

# **Predicting species emergence in simulated complex pre-biotic networks: Supporting information**

Omer Markovitch and Natalio Krasnogor\*

omermar@gmail.com ; natalio.krasnogor@newcastle.ac.uk

## **1. MATLAB code**

GARD simulations were ran using GARD10 MATLAB code [Markovitch, O. and Lancet, D. Excess Mutual Catalysis Is Required for Effective Evolvability. *Artificial Life* 18, 3 (2012) (DOI: 10.1162/artl\_a\_00064)].

The following code is provided as attachment to the present manuscript:

Generating a lognormal  $\beta$  and performing community detection:

- Beta\_louvain.m; receives a random-seed and generates a lognormal  $\beta$  using GARD10, then Louvain's MATLAB function *cluster\_jl\_orient* is run. Output out.indices is an array holding for each molecule the index of the community it belongs to.
  - Beta\_infomap.m: receives a random-seed and generates a lognormal  $\beta$  using GARD10, then saves the  $\beta$  to disk and run *Infomap* as an externally program (UNIX or Windows). Output out.indices is an array holding for each molecule the index of the community it belongs to.
  - Beta\_oslom.m: similar to Beta\_infomap.m, running program *oslom\_dir*. Ouput out.indices is an array of *cells*, holding for each molecule a list of indices of communities it belongs to.
- getmaxrealevec.m receives a  $\beta$  matrix (or  $\beta^*$ ) and return as its first output the eigenvector with the highest real eigenvalue.

Are all available at <http://ico2s.org/data/extras/gard/> and [Markovitch, O. and Krasnogor, N. Accompanying dataset for: Predicting Species Emergence in Simulated Complex Pre-Biotic Networks. Zenodo (2016) (DOI: 10.5281/zenodo.56534)]

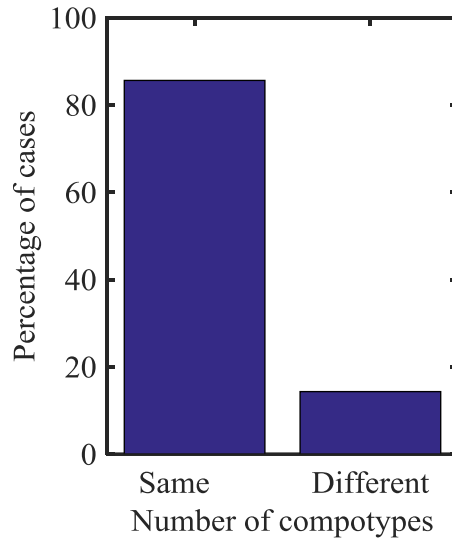
## **2. Accompanying dataset**

Full dataset with the entire 10,000 networks and simulations analysed is available freely at <http://ico2s.org/data/extras/gard/> and [Markovitch, O. and Krasnogor, N. Accompanying dataset for: Predicting Species Emergence in Simulated Complex Pre-Biotic Networks. Zenodo (2016) (DOI: 10.5281/zenodo.56534)]. The dataset includes:

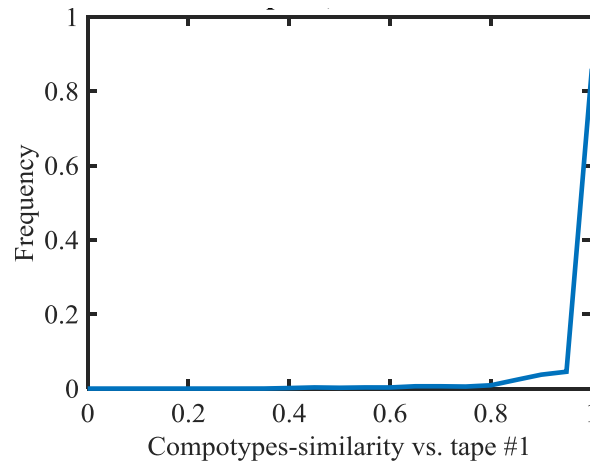
- 2.1. Images such as Fig 3 in the main text, for 100 cases studied for GARD tapes.
- 2.2. For each of the 10,000  $\beta$ 's and simulations used in the rest of the paper a text file with the following details:
  - 2.2.1. Seed number that was used to generate the specific  $\beta$  network
  - 2.2.2. For each comptype observed in the GARD simulation under this  $\beta$  – a vector of its composition (elements normalized to have sum = 1.0).
- 2.3. For each of the 3 community detection algorithms (Louvain, Infomap and OSLOM):
  - 2.3.1. Indices of which communities were assigned to which comptype.
  - 2.3.2. For each detected community – a list of the indices stating for each node to which community it belongs to.
- 2.4. A Matrix with full values of  $\beta$ .

