model predictions and experimental data. MeGro parameters can be found in the caption of Fig. 2. Most parameters are chosen by following the criteria developed in [12], with minor modifications, mostly related to the use of different units.

2.2 The Growth and Cycle model (GroCy)

In yeast, the critical cell size required to enter S phase (P_S) is modulated by nutrient availability [13]. It remains small and nearly constant when glucose is utilized by respiration. In contrast, P_S and hence average protein content increases as cells shift their metabolism towards fermentation [14]. Cells forced to ferment under slow-growing conditions show the same increase [15].

GroCy is composed of three modules: (1) a dynamical cell growth model in which a set of ordinary differential equations describes dynamics of synthesis and degradation of ribosomes and proteins; (2) a molecular triggering mechanism that links cell growth and cell cycle. It exploits a set of ordinary differential equations which detail the dynamics of the growth-controlled activator Cdk1Cln3 and of its cognate inhibitor Far1; (3) a cell cycle module, that consists of three consecutive timers (T_{1b}, T_2 and T_B) that describe the cycle progression after the triggering mechanism activates the first timer T_{1b}. The period that leads from the birth of the cell up to the time instant when the molecular machinery triggers the first timer is denoted by T_{1a}.

The growth module. The growth module deals with the ribosome content *R*, expressed as number of ribosomes per cell (rib), and the protein content *P*, expressed as number of polymerized amino acids per cell (aa), and is taken from [16] (where the reader can find the equations and the details which are below briefly recalled). Both ribosome and protein dynamics are described by the balance between production and degradation rates.

ρ, the target ribosome/protein ratio for each steady-state growth condition is an output from MeGro that can be directly fed into GroCy, providing the link between the two models. According to the model, when the ratio R/P is greater than ρ, then there is no ribosome production; otherwise, the ribosome production rate is proportional to the (positive) difference $\rho P - R$. Denoting with K_2, τ_2, the average translational efficiency and the protein dynamics time constant, respectively, it can be shown that, provided the exponential growth condition is satisfied, ρK_2 - $1/\tau_2 > 0$, the ratio R/P asymptotically converges to the value of parameter ρ. The exponential growth condition ensures that both ribosomes and proteins grow according to the same exponential law, with an exponential growth rate λ (min^{-1}) given by: $\lambda = \rho K_2$ - $1/\tau_2$.

λ is not hard-wired in the model, but rather it is linked to the macromolecular composition and biosynthetic activity of the cells, a connection that is made possible by the appropriate choice of the measurement units for ribosome and protein content, synthesis and degradation.

The molecular triggering mechanism. In budding yeast the entrance into S phase and budding starts when cells reach a critical cell size, thus connecting growth and cycle [17, 18]. Cln3 is an activator of S phase entrance, whose amount is proportional to the overall protein content, eq. (21), therefore linking the growth and cycle modules. Despite some discordant results discussed in [19], we take that Cln3 accumulation is constant during G1. Cln3 production takes place in the cytoplasm. Cytoplasmic Cln3 is defined straightforwardly by eq. (21). Nuclear volume is a constant fraction of total cell volume throughout the cycle [20].

The cyclin-dependent kinase inhibitor Far1 is involved in the cell size control mechanism in cycling cells by inhibiting Cln3 in early G1 [21, 22]. After mitosis, newly synthesized Far1 is endowed to each nucleus [23]. Ensuing Cln3 nuclear transport and accumulation allows overcoming of Far1 inhibition, which is made irreversible by Far1 degradation primed by the rising Cln3 activity [24, 25].

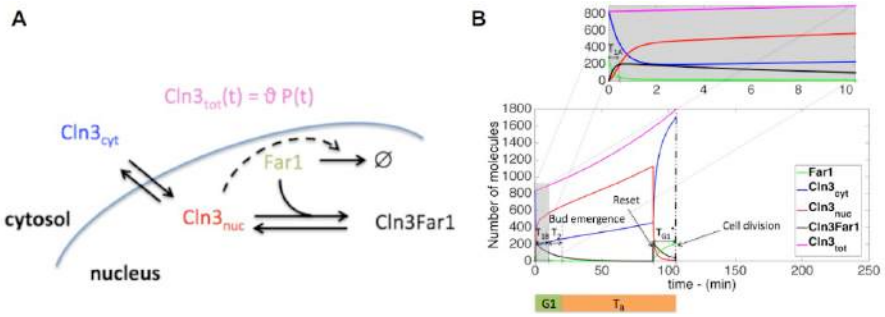


Fig. 3. A: cartoon of the molecular players involved in the molecular triggering mechanism of GroCy. **B:** time evolution of the players involved in the molecular triggering module of GroCy, for fast growth conditions (2% glucose). GroCy parameters are: ρ = 2.0e-5rib/aa, Far1(0) = 240, Cln3nuc(0) = Cln3Far1(0) = 0, K_1 = 1min^{-1}, τ_1 = 4000min, K_2 = 380aa/rib/min, $K_2^{\,1}$ = 342aa/rib/min, $K_2^{\,2}$ = 178aa/rib/min, $K_2^{\,3}$ = 69aa/rib/min, $K_2^{\,4}$ = 51aa/rib/min, $K_2^{\,5}$ = 42aa/rib/min, $K_2^{\,s}$ = 35aa/rib/min, $s > 5$, τ_2 = 3000min, $\tau_2^{\,s}$ = 1500min, s = 1, 2, …, h = 0.07, H = 7.1e23aa/L, k_{off} = 25min^{-1}, k_{on} = 1.5e-15(molec/L)$^{-1}$/min, nF = 10, η* = 1min^{-1}, θ = 3.0e-8molec/aa, k_{cn} = 1.5min^{-1}, k_{nc} = 0.6min^{-1}, $T_{1b,min}$ = 1min, W_0 = 1503min, W_1 = 62.1min, T_2 = 10min, T_B = 85min.

Here we use a simplified version of the equations used in [26] to model the molecular interplay between Cln3 and Far1, which we call the molecular triggering mechanism. Cdk1 – present in excess over its regulatory subunits – is implied, but not explicitly modeled (see the cartoon in Fig. 3A). Cln3 transport in the nucleus and Cln3/Far1 interaction follow mass action kinetics. Far1 degradation is governed by rate η (min^{-1}) modeled according to a Hill function (eq. (21)), that increases from 0 to a high level η^* as soon as the free nuclear Cln3 exceeds Cln3Far1. nF is the Hill coefficient, modeling the steepness of the Hill function. Eq. (19) accounts for reversible nucleo-cytoplasmic transport of Cln3 and interaction with Far1.

$$dCln3Far1/dt = (k_{on}/V_{nuc})\, Clin3_{nuc}\, Far1 - k_{off}\, Clin3Far1, \tag{17}$$

$$dFar1/dt = -(k_{on}/V_{nuc})\, Cln3_{nuc}\, Far1 + k_{off}\, Cln3Far1 - \eta(Cln3_{nuc}/\, Cln3Far1)\, Far1, \tag{18}$$

$$dCln3_{nuc}\,/dt = -(k_{on}/V_{nuc})\, Cln3_{nuc}\, Far1 + k_{off}\, Cln3Far1 + k_{cn}\, Cln3_{cyt} - k_{nc}\, Cln3_{nuc}, \tag{19}$$

$$Cln3_{cyt} = Cln3_{tot} - (Cln3_{nuc} + Cln3Far1), \qquad Cln3_{tot} = \theta\, P, \tag{20}$$

$$V_{nuc} = h V_{cell}, \qquad V_{cell} = P/H, \qquad \eta(x) = \eta^*\, x^{nF}/\,(1 + x^{nF}). \tag{21}$$

When 80% of the budded period has elapsed, a RESET function takes place and the G1* phase begins. RESET denotes the time instant when nuclear division (but not cell division) occurs: i.e. during the G1* phase each cell has two G1 nuclei and an undivided cytoplasm. At RESET the nuclear players show a discontinuity because they no more represent the whole (and unique) nuclear content and also because Far1 has been reset to a higher value. The RESET function includes instantaneous synthesis of Far1 and equal partition of Cln3Far1, Far1, Cln3$_{\mathrm{nuc}}$ in two nuclei whose volume is half of the original volume before RESET: $V_{nuc} = hV_{cell}/2$. Far1 degradation is inhibited (η = 0 in eq. (18)), and Cln3 diffusion from the cytoplasm into the nucleus is strongly reduced (k_{cn} in eq. (19) reduces of 5 orders of magnitudes during the G$_1^*$).

Fig. 3B shows the time course for the different molecular players throughout the whole cycle of an average size cell in fast nutrient conditions (2% glucose): when - very early after division - free nuclear Cln3 overcomes its inhibited form Cln3Far1 the first of the three consecutive Timers related to the cell cycle module is triggered. The time period spanning from the birth of the cell up to the aforementioned time instant is named T_{1a}. In extra-small newborn cells the initial amount of Cln3 is much lower, so that quite a long period may be required before free nuclear Cln3 exceeds Cln3Far1. The kinetic parameters of the molecular trigger do not vary in different nutrient environments, except for the total amount of Far1, known to diminish in poor media [21], and for the parameters H, θ assumed to decrease in case of poor growth.

The cell cycle module. In *S. cerevisiae* cell mass at division is unequally partitioned [18] between a larger, old parent cell (P) and a smaller, newly synthesized daughter cell (D). The degree of asymmetry of cell division in *S. cerevisiae* is modulated by nutrients: poor media – such as ethanol - yield a high level of asymmetry with large parent cells and very small daughter cells, whereas in rich media - such as glucose - parents and daughters at division are very close in size (reviewed in [14]). Since cells have to grow to a critical cell size before entering S phase and budding, small daughter cells have a longer cycle time than the corresponding parent cells, most notably in poor media. This difference in cycle time between daughter and parent cells is due to differences in the G1 phase, whilst the budded period T_B has essentially the same length in both parents and daughter cells [17]. Differences in growth rate have marginal effects on the length of T_B and dramatic effects on the length of G1 (reviewed in [14]).

As explained, T_{1a} is the period from the birth of a cell till the time instant when free nuclear Cln3 exceeds its inhibited form Cln3Far1; the rest of the cycle is modeled by the sequence of three consecutive timers T_{1b}, T_2, and T_B. The sum of the period T_{1a} + timer T_{1b} corresponds to timer T_1 in [17]. The G1 phase is given by $T_1 + T_2$. Timer T_B encompasses the budded phase.

The first timer T_{1b} starts when free Cln3 exceeds its inhibited form Cln3Far1. The length of T_{1b} is related to the size of the cell, so that larger cells have smaller T_{1b} periods, and vice versa. More in details, T_{1b} length is set according to the equation

$$T_{1b} = \max\{T_{1b,min}, W_0 - W_1 \ln(P_{T1a})\}, \qquad (22)$$

with P_{T1a} denoting the size of the cell at the end of T_{1a}. Notice that P_{T1a} plays an active role in the setting of T_{1b} only for cells small enough, i.e. only when:

$$W_0 - W_1 \ln(P_{T1a}) > T_{1b,min} \rightarrow P_{T1a} < \exp\{(W_0 - T_{1b,min}) / W_1\}. \qquad (23)$$

This happens, for instance, with most of daughter cells. In parent cells P_{T1a} is, usually, greater than the upper bound in inequality (23), so that their T_{1b} length is fixed to $T_{1b,min}$ and does not depend on the size.

The length of timer T_2 does not depend on protein content (no distinction for daughters and parents) [17]. At the end of timer T_2, the critical size expressed both as ribosome content and as protein content, R_s and P_s respectively, is estimated.

T_B, refers to the budded period, and encompasses the S, G2, M and G1* phases. This last phase has been modeled as the last 20% period of the whole T_B phase. The end of the timer results in cell division. Like timer T_2, timer T_B length does not depend on protein content, (no difference between daughters and parents). Part of the GroCy parameters are influenced by - and vary according to - the nutrient environment

2.3 Genealogical age heterogeneity

When a yeast cell buds, a chitin ring, called bud scar, builds up at the bud isthmus and remains on the parent cell after the bud has separated (reviewed in [14]). Since each new bud starts at a new site, it is possible to determine the number of bud scars ‘s’ present on the surface of a parent cell and consequently to establish the genealogical age ‘k’ of the parent cell. So, denoting by “P_k” a Parent cell of age ‘k’, a cell P_1 has one bud scar since it has completed a cycle, a cell P_2 has two bud scars since it has completed two cycles, and so on. On the other hand, a cell without bud scars (s = 0) is a Daughter cell and it has not yet completed a cycle. GroCy model, however, distinguishes the genealogical age of the daughter cells: it can be 1 if the daughter is born from another daughter, while it is k > 1 if the Daughter is born from a parent P_{k-1}. We denote by “D_k” a Daughter of genealogical age ‘k’.

