Chaos, Solitons & Fractals 72 (2015) 99–106

Contents lists available at ScienceDirect

Chaos, Solitons & Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Dynamical community structure of populations evolving on genotype networks

José A. Capitán [1], Jacobo Aguirre [1], Susanna Manrubia *

Centro Nacional de Biotecnología (CSIC), c/Darwin 3, 28049 Madrid, Spain

ARTICLE INFO

Article history:
Available online 26 December 2014

ABSTRACT

Neutral evolutionary dynamics of replicators occurs on large and heterogeneous networks of genotypes. These networks, formed by all genotypes that yield the same phenotype, have a complex architecture that conditions the molecular composition of populations and their movements on genome spaces. Here we consider as an example the case of populations evolving on RNA secondary structure neutral networks and study the community structure of the network revealed through dynamical properties of the population at equilibrium and during adaptive transients. We unveil a rich hierarchical community structure that, eventually, can be traced back to the non-trivial relationship between RNA secondary structure and sequence composition. We demonstrate that usual measures of modularity that only take into account the static, topological structure of networks, cannot identify the community structure disclosed by population dynamics.

1. Introduction

The biological evolution of populations is conditioned by the availability and attainability of genomic solutions leading to viable organisms. All biological beings are first defined by their genotypes, a sequence of variable length encoding the information of a developing program that eventually ushers in functional organisms. At the highest level, these organisms are characterized by their phenotypes, that is the set of measurable features that determine their biological functions — and their viability, when phenotype is evaluated in a particular environment —, and on which natural selection acts. Any individual is subjected to replication errors due to a non-zero mutation rate: Variability is thus an intrinsic property that produces heterogeneous populations and becomes essential for adaptation and for the discovery of evolutionary innovations.

Not all mutations are of equal value [1]. Some are beneficial, others are neutral, and many are deleterious for an organism when compared to its progenitors. In populations well adapted to constant environments, most mutations are deleterious. Still, a certain decrease in fitness is tolerated (typically depending on the population size) and this permits the appearance of compensatory mutations that guarantee survivability. When populations are not optimized (a frequent situation when environments change, for example), the fraction of beneficial mutations increases [2]. Finally, it is known that many mutations are neutral, such that they can accumulate in genomes and natural selection does not act on these variants. The idea of neutral evolution was first introduced by Kimura [3] in order to account for the known fact that a large number of mutations observed in proteins, DNA, or RNA, did not have any effect on fitness. Soon after, the relevance of neutral evolution to navigate at zero cost the space of genotypes was put forward [4].

To date, all available data and models analysed indicate that there is an enormous redundancy between genotype and phenotype. That is, many different genotypes produce the same phenotype, revealing the existence of a huge number of neutral mutations [5]. In addition, the space of

* Corresponding author.
[1] These two authors contributed equally.

http://dx.doi.org/10.1016/j.chaos.2014.11.019
0960-0779/

genotypes has a very high dimensionality, a condition that favors the existence of contiguous neutral genotypes. A sequence of length l whose components are taken from an alphabet of four letters (as it happens for DNA and RNA), has $3l$ different genotypes as neighbors that differ from it in only one nucleotide. If any of these neighbors yields the same phenotype, the two sequences can be connected through a point mutation. This permits that the genomic composition of a population be changed from one position (in the space of genomes) to that adjacent one without paying any cost in fitness. Actually, the likelihood that this local move can be repeated and leads to very long excursions in the space of genotypes increases with the sequence length. The previous facts have led to the concept of *neutral networks of genotypes*, representing connected ensembles of genotypes following the criterion of accessibility through point mutations. Genotype networks have important implications in the evolutionary process [6,12].

RNA sequences folding into their minimum free energy secondary structures (see Fig. 1) are a widely used model to represent the genotype-phenotype relationship [7–9]. RNA nucleotides A (adenine), U (uracil), G (guanine), and C (cytosine) form pairs that decrease the free energy of the open chain. The most energetic pair is G-C, followed by A-U and finally by G-U. Their energetic contribution is approximately -3 kcal/mol, -2 kcal/mol and -1 kcal/mol, respectively. The two first pairs are analogous to Watson–Crick pairs G-C and A-T in DNA, and the latter is specific of RNA. Analytical studies of the number of sequences of length l compatible with a fixed secondary structure (used as a proxy for the phenotype) have revealed that the average size of the corresponding neutral network grows as $l^{3/2}b^l$, where b is a constant [10]. For example, there should be about 10^{28} sequences compatible with the structure of a transfer RNA (which has length $l = 76$), while the currently known smallest functional RNAs, of length $l \approx 14$ [11], could in principle be obtained from more than 10^6 different sequences. As anticipated, neutral networks are astronomically large even for moderate values of the sequence length. Together with the high dimensionality of the space of phenotypes, that causes most (common) genotype networks to percolate the space of genotypes.