### 3. GARD tapes

For each  $\beta$ , 9 additional simulation-runs (tapes) were repeated, and in each tape the compotypes were identified as before (i.e. k-means clustering). Fig A shows a histogram of the number of compotype species observed under each  $\beta$  of the 100 studied for this part, vs. the first tape. The similarity (H, Eq 3 in the main text) between the compotypes identified in each simulation-run to the first simulation-run was also calculated (Fig B).

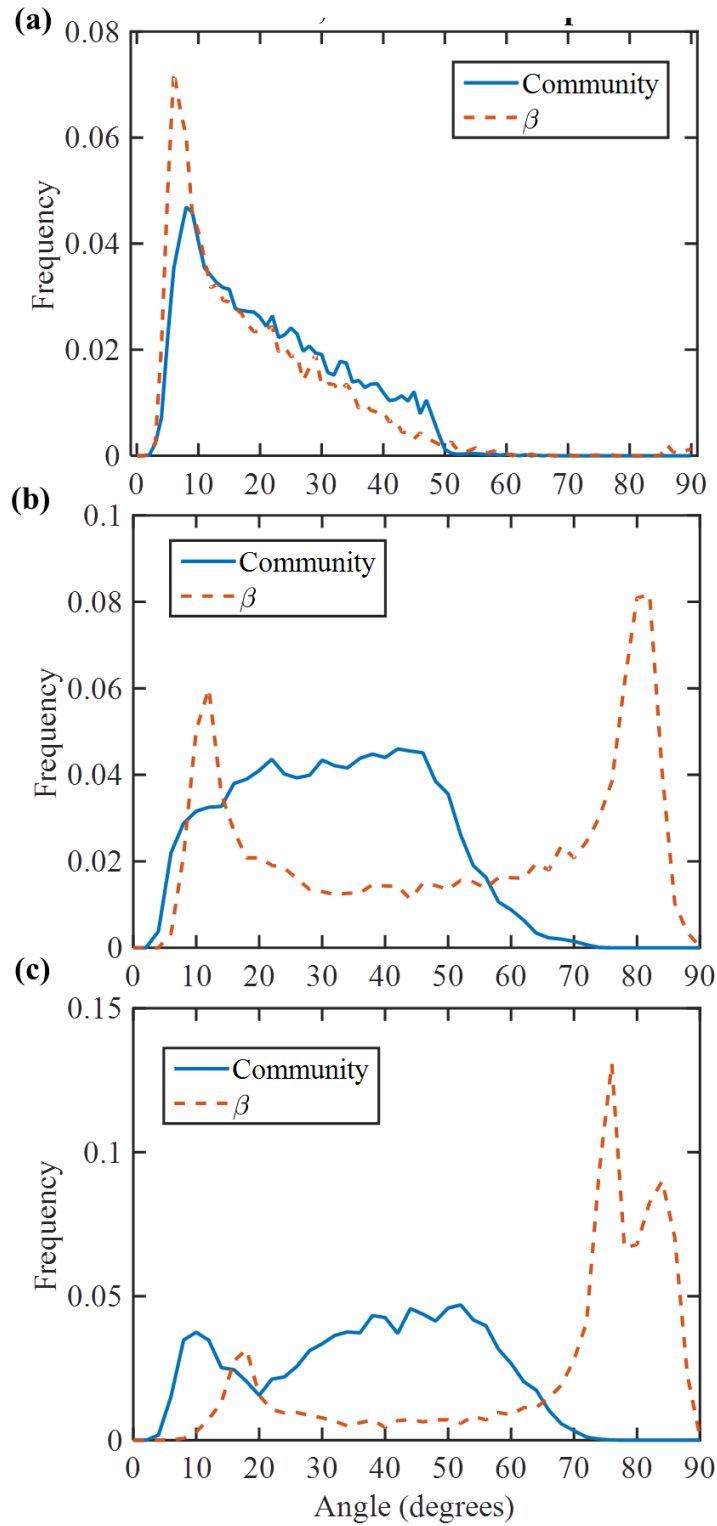


**Fig A. Compotype count with respect to the first tape.** Bar plot of the number of cases a tape showed same or different number of compotype species than the first tape, for the 100  $\beta$ 's studied for this part.  $\%(\text{same})=85.66$ .