Each parent increases in size before starting to bud [27], and at division it receives the mass it had at budding (the mass synthesized during the budding phase going to the newborn daughter), it follows that in parents, cell mass at budding increases with genealogical age. A reduction in cells size increase with genealogical age has been reported [14, 27] (see inset in Fig. 4B) and explained by mechanical stress of the cell wall, which increases with cell size [28]. Both K_2 and τ_2 in the growth module of GroCy (rate of protein synthesis and time constant of protein degradation respectively) decrease in value during the pre-budded period (G1 phase), according to the parent genealogical age, returning to their nominal values at the onset of the budding phase (end of Timer T_2), so that the parent cell P_k grows again with the steady-state exponential rate given by $\lambda = \rho K_2 - 1/\tau_2$. Daughter cells (of any genealogical age) are not affected by such a mechanical stress.

3 MeGro and GroCy integrated in iMeGroCy

Fig. 4A reports a functional scheme with the aim to provide a general procedure to set iMeGroCy parameters according to different glucose nutrient environment.

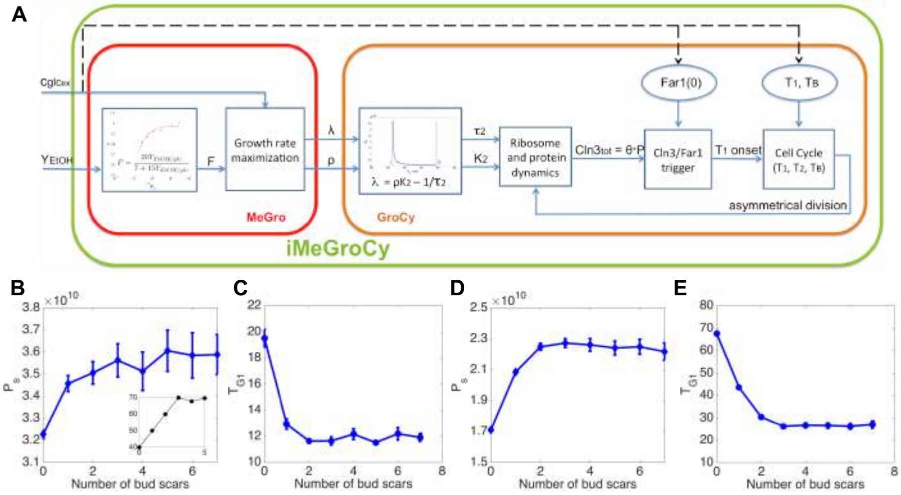


Fig. 4. **A:** Functional iMeGroCy block diagram**. B, D:** critical size Ps, given in number of polymerized amino acid (aa), as a function of the number of bud scars of the parent cells, for fast (panel B, 2% glucose) and slow (panel D, 0.05% glucose) growth conditions. **C, E:** G1 phase (min) as a function of the number of bud scars of the parent cells, for fast (panel C) and slow (panel E) growth conditions. The case of zero scars involves the class of daughter cells. Data in panels B-E report average +/- standard errors obtained from 20 chains of cells D_1, P_1, … , P_7, each seeded starting from 20 different initial cells; timers T_{1b}, T_2, T_B and the initial protein content of the cells have been allowed to vary (log-normal distribution) with a 5% CV over their average values. Panel B reports an inset with experimental data redrawn from [27].

Megro responds to the external glucose $c_{glc,ex}$ and $Y_{EtOH/glc}$ coming from experimental data in order to set the steady-state exponential growth rate λ and ribosome-over-protein ratio ρ as outputs of an optimization algorithm aiming at maximizing the growth rate. The MeGro outputs λ and ρ enter GroCy as inputs, allowing to set the ribosome and protein dynamics parameters, constituting the growth module. As far as the other growth parameters, the exponential growth relationship $\lambda = \rho K_2 - 1/\tau_2$ is used to constrain the GroCy parameters K_2 and τ_2 to the MeGro outputs (λ, ρ). K_1 and τ_1 have been fixed accordingly to [17]: differently from K_2 and ρ, parameters K_1, τ_1 and τ_2 do not change at different glucose concentrations.

iMeGroCy is conceived to replicate pedigree populations at different nutritional conditions. Fig. 4B-D provides a hint of its potentiality by showing protein content and G1 phase length correlated to the genealogical age of cells related by a parental relationship. Fast (glucose 2%) and slow (glucose 0.05%) growth conditions are considered. The inset in Fig. 4B compares simulated data to experimental ones.

4 Conclusions

In this work we presented a coarse grain model for yeast, dealing with metabolism, growth and cycle. We suggest that our low granularity model may act as a scaffold for the construction of a whole cell model for *S. cerevisiae*, Fig. 1B. Adding the modules incrementally, the ability of the model to fit real data could be monitored at any step. By way of example the metabolism module could be substituted by a genome-wide model, appropriately modified to include connections with cell growth and regulation by nutrients. The G_1 timers could be substituted by a recently described G_1/S module [19] or the budded phase by a wave of cyclins [29]. iMeGroCy could then be expanded to include models of signal transduction pathways. In short, top-down definition of the modules to be implemented would allow coherent expansion of the model and favor collaboration among the yeast community, since such an ambitious large-scale project will by necessity require a collaborative effort, that would characterize a new era in life science research [30].

References

1. Sánchez, B.J., Nielsen, J.: Genome scale models of yeast: towards standardized evaluation and consistent omic integration. Integr. Biol. Quant. Biosci. Nano Macro. 7, 846–858 (2015).
2. Botstein, D., Fink, G.R.: Yeast: an experimental organism for 21st Century biology. Genetics. 189, 695–704 (2011).
3. Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., Covert, M.W.: A whole-cell computational model predicts phenotype from genotype. Cell. 150, 389–401 (2012).
4. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G.: Life with 6000 genes. Science. 274, 546, 563–567 (1996).
5. Conrad, M., Schothorst, J., Kankipati, H.N., Van Zeebroeck, G., Rubio-Texeira, M., Thevelein, J.M.: Nutrient sensing and signaling in the yeast Saccharomyces cerevisiae. FEMS Microbiol. Rev. 38, 254–299 (2014).
6. Macklin, D.N., Ruggero, N.A., Covert, M.W.: The future of whole-cell modeling. Curr. Opin. Biotechnol. 28, 111–115 (2014).
7. Molenaar, D., van Berlo, R., de Ridder, D., Teusink, B.: Shifts in growth strategies reflect tradeoffs in cellular economics. Mol. Syst. Biol. 5, 323 (2009).
8. von der Haar, T.: A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst. Biol. 2, 87 (2008).
9. Waldron, C., Jund, R., Lacroute, F.: The elongation rate of proteins of different molecular weight classes in yeast. FEBS Lett. 46, 11–16 (1974).
10. Boehlke, K.W., Friesen, J.D.: Cellular content of ribonucleic acid and protein in Saccharomyces cerevisiae as a function of exponential growth rate: calculation of the apparent peptide chain elongation rate. J. Bacteriol. 121, 429–433 (1975).
11. Waldron, C., Lacroute, F.: Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. J. Bacteriol. 122, 855–865 (1975).

12. Wortel, M.T., Bosdriesz, E., Teusink, B., Bruggeman, F.J.: Evolutionary pressures on microbial metabolic strategies in the chemostat. Sci. Rep. 6, srep29503 (2016).
13. Alberghina, L., Mavelli, G., Drovandi, G., Palumbo, P., Pessina, S., Tripodi, F., Coccetti, P., Vanoni, M.: Cell growth and cell cycle in Saccharomyces cerevisiae: basic regulatory design and protein-protein interaction network. Biotechnol. Adv. 30, 52–72 (2012).
14. Porro, D., Vai, M., Vanoni, M., Alberghina, L., Hatzis, C.: Analysis and modeling of growing budding yeast populations at the single cell level. Cytom. Part J. Int. Soc. Anal. Cytol. 75, 114–120 (2009).
15. Porro, D., Brambilla, L., Alberghina, L.: Glucose metabolism and cell size in continuous cultures of Saccharomyces cerevisiae. FEMS Microbiol. Lett. 229, 165–171 (2003).
16. Alberghina, L., Mariani, L., Martegani, E.: Cell cycle modelling. Biosystems. 19, 23–44 (1986).
17. Di Talia, S., Skotheim, J.M., Bean, J.M., Siggia, E.D., Cross, F.R.: The effects of molecular noise and size control on variability in the budding yeast cell cycle. Nature. 448, 947–951 (2007).
18. Hartwell, L.H., Unger, M.W.: Unequal division in Saccharomyces cerevisiae and its implications for the control of cell division. J. Cell Biol. 75, 422–435 (1977).
19. Palumbo, P., Vanoni, M., Cusimano, V., Busti, S., Marano, F., Manes, C., Alberghina, L.: Whi5 phosphorylation embedded in the G_1/S network dynamically controls critical cell size and cell fate. Nat. Commun. 7, ncomms11372 (2016).
20. Jorgensen, P., Edgington, N.P., Schneider, B.L., Rupes, I., Tyers, M., Futcher, B.: The size of the nucleus increases as yeast cells grow. Mol. Biol. Cell. 18, 3523–3532 (2007).
21. Alberghina, L., Rossi, R.L., Querin, L., Wanke, V., Vanoni, M.: A cell sizer network involving Cln3 and Far1 controls entrance into S phase in the mitotic cycle of budding yeast. J. Cell Biol. 167, 433–443 (2004).
22. Fu, X., Ng, C., Feng, D., Liang, C.: Cdc48p is required for the cell cycle commitment point at Start via degradation of the G1-CDK inhibitor Far1p. J. Cell Biol. 163, 21 (2003).
23. McKinney, J.D., Chang, F., Heintz, N., Cross, F.R.: Negative regulation of FAR1 at the Start of the yeast cell cycle. Genes Dev. 7, 833–843 (1993).
24. Chang, F., Herskowitz, I.: Phosphorylation of FAR1 in response to alpha-factor: a possible requirement for cell-cycle arrest. Mol. Biol. Cell. 3, 445–450 (1992).
25. Peter, M., Gartner, A., Horecka, J., Ammerer, G., Herskowitz, I.: FAR1 links the signal transduction pathway to the cell cycle machinery in yeast. Cell. 73, 747–760 (1993).
26. Barberis, M., Klipp, E., Vanoni, M., Alberghina, L.: Cell Size at S Phase Initiation: An Emergent Property of the G1/S Network. PLOS Comput. Biol. 3, e64 (2007).
27. Johnston, G.C., Ehrhardt, C.W., Lorincz, A., Carter, B.L.: Regulation of cell size in the yeast Saccharomyces cerevisiae. J. Bacteriol. 137, 1–5 (1979).
28. Alberghina, L., Vai, M., Vanoni, M.: Probing Control Mechanisms of Cell Cycle and Ageing in Budding Yeast. Curr. Genomics. 5, 615–627 (2004).
29. Barberis, M., Linke, C., Adrover, M.À., González-Novo, A., Lehrach, H., Krobitsch, S., Posas, F., Klipp, E.: Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins. Biotechnol. Adv. 30, 108–130 (2012).
30. Swierstra, T., Vermeulen, N., Braeckman, J., van Driel, R.: Rethinking the life sciences. To better serve society, biomedical research has to regain its trust and get organized to tackle larger projects. EMBO Rep. 14, 310–314 (2013).

Estimating the multi-scale effects of extrinsic noise on genes and circuits activity from an empirically validated model of transcription kinetics

Samuel M.D. Oliveira[1,2], Mohamed N.M. Bahrudeen[1,2], Sofia Startceva[1,2], and Andre S. Ribeiro[1,2]

[1] Laboratory of Biosystem Dynamics, BioMediTech Institute, Tampere University of Technology, Finland. P.O Box 553, 33101 Tampere, Finland.
[2] Multi-scaled biodata analysis and modelling Research Community, Tampere University of Technology, 33101, Tampere, Finland.
andre.ribeiro@tut.fi

Abstract. Recent studies of *Escherichia coli* transcription dynamics using time-lapse confocal microscopy and *in vivo* single-RNA detection confirmed that transcription initiation has two main rate-limiting steps. Here, we argue that this allows selective 'tuning' of the effects of extrinsic noise on a multi-scale level that ranges from individual genes to large-scale gene networks. First, using empirically validated stochastic models of transcription and translation, we show that the effects of RNA polymerase numbers' cell-to-cell variability on the cell-to-cell diversity in RNA numbers decrease as the relative time-length of the open complex formation increases. Next, using a stochastic model of a 2-genes symmetric toggle switch, we show that the cell-to-cell diversity of the switching frequency due to cell-to-cell variability in RNA polymerase numbers also depends on the promoter kinetics. Finally, from the binarized protein numbers over time of 50-gene network models where genes interact by repression, we calculate the cell-to-cell variability of the mutual information and Lempel-Ziv complexity of the networks dynamics, and find that, while arising from the cell-to-cell variability in RNA polymerase numbers, these variability levels also depend on the promoter initiation kinetics. Given this, we hypothesize that *E. coli* may be capitalizing on the 2 rate-limiting steps' nature of transcription initiation to tune the effects of extrinsic noise at the single gene, motifs, and large gene regulatory network levels.