In this contribution, we analyse the dynamics of populations on realistic genotype networks using RNA secondary structure neutral networks as example. First, we need to rephrase some previous results regarding dynamics on heterogeneous networks [12] in the current molecular context, paying special attention to the consequences of heterogeneity for the diversity and composition of populations. We present new results regarding the community structure of genotype networks under realistic population dynamics, and introduce dynamical measures of modularity that reveal a complex interrelationship between (dynamical) community size, sequence representation in evolving populations, and RNA sequence composition.

2. Dynamics on genotype networks

A genotype network includes in principle all RNA sequences that fold into the same secondary structure. This ensemble does not necessarily form a single connected network. In what follows, we always reduce the dynamics of the populations to connected components of the phenotype. The topology of such a (connected) genotype network is specified through its corresponding adjacency matrix **C**. The elements C_{ij} take value 1 if genotypes i and j differ in a single letter of their sequences and value 0 if they differ in two or more letters. Sequences in a population replicate (see below) and daughter sequences have a probability to mutate one position in their sequences with a probability μ that is a parameter of the model. These dynamics have been studied in previous works [12–15], where the reader can find additional details.

2.1. Definitions and dynamical equations

Each genotype i in the network is represented by $n_i(t)$ sequences at time t, $i = 1, \ldots, m$, with m the number of

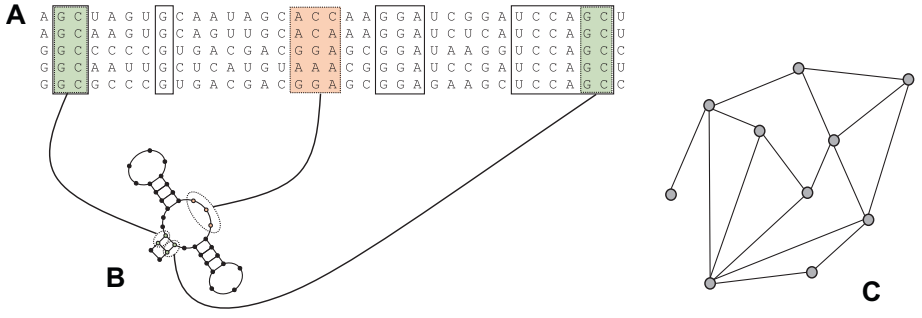

Fig. 1. RNA secondary structure neutral networks. (A) A few examples of the huge number of RNA sequences that fold into the same minimum free energy configuration (B). In order to preserve the structure, there are positions in the sequence that must be conserved (in this case the pairs G-C and C-G signaled by light green boxes), while others are almost free to change (unpaired nucleotides in the light orange box). Other positions might be conserved only in some subsets of sequences, and might correspond to paired or unpaired nucleotides (white boxes). An alternative representation for the structure in (B) is in the form of points (unpaired nucleotides) and parentheses (pairs of nucleotides): ((((..((((....))))...((((..))))).)))) (C) Schematic representation of a genotype network. Nodes correspond to genotypes (RNA sequences in this example) and links join sequences that differ in only one letter and fold into the same secondary structure. There are public servers where several properties of RNA folding can be easily explored, as http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

different genotypes (or nodes) in the network. A fixed population size is used, $N = \sum_i n_i(t)$, and we work in the limit of infinitely large populations, $N \to \infty$, so that finite size effects are discarded and, in practice, we consider the fraction of population at each node. The initial distribution of sequences on the network at $t = 0$ is $\vec{n}(0)$. In particular, we will be interested in homogeneous initial populations, that is, in populations which, at time $t = 0$ are formed by N identical genotypes. The degree k_i of each genotype i specifies how many neutral neighbors it has.