**Fig B. Compotypes similarity (H, Eq 3 in the main text) with respect to the first tape.** Histogram of similarity between compotypes identified at different tapes to those identified in tape #1. Mean =  $0.976 \pm 0.0747$ .

#### 4. Histograms of angle to the eigenvector of the full- $\beta$ and $\beta^*$

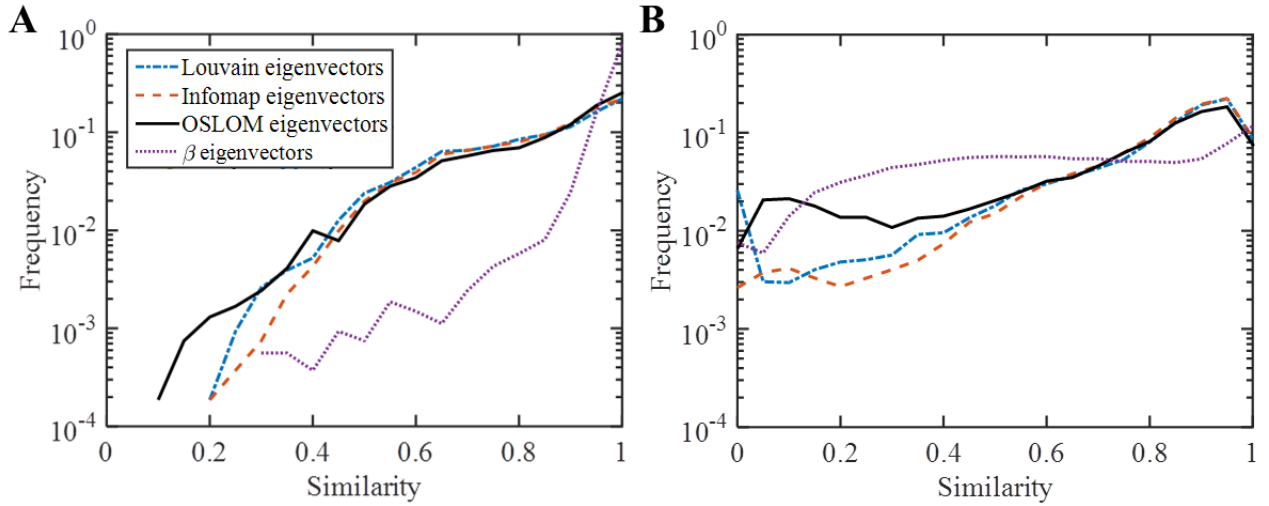


**Fig C. Angles between assemblies and eigenvectors.** Histograms of the angles between the assemblies and the eigenvector of  $\beta^*$  (Eq 2 in the main text) and the full  $\beta$ , for the three examples given in Fig 3 in the main text. Means and standard deviations are given in Table S1.

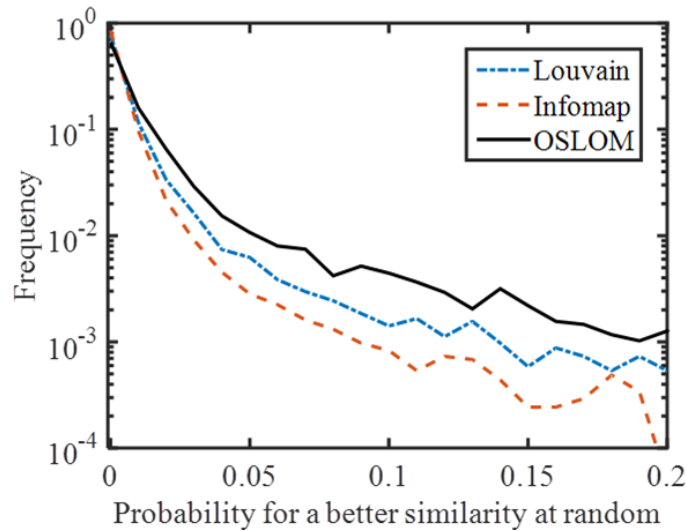
| Panel | Community   | $\beta$     |
|-------|-------------|-------------|
| (a)   | $21 \pm 12$ | $18 \pm 13$ |
| (b)   | $32 \pm 15$ | $51 \pm 27$ |
| (c)   | $37 \pm 17$ | $66 \pm 22$ |