Keywords: Transcription Initiation, Extrinsic Noise, Genetic Circuits, Mutual Information, Lempel-Ziv Complexity.

1 Introduction

When facing changing conditions, *Escherichia coli* cells can perform behavioral changes, that can range from 'smooth' to 'sharp'. This degree of change depends on the changes (how many and by how much) in the regulatory molecules of the transcriptional and translational machineries, such as RNA polymerase (RNAP) core enzymes, promoter sequence, σ factors, transcription factors, and ribosomes [1,2].

Single-cell measurements have shown that, even in monoclonal bacterial populations, cells differ widely in component numbers [3,4]. Consequently, the behavioral changes in response to, e.g., an environmental change, vary widely between individual cells. Such variability in cellular components numbers that causes cell-to-cell variability in the dynamics of cellular processes is usually termed "extrinsic noise". Meanwhile, variability in the dynamics of a system that arises from the stochastic nature of its underlying processes (e.g. the stochastic nature of an event such as two molecules binding to one another) is usually termed 'intrinsic noise'.

Because of the influence of these noise sources in cellular processes, the response of genes, in activity level, to global changes in regulatory molecules numbers is also highly diverse both in time (in a single cell, due to intrinsic noise) and across a cell population (due to intrinsic and extrinsic noise sources) [5]. In the case of σ factors' direct positive regulation, this is believed to be due to a promoter-dependent selectivity for the σ factors [6], and/or the action of transcription factors [5]. Meanwhile, in the case of indirect negative regulation, it has recently been shown to be due to differences in the multi-step kinetics of transcription initiation of the promoters [7].

Following this finding, we recently have made use of stochastic modelling to explore the hypothesis that the dynamics of the rate-limiting steps in transcription initiation [8,9] may influence individual genes' degree of responsiveness to extrinsic noise [10,11].

Here we investigate this phenomenon further, on a wide multi-scale perspective, namely, from individual genes to large-scale networks involving tens of genes. In particular, we study, at each level of complexity, the response to changing extrinsic noise levels as a function of the transcription kinetics of the component genes.

For this, we implement stochastic models of individual genes, genetic toggle switches, and 50-gene networks accounting for cell-to-cell diversity in RNAP numbers. Parameter values used in the models are obtained from recent microscopy measurements of single-cell RNAP, RNA, and protein numbers. Stochastic simulations [12,13] of these models are performed to assess the extent to which the kinetics of initiation of the component promoters can be used to tune the level of the effects of the cell-to-cell variability in RNAP numbers on the dynamics of individual genes, genetic switches and 50-gene networks.

2 Methods

2.1 Models of Transcription, Translation, genes networks, and source of extrinsic noise.

Our stochastic models of gene expression and genetic circuitry are based on multiple genome-wide studies of cell-to-cell variability in RNA numbers [14,15], transcription dynamics of individual genes [9], translation kinetics at the single protein level [16-18], protein folding and activation kinetics [19], natural genetic switches [20,21], and topology of large-scale circuits [22]. Importantly, the value set for each parameter associated to the core process of gene expression was obtained from empirical data (Table 1).

Assuming a N-genes network, the multi-step transcription process of an active promoter *i*, P^i_{ON}, is modeled by reactions (1), with *i* = {1, ..., N} [23]. The closed complex (RP_c) is formed once an RNAP (R) binds to a free promoter [24]. Subsequent steps follow to form the open complex (RP^i_o) [23,24]. Finally, elongation starts [25], clearing the promoter. In the end, an RNA is produced and the RNAP is released. Elongation is not considered, due to its much shorter time-length when compared to initiation [9].

In (1), k_1 is the rate at which an RNAP (R) finds and binds to promoter P^i, k_{-1} is the rate of reversibility of the closed complex, k_2 is the rate of open complex formation, and k_3 is the rate of promoter escape (expected to be much higher than all other rates, and thus assumed to be 'infinite' [9]):

$$P^i_{\mathrm{ON}} + R \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} RP^i_c \xrightarrow{k_2} RP^i_o \xrightarrow{k_3} R + P^i_{\mathrm{ON}} + \mathrm{rna}^{\mathrm{i}} \quad (1)$$

Reactions (2) and (3) model translation of the RNA and subsequent protein folding (which includes activation, for simplicity), respectively:

$$rna^i + Rib \xrightarrow{k_{rbs}} rna^i + Rib + \mathrm{Prot}^i_{unfolded} \quad (2)$$

$$\mathrm{Prot}^i_{unfolded} \xrightarrow{k_{fold}} \mathrm{Prot}^i_{folded} \quad (3)$$

Reactions (4) and (5) model degradation and dilution due to cell division of RNA and proteins, respectively:

$$rna^i \xrightarrow{kd_{RNA}} \varnothing \quad (4)$$

$$\mathrm{Prot}^i_{folded} \xrightarrow{kd_P} \varnothing \quad (5)$$

For simplicity, we assume that genes interact solely via repression mechanisms that block initiation [31]. This suffices to model several known small gene network motifs, such as, e.g., genetic switches. The repression mechanism is modeled by reactions (6), which account for the transitioning of the promoter to active/inactive (P^i_{ON} / P^i_{Rep}) due to the unbinding/binding of an active repressor protein (Prot^j_{folded}), produced by gene *j*:

$$P^i_{\mathrm{ON}} + \mathrm{Prot}^j_{folded} \underset{k_{ON}}{\overset{k_{\mathrm{Rep}}}{\rightleftarrows}} P^i_{\mathrm{Rep}} \quad (6)$$

Finally, we assume a mean cell lifetime ('*div*') of 1 hour [9]. The dilution rate (Dil) of RNA and proteins due to cell division thus equals:

$$Dil = div^{-1} \times \log(2) \quad (7)$$

Taking into account the dilution due to cell division, along with the molecules' natural degradation rate (Deg), one has that the overall decay rate of, e.g., RNA molecules (k_d) along one line of a cell lineage will be:

$$k_d = Dil + Deg \tag{8}$$

The same formula is applied to proteins, using the appropriate rate constant (k_{dp}) for the degradation process. It is noted that, in agreement with [9], in the case of the model of single genes, we assume that the cell has a constant amount of active repressors, which, at points, force the promoter to go into the repressed state.

Given these models, assuming an 'active' promoter, we define τ_{prior} as the mean expected time for a successful closed complex formation, which depends on the speed and number of attempts to initiate an open complex formation (which, in turn, depends on the RNAP concentration). Meanwhile, the remaining time to produce an RNA, τ_{after}, includes the steps following commitment to open complex formation (e.g. isomerization [31]), and prior to transcription elongation. The mean time interval between consecutive RNA productions (Δt_{active}) of a fully active promoter is thus given by:

$$\Delta t_{active} = \tau_{prior} + \tau_{after} \tag{9}$$

Relevantly, in this model, τ_{after} does not depend on the RNAP intracellular concentrations. This is of significance in that, e.g., fluctuations in this concentration will only cause fluctuations in τ_{prior} and thus, will only cause '*partial'* fluctuations in Δt_{active}, whose intensity will depend on the ratio τ_{after} / Δt_{active}. Note also that we do not expect this formula to describe RNA production of genes in circuits, since the repression mechanism will cause significant changes to the RNA production kinetics prior to a successful closed complex formation.

The model above, as it is based on chemical reactions, and it is simulated in accordance with the stochastic simulation algorithm (SSA) [12], will result in systems whose dynamics are inherently stochastic due to two intrinsic noise sources, namely, the variability in the time moments that reactions occur and which reaction occurs next [12].

In addition to this, at the cell population level, the model possesses an extrinsic source of noise, which consists of a variability in the RNAP numbers of individual cells. This variability, as explained in the next section, is based on empirical data.

2.2 Cell-to-cell Variability in RNA Polymerase Numbers

RNAP numbers in individual model cells are set based on measurements of RNAP fluorescence intensity in individual *E. coli* RL1314 cells with fluorescently tagged β' subunits [9]. In particular, we set the mean RNAP fluorescence in individual cells arbitrarily to 1 and obtain the fraction of cells with a given relative fluorescence level. The 2.5% cells with lowest and highest fluorescence intensity were discarded as outliers.

To obtain empirical values, we measured the cell-to-cell variability in RNAP numbers by calculating the squared coefficient of variation (CV^2) of the RNAP fluorescence intensity levels in individual cells. Next, to obtain the CV^2 of RNAP relative levels in individual cells, we fitted a normal distribution to the data (MATLAB package Statistics and Machine Learning Toolbox™). The CV^2 of the fit equaled 0.03, in agreement with [9]. To validate the fitting, we performed a Kolmogorov-Smirnov (KS) test between the empirical and best fit distributions, which showed that they cannot be statistically distinguished (*p*-value of 0.69, which, by being much higher than 0.01, clearly indicates that the null hypothesis that the two sets of data are from the same distribution cannot be rejected). Finally, we used this best fit distribution to set random RNAP numbers in each model cell, unless stated otherwise.

2.3 Detecting Switches in the Dynamics of the Toggle Switch and Switching Frequency Quantification

To detect 'switches' in the two protein numbers over time in toggle switches (where a switch is a change in which protein is more abundant), at each moment of the simulation, we calculate the difference between the numbers of the two proteins, denoted as 'Prot1' and 'Prot2'.To not account short, transient switches, we use the following filter: if the absolute difference between Prot1 and Prot2 is smaller than 100 (~10% of the mean protein numbers of an active gene in our models), we set the difference to 0. The number of switches during the time series is the number of times the difference between the two protein numbers changes from positive to negative or vice-versa. Given that n is the number of switches and Δt is the observation time, the switching frequency (F) is quantified as:

$$F = \frac{n+1}{\Delta t} \tag{10}$$

2.4 Parameter Values and Simulations

Simulations are performed by SGNS [13], a simulator of chemical reaction systems based on the Delay Stochastic Simulation Algorithm [12, 32].

Each model cell 'contains' the systems of reactions (1)-(6). These reactions have the parameter values shown in Table 1 (unless stated otherwise). According to the model, e.g., increasing k_1 decreases the closed complex formation, while increasing k_2 shortens the open complex formation time. Here, we tune k_1 and k_2 so that the mean RNA production rate is kept constant, using the following formula [9]:

$$I(R) = \frac{\left(k_{\mathrm{ON}} + k_{\mathrm{rep}}\right)\left(k_{-1} + k_2\right)}{Rk_1k_2k_{\mathrm{ON}}} + \frac{1}{k_2} + \frac{1}{k_3} \tag{11}$$

where I(R) is the mean interval between consecutive RNA productions in individual cells, assuming infinite cell lifetime. In (12), we first set all parameter values to the values shown in Table 1, to obtain the value of I(R) in the control condition. Next, again using this formula, we alter k_I and k_2, so that I(R) is kept constant and equal to the I(R) of the control condition. This allows changing the ratio $\tau_{after}/\Delta t$, which is given by:

$$\frac{\tau_{after}}{\Delta t} = 1 - \frac{\left(k_{\mathrm{ON}} + k_{\mathrm{OFF}}\right)\left(k_{-1} + k_2\right)}{Rk_1k_2k_{\mathrm{ON}}} \times I(R)^{-1} \qquad (12)$$

Table 1. Parameter values of the models (control condition). k_I and k_{rbs} values are set assuming that the number of available RNAP and ribosomes equal 1 (and are never depleted).

Parameter	Value (s^{-1})	Reference
k_{ON}	0.01	[9]
k_{rep}	281	[9]
k_I	6469	[9]
k_{-1}	1	[9]
k_2	0.005	[9]
kd_{rna}	0.0033	[14], Eq. (7)
k_{rbs}	0.637	[16-18]
k_{fold}	0.0024	[19]
k_{dp}	0.0019	[19], Eq. (7)

The range of possible values of $\tau_{after}/\Delta t$ is set to be [0.05, 0.95], due to high diversity of empirical values of different promoters and promoters subject to different induction settings. These values, reported in [7], are here shown in Table 2:

Table 2. Empirical values of $\tau_{after}/\Delta t$ of various promoters under various induction levels.

Promoter and Induction	$\tau_{after}/\Delta t$	Reference
BAD (0.1% arabinose)	0.29	[7]
BAD (0.01% arabinose)	0.45	[7]
BAD (0.001% arabinose)	0.83	[7]
Lac-O_1O_3 (1 mM IPTG)	0.45	[7]
Lac-O_1O_3 (0.05 mM IPTG)	0.54	[7]
Lac- O_1O_3 (0.005 mM IPTG)	0.88	[7]
TetA (no inducers)	0.93	[7]
Lac-O_1 (1 mM IPTG)	0.95	[7]
Lac-ara1 (1 mM IPTG and 0.1% arabinose)	0.51	[7]

In the case individual genes and small motifs, we simulate models whose promoters differ in $\tau_{after}/\Delta t$ (by changing k_I and k_2 while keeping Δt constant) by 0.1, from 0.05 to 0.95 (i.e. 10 conditions). Meanwhile, given that the RNAP cell-to-cell variability (CV^2(RNAP)) is known to be equal to 0.03 in optimal growth conditions [9], and assuming that it is likely higher in sub-optimal conditions, we simulate models that differ in CV^2(RNAP) by 0.015, from 0 to 0.09 (i.e. 7 conditions). As such 70 different models are simulated. Specifications of the large-scale circuits (50-gene networks) simulated are described in the results section.

2.5 Mutual Information and Lempel-Ziv Complexity

In the case of large-scale gene networks (i.e. with 50 nodes and mean number of input connections of 2, see below), we calculate the Mutual Information (MI) and Lempel-Ziv (LZ) complexity of their dynamics. We consider such dynamics to correspond to the protein numbers of each gene over time (as in [26]). For that, we first define time windows (each window containing 10 consecutive time moments where protein numbers were collected) and calculate the mean protein numbers within that window, for each gene. Next, we binarize these numbers in accordance with a fixed threshold. If, in a given window, the mean protein numbers is smaller than 200, then the 'binary protein value' is set to zero. Else, it is set to 1. The threshold value of 200 for protein numbers was set based on our observation that, when the corresponding gene is repressed, its proteins usually tended to be smaller than 100, while when the gene is unrepressed, they tend to be larger than 300.

Based on this binarized data, to study the global propagation of information in the gene network, we make use of the average pairwise MI, which is a measure of the degree of correlation between the dynamics over time of all genes of the network. In particular, we defined MI as follows. Let S_a be a process that generates 0 with probability p_0 and 1 with probability p_1. We define the entropy of S_a as:

$$H[s_a] = -p_0 \cdot \log_2 p_0 - p_1 \cdot \log_2 p_1 \tag{13}$$

Similarly, for a process S_{ab} that generates pairs xy with probabilities p_{xy}, where x, y $\in \{0,1\}$, we define the joint entropy as:

$$H[s_{ab}] = -p_{00} \cdot \log_2 p_{00} - p_{01} \cdot \log_2 p_{01} - p_{10} \cdot \log_2 p_{10} - p_{11} \cdot \log_2 p_{11} \tag{14}$$

Finally, the MI of the pair of genes *i* and *j* is [26]:

$$MI_{ij} = H\left[s_i\right] + H\left[s_j\right] - H\left[s_{ij}\right] \tag{15}$$

Given this definition, MI_{ij} measures the extent to which information about node *i* at time t influences, directly or not, node *j* one time step later. From this, to quantify the efficiency of information propagation throughout the entire network, assuming N to be the number of nodes, we define the average pairwise MI of a network as:

$$MI = N^{-2} \cdot \sum_{i,j=1,..,N} MI_{ij} \tag{16}$$

In addition to the average pairwise MI, again using the windowed, binarized protein numbers data, we further calculate the Lempel-Ziv (LZ) complexity of each gene's protein numbers over time (averaged over all genes) [27], as a means to quantify the degree of complexity of the signals that each gene of the network can generate.

In general, LZ measures a sequence's complexity over a finite alphabet (here {0,1}) by counting the number of new sub-strings (words) found, as the sequence is read (usually from left to right). For this, the algorithm used here [29] separates the sequence into shortest words that haven't occurred yet, and the complexity equals the number of unique words, except for the last word, which may not be unique [27,28]. Finally, assuming that n is the length of the time series of protein numbers from which the absolute LZ is calculated, we divide this absolute quantity by $\log_2(n)$ so as to scale it by the length, thus obtaining the scaled LZ for each gene. Then, we calculate the *average* scaled LZ of the network by summing the scaled LZ of each gene i and dividing by the total number of genes (N):

$$LZ = \frac{1}{N} \cdot \sum_{i=1,..,N} \left\{ LZ(i) \cdot \frac{\log_2(n)}{n} \right\} \tag{17}$$

3 Results and Conclusions

3.1 $\tau_{after}/\Delta t$ tunes the generation of cell-to-cell variability in RNA numbers from the cell-to-cell variability in RNAP numbers

We simulated the 70 models of individual genes described in the Methods section. Each model was simulated 100 times, each time for 2x10^4 seconds. From each simulation, we extracted the total number of produced RNA molecules during that time period.

As described in Methods, the models differ in such a way that the mean rate of RNA production should not differ. This was verified to be true. In all models, the total number of RNAs produced per cell equals ~20, as expected (Table 1).

In Figure 1, we show the CV^2 of the number of produced RNAs (CV^2(RNA)) in individual cells in each model. We find that, as $\tau_{after}/\Delta t$ increases, the CV^2(RNA) decreases. Meanwhile, the CV^2(RNA) grows with increasing CV^2(RNAP). More importantly, visibly, both $\tau_{after}/\Delta t$ and CV^2(RNAP) need to be tuned in particular ways so that the CV^2(RNA) reaches a maximum and a minimum, which is not possible otherwise.

We conclude that, while the CV^2(RNAP) 'propagates' to the CV^2(RNA), as expected [33], the degree with which it does so strongly depends on the promoter initiation kinetics (specifically, it differs depending on the value of $\tau_{after}/\Delta t$).

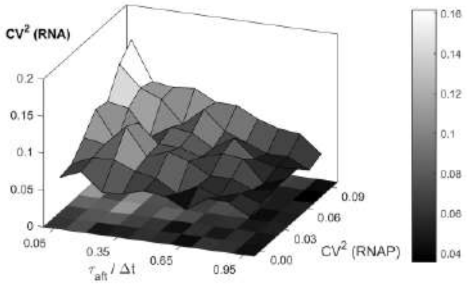


Fig. 1. CV^2 of number of produced RNAs in model cells versus $\tau_{after}/\Delta t$ and CV^2(RNAP).

3.2 $\tau_{after}/\Delta t$ tunes the influence of the cell-to-cell variability in RNAP numbers on the cell-to-cell variability in switching frequency of a 2-gene toggle switch

A 2-gene toggle switch consists of a genetic circuit of 2 genes that repress each other. Here, we model symmetric circuits, i.e., the 2 genes are identical. We simulated 70 models of toggle switches, differing in the dynamics of the component genes as described in Methods. Each model was simulated 100 times, each for 5x10^7 seconds, and protein numbers were assessed at a sampling interval of 10^4 s. In each simulation, we determined the moments when the protein numbers 'switched', as described in the Methods section, from which we obtained the mean and CV^2 of the switching frequency (F) for each model toggle switch. Results are shown in Figure 2.

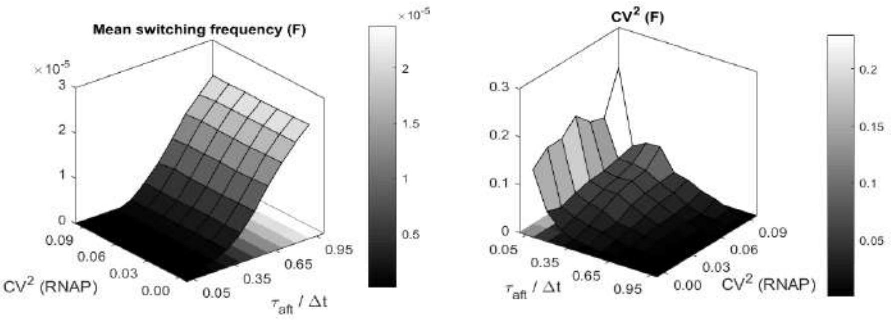


Fig. 2. (Left) Mean switching frequency (F) as a function of $\tau_{after}/\Delta t$ and CV^2(RNAP). (Right) Cell-to-cell diversity of the switching frequency (CV^2(F)) as a function of $\tau_{after}/\Delta t$ and CV^2(RNAP).

From Figure 2 (Left), decreasing $\tau_{after}/\Delta t$ decreases mean *F*, due to increased robustness of the 'noisy attractors' of the toggle switch [30], caused by a reduced probability of finding the 'repressed' promoter in a state that allows transcription to initiate. Meanwhile, changing CV^2(RNAP) does not affect the mean *F*, as this variable should not influence the mean behavior of the cell population, only the behavior diversity.

From Figure 2 (Right), we find that as the CV^2(RNAP) increases, so does the CV^2(F) (although mildly), provided that $\tau_{after}/\Delta t$ is smaller than ~0.7-0.8. This is because, the smaller is $\tau_{after}/\Delta t$, the weaker is the filtering of the extrinsic noise.

The CV^2(RNAP) affects the CV^2(F) weakly since, first, the CV^2(RNAP) only affects the cell-to-cell variability of the mean $\tau_{prior}/\Delta t$ and, thus, if this step is short-length, the effects on the variability in RNA production kinetics will be weak. Second, for relatively small values of $\tau_{prior}/\Delta t$ compared to $\tau_{after}/\Delta t$, the switch becomes less stable [11], causing the effects of increasing CV^2(RNAP) to become negligible.

3.3 Large-Scale Circuits. Information Generation and Propagation

We simulate networks with 50 nodes and mean number of input connections (i.e. mean connectivity) of 2. The network topologies are generated using the 'Erdös Random 2' algorithm proposed in [34] that allow producing Erdös random graphs [35].

In these networks, each node is a gene whose expression dynamics and interactions (i.e. connections) are defined according to reactions (1-6). Consequently, all interactions between genes consist of repression mechanisms, as those used to model the genetic switches above. Given this modelling strategy, at any given moment, a gene is expected to be either actively expressing (in the absence of its repressor proteins, which can be more than 1) or to be repressed. We hypothesize that the degree of repression should depend on the number of genes that can directly repress a given gene, on the relative time length during which the repressing genes are active, and, importantly, as seen above, on the value of $\tau_{after}/\Delta t$ of the repressed gene.

To study how changing $\tau_{after}/\Delta t$ and CV^2(RNAP) affects the networks ability to generate and propagate information, for each set of values of $\tau_{after}/\Delta t$ and CV^2(RNAP), we generate 10 topologies. Then, we simulated each topology 10 times, with individual simulations differing in RNAP numbers as above.

Each simulation lasts 10^6 seconds, and the numbers of each protein were collected each 10^3 seconds. Each such set of collected protein numbers constitutes a network 'time moment'. From these, network 'states' over time are obtained as follows. First, time windows with a length of 10 time moments are defined (the first window is not considered since the network is initialized without proteins). Next, the protein numbers in each state are obtained by averaging these numbers from all 10 time moments composing the window. Finally, for each protein, we binarize its mean numbers of each time window using a fixed threshold (see Methods).

Thus, from each simulation, we obtained 100 consecutive 'binary states' of the given network. From this data, we can then calculate the average pairwise MI and the average scaled LZ, so as to measure, respectively, the degrees of information propagation and generation of the network during that time period. Finally, for each condition, we obtained the averages of these two quantities for all networks.

Six models of gene expression differing in $\tau_{after}/\Delta t$ and CV^2(RNAP) were considered (Table 3). Note that, as described in Methods, all genes of a given network share the same value of $\tau_{after}/\Delta t$ and all networks of a given model share the same value of CV^2(RNAP). The values for these two parameters were chosen so as to test whether they can affect the mean and variability of the information generation and propagation capabilities of the networks. Results are shown in Table 3.

Table 3. Pairwise mutual information (MI) and scaled Lempel-Ziv complexity (LZ) mean (μ) and CV^2 as a function of $\tau_{after}/\Delta t$ of the promoters and the CV^2(RNAP) of the cell populations.

Condition	$\tau_{after}/\Delta t$	CV^2(RNAP)	μ(MI)	μ(LZ)	CV^2(MI)	CV^2(LZ)
1a	0.1	0.01	0.004	0.18	3.4	0.0016
1b	0.1	0.08	0.005	0.18	1.6	0.0018
2a	0.5	0.01	0.008	0.48	0.002	0.0007
2b	0.5	0.08	0.008	0.48	0.002	0.0009
3a	0.9	0.01	0.0007	0.17	0.05	0.0016
3b	0.9	0.08	0.0007	0.17	0.04	0.0016

From Table 3, we find that the value of $\tau_{after}/\Delta t$ of the component genes affects the values of μ(MI), μ(LZ), CV^2(MI), and CV^2(LZ). This is expected, since this parameter affects the degree to which a gene's activity is affected not only by variability in RNAP numbers but also by its repressor genes' activity levels.

Meanwhile, the CV^2(RNAP) affects the CV^2(MI) and CV^2(LZ), provided small values of $\tau_{after}/\Delta t$, as expected given the results for the toggle switch model.

4 Discussion

We performed simulations of stochastic models of single gene, 2-gene toggle switches, and large-scale (50 genes) genetic circuits, all of which include the multi-step process of transcription, whose parameter values have been obtained from empirical data extracted from *in vivo*, single-cell measurements on *E. coli* cells.

Overall, we find that the relative time that the gene spends in the rate-limiting steps *after* initiation of the open complex formation, here quantified by the ratio $\tau_{after}/\Delta t$, significantly affects the degree to which individual genes and circuits (small- and large-scale) are affected by extrinsic noise.