**Table S1.** Mean and standard deviations of angle distributions in Fig C.

## 5. Comptype-community assignment



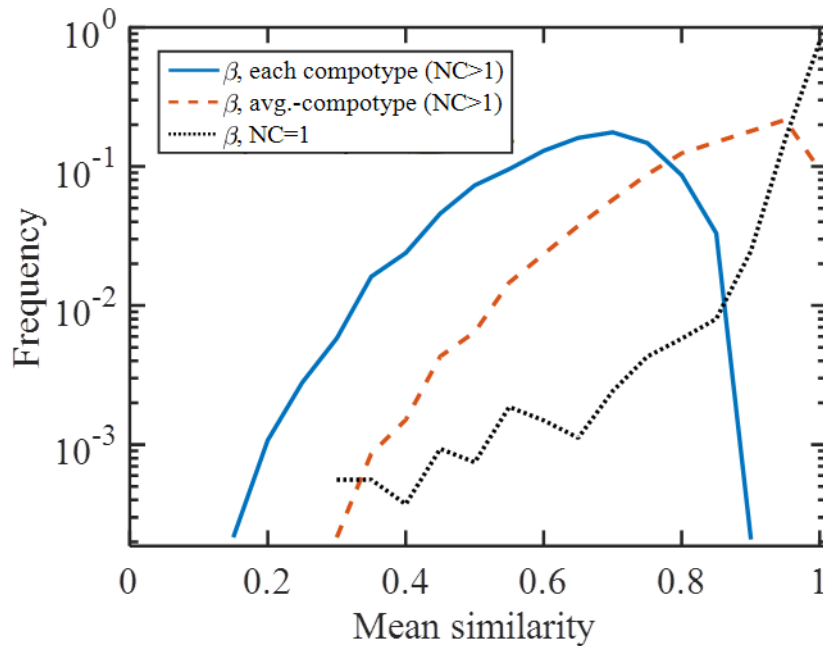
**Fig D. Histograms of compositional similarity when predicting comptypes using the eigenvectors of  $\beta^*$  and full- $\beta$ .** (A) Comparing all the cases when a single comptype was observed ( $NC=1$ ). (B) Comparing all the comptypes from all the cases where two or more comptypes were observed ( $NC>1$ ). Mean and standard deviation are given in Table 1 in the main text.