It is of interest that this property of the promoter initiation kinetics has a clear multi-scale effect, ranging from effects on RNA numbers of individual genes over time, to effects on the dynamics of small network motifs, to effects on the capacity of large networks to produce and propagate information. The above suggests that this feature of the kinetics of transcription may be used as a 'master regulator' of the functioning of the genetic circuits in *E. coli*, perhaps as influent as the global and local topological structures formed by promoter-protein, protein-protein, and RNA-RNA interactions.

Importantly, the kinetics of transcription initiation of each gene in the network is both sequence dependent as well as subject to regulation, both by transcription factors as well as by global regulatory molecules, such as σ factors. As such, this mechanism is both, respectively, evolvable as well as adaptive at the single gene level. We hypothesize that, given this, the 2 rate-limiting step nature of the transcription process may confer *E. coli* rapid evolvability as well as plasticity in fluctuating environments.

In the future, we plan to perform a wide range of experiments to validate our findings, as well as to make use of additional simulation and more detailed models to further explore how the dynamics of transcription of individual genes may act as a regulator of the degree of influence of extrinsic noise on genetic networks.

5 Acknowledgements

Work supported by Academy of Finland (295027 ASR), Academy of Finland Key Project Funding (305342 ASR), Jane and Aatos Erkko Foundation (610536 ASR), Finnish Academy of Science and Letters (SO), and Tampere University of Technology President's Graduate Program (SS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Jishage, M., Iwata, A., Ueda, S., Ishihama, A.: Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions. J. Bacteriol. 178, 5447–5451 (1996).
2. Rahman, M., Hasan, M.R., Oba, T., Shimizu, K.: Effect of rpoS gene knockout on the metabolism of *Escherichia coli* during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations. Biotechnol. Bioeng. 94, 585–595 (2006).
3. Megerle, J. A., Fritz, G., Gerland, U., Jung, K. & Rädler, J. O. Timing and dynamics of single cell gene expression in the arabinose utilization system. Biophys. J. 95, 2103–2115 (2008).
4. Jones, D. L., Brewster, R. C. & Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. Science 346, 1533–1537 (2014).
5. Farewell, A., Kvint, K., Nyström, T.: Negative regulation by RpoS: A case of sigma factor competition, Mol. Microbiol. 29, 1039–1051 (1998).
6. Hengge-Aronis, R.: Recent insights into the general stress response regulatory network in *Escherichia coli*. J. Mol. Microbiol. Biotechnol. 4, 341–346 (2002).
7. Kandavalli, V.K., Tran, H., Ribeiro, A.S.: Effects of σ factor competition on the in vivo kinetics of transcription initiation in *E. coli*. BBA Gene Regul Mech 1859, 1281–8 (2016).
8. McClure, W.R.: Rate-limiting steps in RNA chain initiation. Proc. Natl Acad. Sci. USA 77, 5634–5648 (1980).
9. Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J., Goncalves, N., Oliveira, S.M.D., Häkkinen, A., Ribeiro, A.S.: Dissecting the stochastic transcription initiation process in live *Escherichia coli*. DNA Res. 23(3), 203-214 (2016).
10. Bahrudeen, M.N.M., Startceva, S., Ribeiro, A.S.: Effects of Extrinsic Noise are Promoter Kinetics Dependent. In: The 9th International Conference on Bioinformatics and Biomedical Technology on Proceedings, pp. 44-47. ICBBT 2017, Lisbon, Portugal (2017).
11. Bahrudeen, M.N.M., Startceva, S., Ribeiro, A.S.: Tuning extrinsic noise effects on a small genetic circuit. In: The European Conference on Artificial Life on Proceedings. ECAL 2017, Lyon, France (2017). *In press*.

12. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81(25), 2340–2361 (1977).
13. Lloyd-Price, J., Gupta, A., Ribeiro, A.S.: SGNS2: A Compartmentalized Stochastic Chemical Kinetics Simulator for Dynamic Cell Populations. Bioinformatics 28, 3004-5 (2012).
14. Bernstein, J.A., Khodursky, A.B., Pei-Hsun, L., Lin-Chao, S. Cohen, S. N.: Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl Acad. Sci. USA 99, 9697–9702 (2002).
15. Taniguchi, Y., Choi, P.J., Li, G.-W., et al.: Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329, 533–538 (2010).
16. Mitarai, N., Sneppen, K., Pedersen, S.: Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. J. Mol. Biol. 382, 236-245 (2008).
17. Bremer, H. and Dennis, P.P.: Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. In: Neidhardt, F.C., (ed.), *Escherichia Coli* and *Salmonella*, 2nd ed. ASM Press, Washington, DC, 1553–1569 (1996).
18. Kennel, D., and Riezman, H.: Transcription and translation initiation frequencies of the *Escherichia coli* lac operon. J. Mol. Biol. 114(1), 1-21 (1977).
19. Cormack, B.P., Valdivia, R.H., Falkow, S.: FACS-optimized mutants of the green fluorescent protein (GFP). Gene 173(1), 33-38 (1996).
20. Neubauer, Z., and Calef, E.: Immunity Phase-shift in Defective Lysogens: Non-mutational Hereditary Change of Early Regulation of λ Prophage. J. Mol. Biol. 51, 1-13 (1970).
21. Arkin, A., Ross, J., McAdams, H.: Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage λ-Infected *Escherichia coli* Cells. Genetics 149, 1633-1648 (1998).
22. Fu, Y., Jarboe, L. R., Dickerson. J.A.: Reconstructing genome-wide regulatory network of *E. coli* using transcriptome data and predicted transcription factor activities. BMC Bioinformatics 12(233), DOI: 10.1186/1471-2105-12-233 (2011).
23. Saecker, R.M., Record, M.T. Dehaseth, P.L.: Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. J. Mol. Biol. 412, 754–771 (2011).
24. Chamberlin, M.: The selectivity of transcription. Annu. Rev. Biochem. 43, 721–75 (1974).
25. deHaseth, P.L., Zupancic, M.L. and Record, M.T.: RNA polymerase promoter interactions: The comings and goings of RNA polymerase. J. Bacteriol. 180, 3019–3025 (1998).
26. Ribeiro, A.S., Kauffman, S.A., Lloyd-Price, J., Samuelsson, B., Socolar, J.E.S.: Mutual information in random Boolean models of regulatory networks. Phys. Rev. E 77, 011901 (2008).
27. Lempel, A., and Ziv, J.: On the Complexity of Finite Sequences. IEEE Trans. Inform. Theory 22, 75-81 (1976).
28. Shmulevich, I., Kauffman, S.A., Aldana, M.: Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc. Natl Acad. Sci. USA 102(38), 13439-13444 (2005).
29. Borowska, M., Oczeretko, E., Mazurek, A., Kitlas, A., Kuć, P.: Application of the Lempel-Ziv complexity measure to the analysis of biosignals and medical images. Annual Proceedings of Medical Science, Suppl. 2, 50 (2005).
30. Ribeiro, A.S., and Kauffman, S.A.: Noisy Attractors and Ergodic Sets in Models of Gene Regulatory Networks, J. of Theor. Biol. 247(4), 743-755 (2007).
31. McClure, W.R.: Mechanism and control of transcription initiation in prokaryotes. Annu Rev Biochem. 54, 171-204 (1985).
32. Roussel, M.R. and Zhu, R.: Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. Phys. Biol. 3, 274–84 (2006).
33. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S.: Stochastic gene expression in a single cell. Science 297, 1183-1186 (2002).

34. Airoldi, E.M., and Carley, K.M.: Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings. ACM SIGKDD Explorations Newsletter 7(2), 13-22 (2005).
35. Bollobas. B. Random Graphs. Academic Press, New York, 2nd edition, (2001).

SB-AI: How the synthetic biology paradigm is impacting AL and AI research

Luisa Damiano[1,2] and Pasquale Stano[3]

[1] Epistemology of the Sciences of the Artificial Research Group (ESARG),
Department of Ancient and Modern Civilizations, University of Messina, Messina,
Italy
[2] Centre for Research on Complex Systems (CERCO), University of Bergamo, Italy
[3] Department of Biological and Environmental Sciences and Technologies
(DiSTeBA), University of Salento, Lecce, Italy
luisa.damiano@unime.it
pasquale.stano@unisalento.it

Abstract. A novel scenario is emerging from synthetic biology advancements of the last fifteen years, namely, a well-defined multidisciplinary sci-tech arena focused on the construction of biological-like systems, and in particular microscopic cell-like systems. The challenge of assembling a minimal cell from its separated parts is generally considered one of the most important Holy Grail of biology, but an accurate analysis, based on the theory of autopoiesis and its implications, discloses a fecund field from which artificial life and artificial intelligence might benefit and profit for a perspective change. Here we comment some recent trends and experimental results in synthetic biology, explaining in general terms how these can impact artificial life and artificial intelligence research, in particular with respect to embodied cognition.

1 The Vision

In 1943 Arturo Rosenblueth, Norbert Wiener and Julian Bigelow concluded their article on *Behavior, Purpose and Teleology* [31], inaugurating the era of cybernetics, with a visionary remark:

> "If an engineer were to design a robot, roughly similar in behaviour to an animal organism, he would not attempt at present to make it out of proteins and other colloids. He would probably build it out of metallic parts, some dielectrics and many vacuum tubes. The movements of the robot could readily be much faster and more powerful than those of the original organism. Learning and memory, however, would be quite rudimentary. In future years, as the knowledge of colloids and proteins increases, future engineers may attempt the design of robots not only with a behaviour, but also with a structure similar to that of a mammal. The ultimate model of a cat is of course another cat, whether it be born of still another cat or synthesized in a laboratory."

Today, while engineering research on mechanical robots is impressively advancing, biology is integrating fifty years of progress in chemistry, biophysics and, in general, the sciences of life to proceed in the direction indicated by these pioneers of cybernetics a way leading to bio-robots, and maybe, one day, even to approximations of what they called 'ultimate' models of living beings. Starting from 1953, the *annus mirabilis* of biology (due to the Watson-Crick discovery of DNA helical structure, the completion of the first protein – insulin – sequencing by Sanger, and the Urey-Miller experiment), our understanding of 'proteins and other colloids' increased at a stunning level, especially in the last decades. After having explored and analysed the biological systems according to the analytic or 'taking apart' methodology, in the early 2000 a new wave of biological studies emerged, explicitly inspired to engineering. This is synthetic biology (SB), a branch of biology and bioengineering aiming at *constructing*, or 'putting together', biological parts, devices and systems, which do not currently exist in the natural world, for useful purposes.

Hence, SB targets the design and the construction of programmable synthetic cells by developing first, and using later, parts-devices-systems that ultimately work by exploiting the so-called intrinsic 'computational' capability of molecules. SB exploit is generally seen as a technical one, based on the availability of powerful bio-analytical techniques, especially high-throughput ones, progress in synthetic capability (synthesis of genes), and the attitude of a young generation of scientists toward the blending and the convergence of biology and engineering.

However, SB also has another soul, due to its very peculiar manner of contributing to the scientific understanding of life, and for this reason it stimulates philosophers to study the epistemological and theoretical roots of SB approaches [26, 21]. The attention is primarily focused on the way in which SB generates knowledge, based not on developing abstract representations, but on creating material models - that is, concrete physico-chemical models - of the biological processes under inquiry, and the underlying biological mechanisms. Such an 'understanding-by-building' methodological approach [29, 5] has a long tradition. Introduced by proto-cybernetic movements between the 1910s and the 1930s for the modelling of biological and cognitive processes through mechanical artefacts (hardware models), it has become a recognised scientific method with the birth of cybernetics in the 1940s, and has been developed by classical Artificial Intelligence (AI) and Artificial Life (AL) through computer simulations (software models), and by the emerging 'embodied AI' through new generations of mechanical robots (again, hardware models).

The novelty generated by SB, which makes the link with the above quote, is that now we can build chemical models (wetware models) of natural processes at the molecular, supramolecular and cellular levels, precisely because our knowledge of 'colloids and proteins' has increased so much in the past years, and SB has developed a variety of constructive approaches.

In this short essay we will discuss with a certain detail the so-called *bottom-up semi-synthetic approach*, which was introduced in the early 1990s by researchers in origins-of-life (and in particular by the seminal work of P. L. Luisi, P. Walde and T. Oberholzer at the ETH-Zürich [19]). The article is composed of three parts, respectively dedicated to: (i) theory and construction of semi-synthetic minimal cells; (ii) chemical communications between synthetic cells and biological cells; (iii) relevance and implications of the recent advancements in SB for the creation of fecund interdisciplinary research embracing SB and AI, or AL. The theoretical framework defining our perspectives on these developments relies on autopoietic cognitive biology [24, 23, 22] and embodied cognitive science [39].

2 Theory and construction of semi-synthetic minimal cells

Mainstream SB operates according to the biobrick philosophy (`parts.igem.org`), namely, the construction of molecular devices – generally genetic circuits – from standardized parts, whereby by parts a DNA molecule is intended. Just like electronic engineers design and build devices from electronic parts [17, 9], synthetic biologists operate with genes and regulatory proteins, and plug them in the biological chassis, i.e., the cell with its core set of genetic-metabolic circuitry.

An authentic new wave in SB considers the possibility of constructing simplified (minimal) cells [18] that might be somehow functional, despite a strong reduction of biological complexity, to a level that the resulting system looks more like a primitive (ancient) cell than a modern evolved – and therefore highly sophisticated – cell. These cell-like compartments, called semi-synthetic minimal cells (SSMCs), are interesting in many respects, such as:

1. in origins-of-life research, they help understanding basic mechanisms leading to the transition from inanimate to living matter;
2. in basic science, they allow understanding biochemical-biophysical processes without the interference of the other cellular processes in background;
3. in SB, they can be used to develop systems performing specific functions;
4. they are composed using dozens of separated, yet well-characterized, molecular parts (proteins, nucleic acids, lipids, etc.);
5. they can be built in the laboratory according to a novel technology based on the convergence of cell-free systems, liposome technology, and microfluidics;
6. a precise mathematical modelling can be applied to SSMCs processes;
7. despite their simplicity, SSMCs have several cell-like properties (though in a rudimentary form) and might display emergent properties;
8. although creating *living* SSMCs is the long-term goal of this research, non-living SSMCs work well for most biotechnological applications;
9. SSMCs might represent a concrete attempt to experimentally investigate autonomy, autopoiesis and cognition.

The SSMCs construction is generally acknowledged as the SB bottom-up route, as opposite to the mainstream approach that is generally dubbed as top-down SB (for a critical discussion, see [34]). For constructing SSMCs, the minimal number of biological macromolecules (DNA, ribosomes, enzymes) are encapsulated inside liposomes (Figure 1). Early attempts started in the 1990s, and today

SSMCs can be constructed in a rather complex way, produce proteins [35], and perform simple biological-like functions.

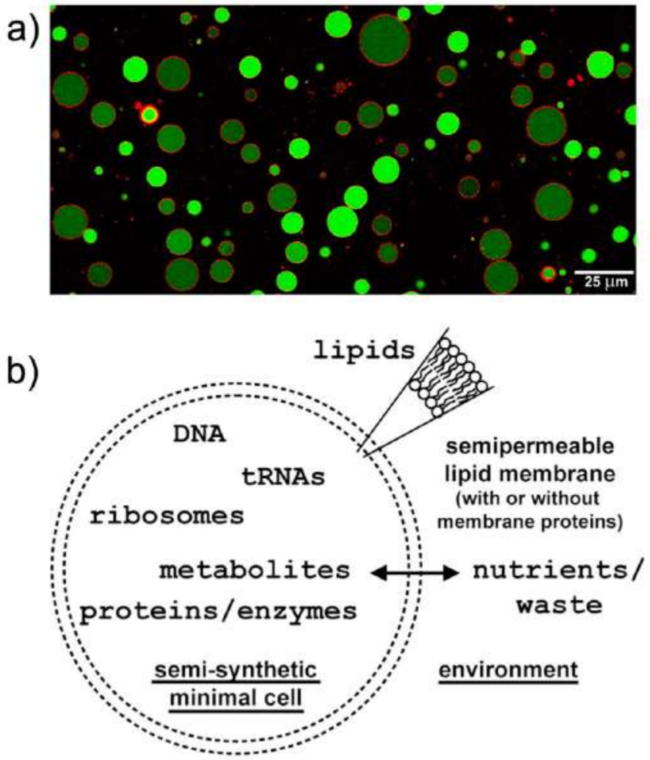


Fig. 1. SSMCs made of solute-containing liposomes. (a) Giant lipid vesicles are often used for constructing cell-like systems. The picture shows calcein-filled vesicles whose membranes have been stained by Trypan Blue.Reproduced from [36] according to the CC-BY license. (b) The minimal number of genes, enzymes, RNAs, and low molecular-weight compounds are encapsulated into synthetic lipid-based compartments, such as in the case of lipid vesicles. The membrane acts as a boundary to confine the interacting internalized molecules and allow selective passage of some molecules (nutrients, waste). Reproduced from [33] according to the CC BY-NC-SA 3.0 license.

Autopoiesis (self-production) is the theory that guide us in designing and constructing SSMCs [23, 20]. The theory of autopoiesis deals with the question “what is life?” and finds the answer in the specific form of dynamical organization that, according to this theory, defines all biological systems. Autopoiesis argues that:

1. the distinctive property of living systems is its autopoiesis, i.e., their capability of producing and maintaining their material identity (themselves) by producing their own components (metabolism);
2. being autopoiesis a global property, its realization does not rely on the systems' components taken separately, but on the way in which these components are organised within living systems;
3. in its minimal manifestation, given at the level of minimal cells, the autopoietic organization is a self-regenerating network of operations of synthesis and destruction of components (metabolism) that defines by itself its topological limits through the creation of a material separation from the external environment;
4. the intrinsic and observer-independent 'purpose' of an autopoietic system is just its own self-maintenance;
5. the changes in the environment affecting an autopoietic system have to be considered as perturbations of its internal dynamic, to which the system can react through conservative processes of self-regulation;
6. the self-regulative adaptive activity of autopoietic systems can be interpreted as a cognitive activity, and implies the possibility of recognising the minimal form of natural cognition in the self-regulative processes of self-production of the minimal living units (that is, the self-production of minimal biological bodies), which entails a strong link between autopoiesis and the concept of *embodied cognition.*

According to our view, autopoiesis constitutes the best theoretical framework for the development of any man-made synthetic cell, and SSMCs in particular. Early work was dedicated to realize simple systems, like the self-reproducing reverse micelles [2] and vesicles [40], which, owing to their simplicity and to the exploitation of surfactants self-organization, displayed (quasi-)autopoietic, or autopoietic-like, dynamics.

Moving from very basic chemical systems (with a dozen or different molecules, or so) to more complicated cell-like ones, like the SSMCs, the attempts of creating autopoietic dynamics becomes significantly more difficult to be implemented. This is due to the fact that the production of such complex components (protein, nucleic acids, ...) relies on complex biochemical mechanisms based on other macromolecules, which in turn needs to the produced according to the autopoietic (self-production) scheme.

To be more specific, current SSMCs are assembled by combining liposome technology and cell-free technology [35, 13, 3], for example by the incorporation of the PURE system (a transcription-translation system) inside liposomes, aiming at synthesizing proteins and thus letting the SSMCs performing a specific function. However, for genuine autopoietic SSMCs, the reaction network should also produce the PURE system itself, and at the expenses of external resources.

It is important to remark, thus, that even if the final goal of building an autopoietic synthetic cell from scratch remains the principal purpose (with important conceptual consequences for science in general, an in particular for chemistry and biology), all intermediate structures which do not necessarily display an autopoietic organization, or display only parts of it, are *per se* interesting milestones to reach, both to contribute to our understanding of biological systems (understanding-by-building), and to develop novel tools and applications in biotechnology.

One of these aspects will be discussed in the next section, dealing with synthetic cells capable of exchanging chemical signals with biological cells.

3 Chemical communications between synthetic cells and biological cells

Synthetic cells can be built in order to recognize a chemical signal and behave accordingly, as it happens between biological cells. This is generally discussed, also by us, in terms of performing a kind of logical operation (i.e., computing), and thus as a route to the novel bio-chemical information and communication technologies (bio-chem ICTs). On the other hand, developing an autopoiesis-based view, these mechanisms are better interpreted as attempts of autopoietic systems to conservatively (i.e., self-regulatively) react to externally-generated 'perturbations'. In 2012, we specified that [37]:

> "... [autopoietic systems] can perceive some external variations as perturbations of their internal process of self-production. Besides, they can react to them through an activity of self-regulation, that is, through changes in their elementary processes that compensate the alteration. In this sense, these systems can be conceived able of generating internal operational meanings for the perceived external variations. These meanings are expressed in terms of dynamical schemes of self-regulation, which externally appears as actions oriented to conservation (e.g., absorbing a molecule of sugar, overcoming an obstacle...). This 'meaning generation' behaviour – for Maturana and Varela the basic 'cognitive' behaviour – grounds what the two researchers called 'structural coupling' with the environment: a dynamic of reciprocal perturbations and compensation, in which the autopoietic system continuously generate and associate to exogenous variations operational meanings of self-regulation that allows it to keep its process of self-production in an ever-changing environment."

Thus, the recognition capabilities which are intrinsic to the molecular domain can be used (1) to build this kind of synthetic cells (capable of exchanging chemical signals, or, more in general, of expressing adaptive responses to perturbations), and, on this basis, (2) to further explore the radically embodied autopoietic views on cognition mentioned above. An analysis of the deep implications of these perspectives exceeds the scope of this contribution, and we ill develop it in future publications. Here we ould like to underline that, as remarked also in Section 4, these developments might profoundly impact AI research and its underlying research paradigms.

Coming back to traditional bio-chem ITCs, the relevance of inter- and intracellular molecular communication has been put forward by the Suda-Nakano group [28], and its application to nanomedicine has been lucidly defined by [14] (Figure 2). For example, SB genetic, regulatory, metabolic circuitry could be adapted to this scope – consider for instance SSMCs endowed with plugged-in circuits.

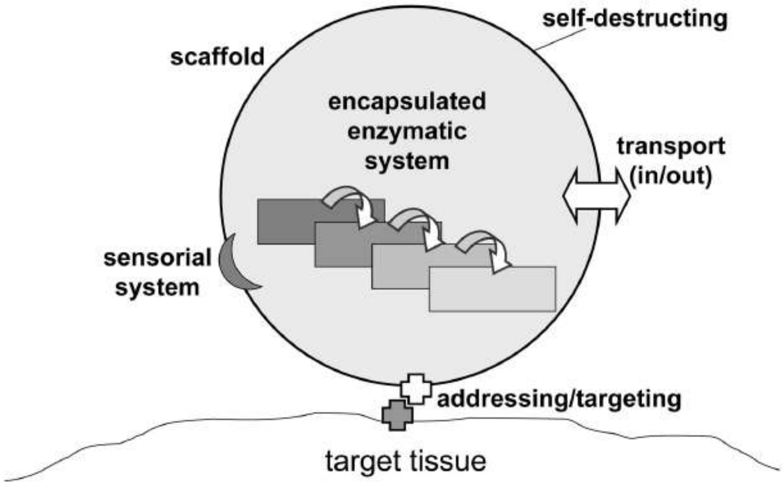


Fig. 2. By advancing the current biotechnology for SSMCs construction, medical potential applications of more sophisticated, communicating and programmable synthetic cells can be envisaged. A goal would be to construct cell-like systems that, once injected in human body, reach a specific target region and thanks to chemical information processing, is able to act accordingly, for example producing in situ a cytotoxic drug or a stimulus to trigger a cellular response. Very recently, the concept of 'pseudo-cell factories' or 'nanofactories' (illustrated above) has been proposed by Le Duc and collaborators [14]. Figure reproduced from [38] with the permission of Elsevier. The drawing follow what has been sketched in [14].

Experimental work has been carried out in recent years. In 2009 Ben Davis and collaborators published a paper where it was shown that vesicles could synthesize a sugar-like molecule by the formose reaction [10]. The molecule could be released in the medium, converted to a borate derivative, and could thus stimulate a response in receiving cells (bacteria). Mansy and collaborators built synthetic cells acting as 'translators' for *E. coli* [16], and more recently reported a two-way chemical communication between synthetic cells and bacteria [15], based on acyl homoserinelactones. Furthermore, Adamala's research group demonstrated chemical communication between synthetic cells [1]. Our group is also involved in this kind of research since 2012, promoting this approach [37, 30] also based on numerical simulations [25], and reporting case studies (Rampioni et al., manuscript in preparation).

3.1 A Turing test for synthetic cellularity?

As we commented in a previous work [30, 4], an interesting and somehow provocative scenario was proposed by Cronin, Krasnogor, Davis and collaborators in 2006 [6], arguing that a sort of Turing test, targeting minimal communication skills, could help to determine whether a system is alive (at the cellular level), bypassing the issue of clarifying what life is, as the original Turing test was devised to determine whether a system is intelligent, bypassing the problem of defining what intelligence is. As related literature pointed out, in AI, the artificial imitation of a cognitive behaviour might not be based on the *generative mechanism* producing the target cognitive behaviour, abrogating the actual relevance of a Turing test). Anyhow, in the case of synthetic cells built by SB, and due to their constitutive molecular nature, this kind of imitation game might be significant. At this minimal level, will the synthetic/artificial cells really reproduce the cognitive pattern of a natural-biological partner? Will the synthetic vs. natural barrier indeed become ill-defined and possibly disappear? [4]

It is interesting to shortly consider the attempt, by Mansy and colleagues [15], of *quantifying* the life-likeness (or better, *Vibrio fisheri*-likeness) of synthetic cells capable of sending and receiving signals to/from th bacterium *V. fischeri*. The quantification was done on the basis of RNA sequencing (i.e., determining the gene expression profile of natural cells in response to the activity of the synthetic cells). From a series of comparative tests (*V. fischeri* vs *V. fischeri*, *V. fischeri* vs non-functional synthetic cells, and *V. fischeri* vs functional synthetic cells), it was determined that their synthetic cells were 39% life-like. As the same authors remarked, the genetically-encoded function-related elements in synthetic cells were only two (LuxR and LuxI), whereas a synthetic cell, at least conceptually, is based on more than 100 genes only to encode the required transcription-translation machineries. The 39% value refers therefore to an ideal extrapolation of the observed behaviour to a synthetic cell that produce all its proteins and nucleic acid. As mentioned above, such system is still out of reach, however.