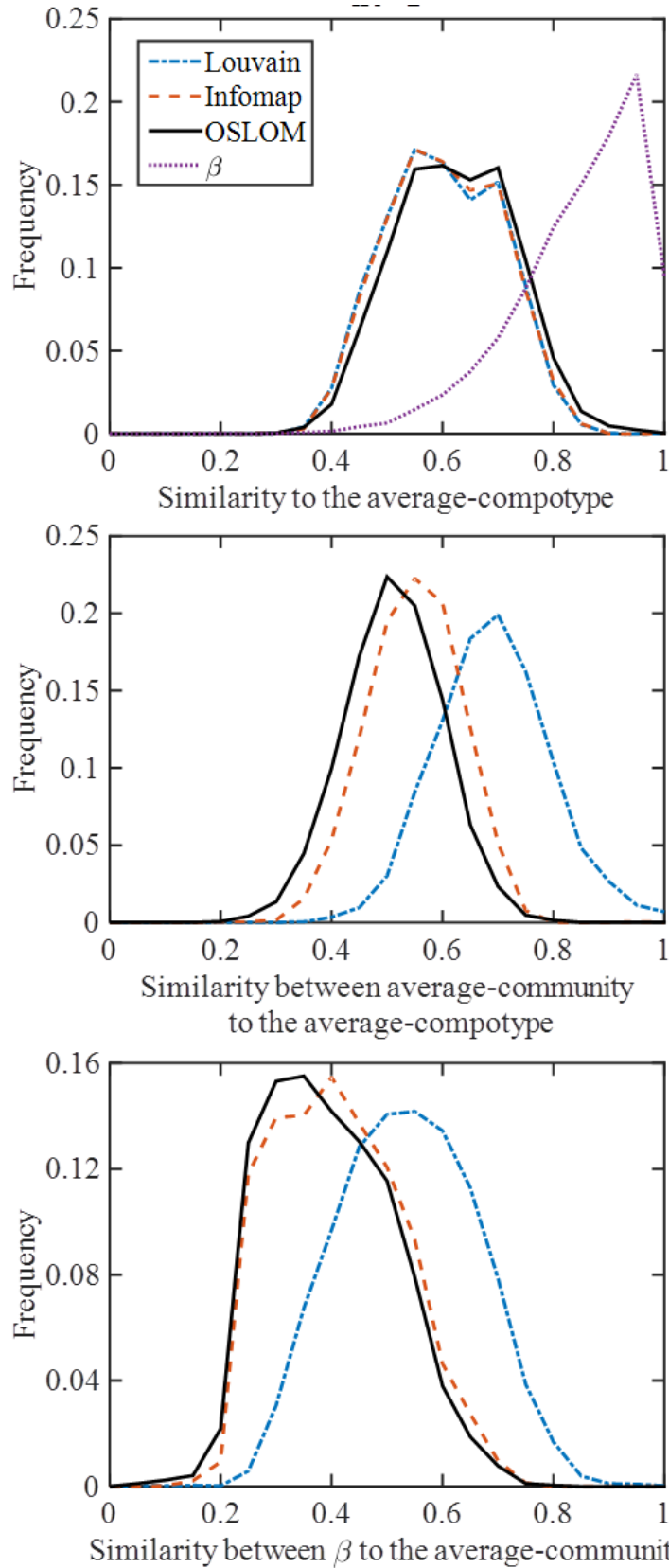
**Fig E. Testing for random assignment of comptype-community.** Histogram of the probability of achieving a higher similarity by a random community (figure truncated at  $p=0.2$ ). Mean and standard deviations are given in Table 1 in the main text.

## 6. On the eigenvector of the full- $\beta$

In the limiting case, where only a single compotype was observed, using Perron-Frobenius theorem directly on the full- $\beta$  resulted in an eigenvector which was highly similar to that compotype (Fig F). When multiple compotypes were observed in a simulation under a given  $\beta$ , the average similarity between eigenvector of the full- $\beta$  to each of those compotypes was low, yet the similarity to the average-compotype was quite high (Fig F). The average-compotype is a vector whose elements are the component-wise average of the compotypes that emerged under a given  $\beta$ . It follows that the eigenvector of the full- $\beta$  is the center of mass of the entire assemblies in the simulation, and it is suggested that the eigenvectors of the different  $\beta^*$  represent potential attractors in the compositional space (i.e. the space of all possible compositions of  $N_{\max}$  molecules out of an alphabet of  $N_G$  molecular types). As a control, the similarity between the eigenvectors of  $\beta^*$  and the average- $\beta^*$ -eigenvectors to the average-compotype were calculated, as well as between the average- $\beta^*$ -eigenvectors to the eigenvector of the full- $\beta$  and all found to be substantially low (Fig G). The average- $\beta^*$ -eigenvectors is a vector whose elements are the component-wise average of all the eigenvectors of all the communities detected under a given  $\beta$ .



**Fig F. Histograms of similarity with respect to the eigenvector of the full- $\beta$ .** The average of similarity to each of the compotypes when multiple compotypes observed (blue solid line), the similarity to the average-compotype when multiple compotypes observed (red broken line) and when only a single compotype is observed (black dotted line). Mean values respectively are:  $0.642 \pm 0.119$ ,  $0.848 \pm 0.116$  and  $0.975 \pm 0.054$ .



**Fig G. Similarities with respect to the average-compotype.** (top) Similarity between the eigenvector of  $\beta^*$  (the one that gives the highest similarity is picked) to the average-compotype. For reference the similarity with the eigenvector of the full- $\beta$  is also shown (Fig F). (middle) Similarity between the average- $\beta^*$ -eigenvector to the average-compotype. (bottom) Similarity between the eigenvector of the full- $\beta$  to the average- $\beta^*$ -eigenvector.