4 SB-AI

The connection between SB 'understanding-by-building' approach and AI research is novel, although it is implicit in the autopoietic theory and the view of living systems that springs from that. Can SB, and in particular research on synthetic cells, be useful in AI inquiries? In principle, the answer is positive in relation to the new embodied AI, which, differently from classic AI, acknowledges a deep unity between cognitive and biological processes. We refer in particular to radical approaches to embodied cognition, which, in line with the autopoietic theory, ground the cognitive mind in the biological processes of construction and maintenance of the biological body. Within these explorative contexts, the SB 'understanding-by-building' research on minimal cells, by offering the possibility of building material (chemical) models of minimal living systems, opens the prospect of research scenarios directed towards experimentally studying minimal embodied cognition and related processes (molecular recognition, organizational closure, self-organization, autopoiesis, self-regulation, dynamical coupling between self-organizing or autopoietic systems and their environments, among others). These processes stem from the molecular dimension, molecular dynamics, strength of entropy at the molecular level, molecular energies and interplay with the thermal background [12], which cannot be found in macroscopic objects. It is not by chance that life originated at the molecular and supramolecular level. Moreover, and differently from electronic computers, molecular systems and their operations can be interpreted as forms of computing, although with properties which differ from the one we are used to consider. These are well specified and commented in the work of Nakano [28, 27].

We have recently promoted, by means of a series of workshops and of a journal special issue, the possibility of a synergic cross-fertilization between SB and AI [7] focused on minimal embodied cognition. In what follows, we draw the main lines of a possible SB-AI approach to the study of minimal cognition based on the theory of autopoiesis an approach that we call Chemical Autopoietic AI.

4.1 Chemical Autopoietic AI: drawing the basic lines of a SB-AI research program

The potentialities that autopoiesis can express in AI rely on its definition of life, based on two main reasons. The first, related to the methodological approach of autopoietic biology, refers to the "synthetic" nature of this definition. Maturana and Varela's answer to the question "What is life?" is not analytical, as traditional definitions of life are. Autopoiesis answers to this question not by proposing a list of properties of living systems, but by theoretically defining a mechanism able to generate, from a multiplicity of components, a minimal living system endowed with the features making it potentially able to produce the whole biological domain as we know it. The idea of providing a "synthetic" definition of life is essentially this: theoretically determining a mechanism able to generate, from scratch, the whole biological phenomenology. Its interest for the sciences of the artificial, and AL and AI in particular, is evident: this definitional approach promises that, when these sciences properly implement the mechanism through which autopoiesis defines life, in principle they will be able to artificially generate all biological processes processes that, according to Maturana and Varela, are intrinsically cognitive processes. In the pioneers of cybernetics' words: providing a material model of the autopoietic definition of life would open to the sciences of the artificial the possibility of creating 'ultimate' models of living and cognitive systemsm i.e., artificial, but genuinely living and cognitive systems. The second reason of the relevance of the autopoietic definition of life for AI relies on the theoretical content of the autopoietic definition of life, given by the notion of autopoietic organization, according to which a mechanism able to generate minimal living and cognitive systems, and, through them, all biological and cognitive processes is:

> . . . a network of process of production (transformation and destruction) of components that produces the components which: (i) through

> their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network . . . [23] (p. 79).

The specificity of the notion of autopoietic organization, as it clearly emerges from its definition, is that it is independent from the definition of specific components. This implies that, in principle, the sciences of the artificial engaged in the project of creating material models of the autopoietic organization do not have necessarily to focus on the actual components of life as we know it, but can use all kinds of components able to generate the network described by the autopoietic definition of life. This gives AL and AI the theoretical possibility to implement different material models of life and cognition, that is, (a variety of) forms of biological and cognitive processes that, with respect to their material structure, do not exist in nature. These two characteristics of the autopoietic definition of life its synthetic character and its independence from specific components makes it particularly interesting for SB research on synthetic cell construction, and, in particular, for its subsections focusing on the convergence between liposome technology and cell-free systems, with respect to which, as we previously pointed out, the theory of autopoiesis, as a theoretical framework, defines the long-term goal of implementing the autopoietic definition of minimal living systems building them in labs.

This goal allows SB to aspire to actively contribute to AI research, with a significant advantage with regard to other ways of modeling life and cognition based on autopoiesis. While computer simulations can provide only abstract artificial models of the autopoietic definition of life, and current mechanical robotics appears to be far from the possibility of generating material models of the dynamic network that this definition describes, SB, operating with molecular components that can differ from those of terrestrial life for chemical structure, but not reactivity in general terms, is in principle able to generate embodied modelizations of this kind of network, and, as we showed before, is already engaged in designing primitives versions of it. Its main obstacle, of course, is the complexity of even minimal autopoietic networks. As we will show in a future publication, however, this does not preclude the possibility of building simplified versions of the autopoietic organization.

The promises of this kind of SB-AI approach would be extremely relevant for the evolution of AI:

1. as we showed elsewhere, this approach would provide AI with an experimentally explorable material model of an "intrinsically intentional" cognitive agent namely, a cognitive agent whose relations with the environment are charged with "intrinsic" meanings, depending on the conservation of its organization and its ways of existence , which, as literature emphasizes since Dreyfus' "What Computers Cannot Do" (1979) [8], Searle's "Chinese Room Argument" (1980) [32], and Harnad's "Symbol Grounding Problem" (1990) [11], is one of the main weak point of (both traditional and embodied);

2. the experimental exploration of this notion through its basic wetware modelization could include not only the structural coupling of the minimal synthetic autopoietic system with its environment, but also to the evolutive dynamics of interactions with this niche that could lead it to develop higher level of organizational complexity, as planned by the autopoietic theory;
3. both explorations could have applicative implications leading to hybrid biomechanical robots and, more in general, synthetic cognitive systems based on the autopoietic, and, more in general, a radical embodiment approach.

The above described difficulties in fully implementing this SB-AI are currently leading us to develop a simplified version of it, which we plan to fully describe in future works, and we consider as a first step in the inauguration of the Chemical Autopietic AI approach.

5 Acknowledgements

The authors thank Pier Luigi Luisi (Roma Tre University and ETH Zürich) for inspiring discussions. This work has been stimulated by our involvement in the European COST Action CM-1304 "*Emergence and Evolution of Complex Chemical Systems*" and TD-1308 "*Origins and evolution of life on Earth and in the Universe (ORIGINS)*".

References

1. Adamala, K.P., Martin-Alarcon, D.A., Guthrie-Honea, K.R., Boyden, E.S.: Engineering genetic circuit interactions within and between synthetic minimal cells. Nature Chemistry 9(5), 431–439 (2017)
2. Bachmann, P.A., Walde, P., Luisi, P.L., Lang, J.: Self-replicating reverse micelles and chemical autopoiesis. Journal of the American Chemical Society 112(22), 8200–8201 (1990)
3. Blain, J.C., Szostak, J.W.: Progress Toward Synthetic Cells. Annual Review of Biochemistry 83(1), 615–640 (2014), http://dx.doi.org/10.1146/annurev-biochem-080411-124036
4. Bracciali, A., Cataldo, E., Damiano, L., Felicioli, C., Marangoni, R., Stano, P.: From Cells as Computation to Cells as Apps. In: History and Philosophy of Computing. pp. 116–130. Springer, Cham (2015)
5. Cordeschi, R.: The Discovery of the Artificial. Springer (2002)
6. Cronin, L., Krasnogor, N., Davis, B.G., Alexander, C., Robertson, N., Steinke, J.H.G., Schroeder, S.L.M., Khlobystov, A.N., Cooper, G., Gardner, P.M., Siepmann, P., Whitaker, B.J., Marsh, D.: The imitation game–a computational chemical approach to recognizing life. Nat. Biotechnol. 24(10), 1203–1206 (Oct 2006)
7. Damiano, L., Kuruma, Y., Stano, P.: What can synthetic biology offer to artificial intelligence (and vice versa)? Biosystems 148, 1–3 (2016)
8. Dreyfus, H.L.: What computers cant do: The limits of artificial intelligence. Harper and Row, New York, revised edition edn. (1979)

9. Endy, D.: Foundations for engineering biology. Nature 438, 449–453 (2005)
10. Gardner, P.M., Winzer, K., Davis, B.G.: Sugar synthesis in a protocellular model leads to a cell signalling response in bacteria. Nat Chem 1(5), 377–383 (Aug 2009)
11. Harnad, S.: The symbol grounding problem. Physica D 42, 335–346 (1990)
12. Hoffmann, P.M.: Life's Ratchet. How Molecular Machines Extract Order from Chaos. Basic Books. A member of the Perseus Books Group., New York, 1st edn. (2012)
13. Ichihashi, N., Matsuura, T., Kita, H., Sunami, T., Suzuki, H., Yomo, T.: Constructing partial models of cells. Cold Spring Harb Perspect Biol 2(6), a004945 (Jun 2010)
14. Leduc, P.R., Wong, M.S., Ferreira, P.M., Groff, R.E., Haslinger, K., Koonce, M.P., Lee, W.Y., Love, J.C., McCammon, J.A., Monteiro-Riviere, N.A., Rotello, V.M., Rubloff, G.W., Westervelt, R., Yoda, M.: Towards an in vivo biologically inspired nanofactory. Nat Nanotechnol 2, 3–7 (2007)
15. Lentini, R., Martn, N.Y., Forlin, M., Belmonte, L., Fontana, J., Cornella, M., Martini, L., Tamburini, S., Bentley, W.E., Jousson, O., Mansy, S.S.: Two-Way Chemical Communication between Artificial and Natural Cells. ACS Central Science 3(2), 117–123 (2017)
16. Lentini, R., Santero, S.P., Chizzolini, F., Cecchi, D., Fontana, J., Marchioretto, M., Del Bianco, C., Terrell, J.L., Spencer, A.C., Martini, L., Forlin, M., Assfalg, M., Dalla Serra, M., Bentley, W.E., Mansy, S.S.: Integrating artificial with natural cells to translate chemical messages that direct E. coli behaviour. Nat Commun 5, 4012 (2014)
17. de Lorenzo, V., Danchin, A.: Synthetic biology: discovering new worlds and new words. EMBO Rep. 9(9), 822–827 (Sep 2008)
18. Luisi, P.L., Ferri, F., Stano, P.: Approaches to semi-synthetic minimal cells: a review. Naturwissenschaften 93, 1–13 (2006)
19. Luisi, P.L., Walde, P., Oberholzer, T.: Lipid vesicles as possible intermediates in the origin of life. Current Opinion in Colloid & Interface Science 4(1), 33–39 (1999)
20. Luisi, P.L.: Autopoiesis: a review and a reappraisal. Naturwissenschaften 90(2), 49–59 (Feb 2003)
21. Luisi, P.L.: The Synthetic Approach in Biology: Epistemological Notes for Synthetic Biology. P. L. Luisi, C. Chiarabelli (eds.). In: Chemical Synthetic Biology, pp. 343–362. Wiley, Chichester NY (2011)
22. Luisi, P., Varela, F.: Self-replicating micelles - a chemical version of a minimal autopoietic system. Orig. Life Evol. Biosph. 19, 633–643 (1989)
23. Maturana, H.R., Varela, F.J.: Autopoiesis and Cognition: The Realization of the Living. D. Reidel Publishing Company, 1st edn. (1980)
24. Maturana, H.R., Varela, F.G.: De máquinas y seres vivos: Una teoría de la organizacíon Biológica. Editorial Universitaria, Santiago (1972)
25. Mavelli, F., Rampioni, G., Damiano, L., Messina, M., Leoni, L., Stano, P.: Molecular Communication Technology: General Considerations on the Use of Synthetic Cells and Some Hints from In Silico Modelling. In: Advances in Artificial Life and Evolutionary Computation. pp. 169–189. Springer, Cham (2014)
26. Morange, M.: A Critical Perspective on Synthetic Biology. Hyle 15(1), 21–30 (2009)
27. Nakano, T., Eckford, A.W., Haraguchi, T.: Molecular Communications. Cambridge University Press, Cambridge UK (2013)
28. Nakano, T., Moore, M., Enomoto, A., Suda, T.: Molecular Communication Technology as a Biological ICT. In: Sawai, H. (ed.) Biological Functions for Information and Communication Technologies, pp. 49–86. No. 320 in Studies in Computational Intelligence, Springer Berlin Heidelberg (Jan 2011)

29. Pfeifer, R., Scheier, C.: Understanding Intelligence. MIT Press, Cambridge MA (1999)
30. Rampioni, G., Mavelli, F., Damiano, L., DAngelo, F., Messina, M., Leoni, L., Stano, P.: A synthetic biology approach to bio-chem-ICT: first moves towards chemical communication between synthetic and natural cells. Natural Computing pp. 1–17 (2014)
31. Rosenblueth, A., Wiener, N., Bigelow, J.: Behavior, Purpose and Teleology. Philosophy of Science 10, 18–24 (1943)
32. Searle, J.: Minds, Brains, and Programmes. Behavioral and Brain Sciences 3, 417–424 (1980)
33. Stano, P.: Advances in Minimal Cell Models: a New Approach to Synthetic Biology and Origin of Life. In: Progress in Molecular and Environmental Bioengineering - From Analysis and Modeling to Technology Applications, pp. 23–44. A. Carpi (Ed.), InTech (2011)
34. Stano, P.: The birth of liposome-based synthetic biology: a brief account. In: Liposomes: Historical, Clinical and Molecular Perspectives, B. R. Pearson (Ed.), pp. 37–52. Nova Science Publishers, Inc. (2017)
35. Stano, P., Carrara, P., Kuruma, Y., de Souza, T.P., Luisi, P.L.: Compartmentalized reactions as a case of soft-matter biotechnology: synthesis of proteins and nucleic acids inside lipid vesicles. J. Mater. Chem. 21, 18887–18902 (2011)
36. Stano, P., Mavelli, F.: Protocells Models in Origin of Life and Synthetic Biology. Life 5(4), 1700–1702 (Dec 2015), http://www.mdpi.com/2075-1729/5/4/1700
37. Stano, P., Rampioni, G., Carrara, P., Damiano, L., Leoni, L., Luisi, P.L.: Semi-synthetic minimal cells as a tool for biochemical ICT. Bio Systems 109(1), 24–34 (2012)
38. Stano, P., Rampioni, G., Carrara, P., Damiano, L., Leoni, L., Luisi, P.L.: Semi-synthetic minimal cells as a tool for biochemical ICT. Biosystems 109, 24–34 (Jul 2012)
39. Varela, F.J., Thompson, E.T., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. The MIT Press, new edition edn. (Nov 1992)
40. Walde, P., Wick, R., Fresta, M., Mangone, A., Luisi, P.: Autopoietic self-reproduction of fatty-acid vesicles. J. Am. Chem. Soc. 116, 11649–11654 (1994)

Relevant fluxes in metabolic steady-states

Chiara Damiani[1,2,*], Riccardo Colombo[1,2], Diletta Paone[1,3], Giancarlo Mauri[1,2], and Dario Pescini[1,3]

[1] SYSBIO Centre of Systems Biology, Piazza della Scienza 2, 20126 Milano, Italy
[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
[3] Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy

∗ Corresponding author: chiara.damiani@unimib.it

In the race to characterize the phenotype of different cellular systems, quantification of the rate of turnover of molecules through a metabolic pathway (the metabolic flux) has captured the attention of a large part of the Systems Biology and Systems Medicine [1] community. Metabolic fluxes are a function of gene expression, translation, post-translational protein modifications, and protein-metabolite interactions. When a cell experiences changes in its environment, the metabolism needs to adjusted accordingly. For example, when the carbon source changes, the cells need to down-regulate certain pathways and to up-regulate other parts of the metabolic network [2]. Flux levels represent thus integrative information about phenotypic peculiarities.

C metabolic flux analysis (MFA) is an effective experimental technique for quantitative determination of metabolic fluxes, which involves the application of a labeled substrate to a biological system, measurement of label incorporation within metabolite pools with mass spectrometry methods, and computational estimation of intracellular fluxes that fit the observed data [3]. Computations rely on stoichiometric models of metabolism to elucidate the transfer of moieties containing isotopic tracers from one metabolite into another. The scope of this kind of models can span from the description of a single pathway up to an entire organism. They are based on a steady assumption to describe the rate of transformations of compounds into products (with the proper stoichiometry) for the considered set of reactions, while disregarding enzyme kinetics. Because 13C-labeled substrates are expensive and the overall technique is laborious, fluxomics data are hardly collected in a systematic fashion.

The identification of a smaller subset of fluxes which is supposed to account for the overall dynamics of a metabolic network would mightily improve our understanding of the metabolic changes that a cell undergo, and would more effectively and precisely drive wet-lab investigations. We refer to fluxes belonging to this subset as relevant fluxes.

We have recently proposed a strategy [4, 5] to populate ensembles of metabolic steady-states according to given metabolic properties. Steady states have been retrieved either from mechanism based simulations [4] or from constrain-based simulations [5]. In the former case, we use experimental concentrations from the literature or measured (whenever possible), whereas the kinetic constants are randomly sampled [4]. In the latter case, we sample the space of feasible solutions, by optimizing the network for many random objective functions [5]. The mechanism-based methodology is computationally more demanding as opposed to the constraint-based one, but it has the side benefit of taking into account information on metabolite concentrations.

With the above described procedures, we have been able to determine steady state metabolic fluxes that satisfy the definition of a given metabolic phenotype. It is worth to underline that this approach is independent from knowledge of kinetic constants, as well as from the definition of an objective function. Conversely, the method can be used to predict ensembles of rate constants or metabolic objectives that are in agreement with a given condition of interest, which is defined as a set of flux properties.

In this work we want to perform statistical analysis on the data generated with the above methodologies to assess if relevant fluxes exist that can discriminate whether a flux distribution pertains to one among several ensembles, and hence to a particular phenotype.

As a test case, we used *in silico* data obtained from kinetic simulations of the metabolic network *EColiCore2* of the model bacteria *E. coli* [6]. The network describes the central carbon metabolism of the bacteria and was automatically extracted from and it is consistent with a genome-wide model of the same organism [7].

We used 5 ensembles of fluxes corresponding to 5 different metabolic phenotypes [8]. The ensembles were obtained in [8] by first simulating different supply of carbon source and different oxygenation states (namely: aerobic growth on glucose; anaerobic growth on glucose; aerobic growth on acetate; aerobic growth on succinate; aerobic growth on glycerol) and then filtering out flux distributions that are not in agreement with the metabolism expected for each experimental condition. Each condition was simulated for 10000 random parametrization (i.e. set of random kinetic constants, one for each mass action reaction in the model, whose value has been uniformly sampled in the interval [0,100]) until reaching a steady state. Once steady state flux values were computed, experiments were filtered according to some key pathways that are known to be active in a given condition. The cardinality of the obtained 5 ensembles of selected flux distributions differ, but it is approximately comprised between 1500 and 2200.

In order to look for relevant fluxes we applied non parametric analysis to the above described data. We performed a Kolmogorov-Smirnov [9] test (a non-parametric hypothesis test procedure able to discriminate if two samples derive from the same distribution without investigating the actual shape of the distributions) for each possible pair of conditions and for each flux. The goal was to identify those fluxes that statistically differ for each pair of conditions (with significance 0.05) and that should thus be able to discriminate each of the 5 conditions.

We were satisfied with the results of the analysis as we found a subset of 38 fluxes (over 114) that can be regarded as relevant fluxes. Relevant fluxes are reported (red color) in Figure 1. From their mapping on the metabolic net-

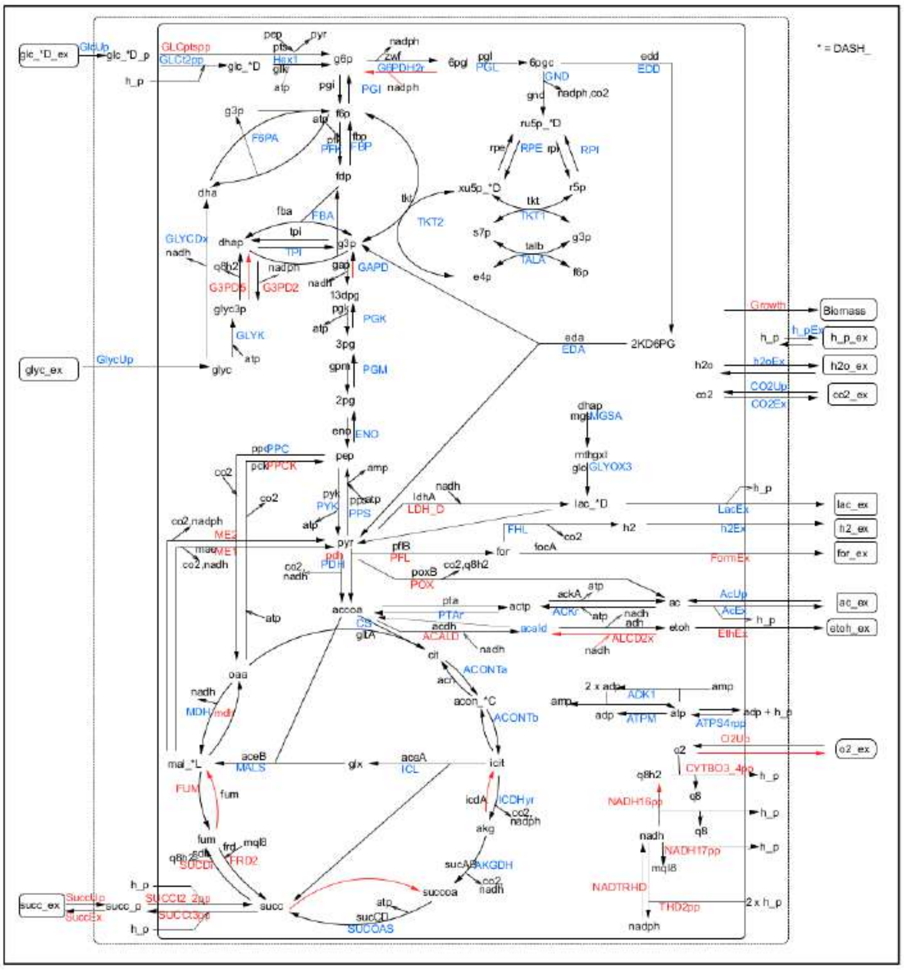


Fig. 1. Wiring diagram of th metabolic network *EcoliCore2*, the map has been modified from Hädicke et al. [6] adding cofactors and reverse reactions. In figure, next to edges representing reactions, reactions names are reported in blue. A dashed contour represents the external environment, while a solid line delimits the cell. Significant reactions emerging from the Kolmogorov-Smirnov test are labeled in red.

work it emerges that these reactions are mainly part of functional elements in the network: in particular exchange reactions (between cellular and external environment) and hubs (i.e. network junctions connecting different pathways). Instead, reactions internal to pathways are less represented among the significant ones. This may suggest that the cell performs a tight regulation of fluxes among different pathways and a less stringent tuning for reactions belonging to the same pathway.

The procedure can be extended to identify potential sets of kinetic constants that significantly differ across phenotypes. Some kinetic constants may indeed assume different values under various environmental conditions because of enzymatic regulation. The parameters identified as *relevant* would correspond to those reactions that must be up or down regulated to achieve given phenotypes, as opposed to reactions that are able to modulate their flux, thanks to variations in substrate or products concentrations, without enzymatic/genetic regulation.

References

1. A. Mardinoglu, J. Nielsen Systems medicine and metabolic modelling, *Journal of internal medicine*, 271(2):142–154, 2012.
2. J. Nielsen It is all about metabolic fluxes. *Journal of Bacteriology*, 185(24):7031–7035, 2003.
3. C. M. Metallo, J. L. Walther, G. Stephanopoulos Evaluation of 13 C isotopic tracers for metabolic flux analysis in mammalian cells. *Journal of biotechnology*, 144(3):167–174, 2009.
4. R. Colombo, C. Damiani, G. Mauri, D. Pescini. Constraining mechanism based simulations to identify ensembles of parametrizations to characterize metabolic features. To appear in *Lecture Notes in Bioinformatics (LNBI)*, 2017
5. C. Damiani, D. Pescini, R. Colombo, S. Molinari, L. Alberghina, M. Vanoni, G. Mauri. An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes. *Natural Computing*, 13(3):321–331, 2014.
6. O. Hädicke and S. Klamt. EColiCore2: a reference network model of the central metabolism of *Escherichia coli* and relationships to its genome-scale parent model. *Scientific Reports*, 7:39647, 2017.
7. J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, B. Ø. Palsson. A comprehensive genome-scale reconstruction of Escherichia coli metabolism2011, *Molecular systems biology*, 7(1):535, 2011.
8. R. Colombo, C. Damiani, D. Gilbert, M. Heiner, G. Mauri, D. Pescini. Emerging ensembles of kinetic parameters to identify experimentally observed phenotypes. Submitted to *BMC Bioinformatics*, 2017.
9. T. W. MacFarland and J. M. Yates. Introduction to nonparametric statistics for the biological sciences using R. Springer, 2016.