At each time step, all sequences replicate synchronously. Daughter sequences mutate to one of the $3l$ nearest neighbors with probability μ, and remain equal to their mother sequence with probability $1 - \mu$. In our representation $0 < \mu \leqslant 1$. The singular case $\mu = 0$ is excluded to avoid trivial dynamics and guarantee evolution towards a unique equilibrium state. With probability $k_i/(3l)$, the mutated sequence exists in the neutral network and it adds to the population of the corresponding neighboring genotype. Otherwise, it falls off the network and disappears. These conditions implicitly represent a peak landscape in phenotypes, that is a phenotype with value 1 and any other possibility with value 0 [6].

The mean-field equations describing the dynamics of the population on the network in matrix form are

$$\vec{n}(t+1) = (2-\mu)\mathbf{I}\vec{n}(t) + \frac{\mu}{3l}\mathbf{C}\vec{n}(t), \tag{1}$$

where $\mathbf{I}$ is the identity matrix. The transition matrix $\mathbf{M}$ is defined as

$$\mathbf{M} = (2-\mu)\mathbf{I} + \frac{\mu}{3l}\mathbf{C}. \tag{2}$$

The set of m eigenvalues (all real) of $\mathbf{M}$ is $\{\lambda_i\}$, and they are ordered such that $\lambda_i \geqslant \lambda_{i+1}$. The corresponding m eigenvectors are $\{\vec{u}_i\}$, and since $\mathbf{M}$ is real and symmetric they can be chosen such that $\vec{u}_i \cdot \vec{u}_j = 0$, $\forall i \neq j$ and $|\vec{u}_i| = 1$, $\forall i$. Matrix $\mathbf{M}$ is irreducible by definition (the underlying network is connected) and has positive values in the diagonal. It is therefore primitive, so the Perron–Frobenius theorem guarantees that the largest eigenvalue of $\mathbf{M}$ is positive, $\lambda_1 > |\lambda_i|$, $\forall i > 1$, and its associated eigenvector is also positive (i.e., $(\vec{u}_1)_i > 0$, $\forall i$) in the interval of μ values used [12].

The dynamics of the system, Eq. (1), can thus be written as

$$\vec{n}(t) = \mathbf{M}^t\vec{n}(0) = \sum_{i=1}^{m} \lambda_i^t \alpha_i \vec{u}_i, \tag{3}$$

where we have defined α_i as the projection of the initial condition on the ith eigenvector of $\mathbf{M}$,

$$\alpha_i = \vec{n}(0) \cdot \vec{u}_i. \tag{4}$$

Furthermore, as $\lambda_1 > |\lambda_i|$, $\forall i > 1$, there exists a unique asymptotic state of the population that is independent of the initial condition $\vec{n}(0)$ and is proportional to the eigenvector that corresponds to the largest eigenvalue, $\vec{u}_1$:

$$\lim_{t\to\infty}\left(\frac{\vec{n}(t)}{\lambda_1^t\alpha_1}\right) = \vec{u}_1, \tag{5}$$

while the largest eigenvalue λ_1 yields the growth rate of the population at equilibrium (in the absence of rescaling). For convenience, in the following, and without any loss of generality, we normalize the population $\vec{n}(t)$ such that $|\vec{n}(t)| = 1$ after each generation. With this normalization, $\vec{n}(t) \to \vec{u}_1$ when $t \to \infty$.

It is easy to demonstrate that the eigenvalues λ_i of the transition matrix $\mathbf{M}$ are related to the eigenvalues γ_i of the adjacency matrix $\mathbf{C}$ through $\lambda_i = (2-\mu) + \frac{\mu}{3l}\gamma_i$. Furthermore, the eigenvectors of both matrices are identical [12]. This result implies that the asymptotic state of the population only depends on the topology of the genotype network.

2.2. Time to equilibrium

Eq. (3) describes the dynamics towards equilibrium from an initial condition $\vec{n}(0)$. The distance $\Delta(t)$ to the equilibrium state can be written as

$$\Delta(t) \equiv \left|\frac{\mathbf{M}^t\vec{n}(0)}{\lambda_1^t\alpha_1} - \vec{u}_1\right| = \left|\sum_{i=2}^{m}\vec{\Delta}_i(t)\right| = \left|\sum_{i=2}^{m}\frac{\alpha_i}{\alpha_1}\left(\frac{\lambda_i}{\lambda_1}\right)^t\vec{u}_i\right|. \tag{6}$$

In order to estimate how many generations elapse before equilibrium is reached, we fix a threshold ϵ, and define the *time to equilibrium* t_ϵ as the number of generations required for $\Delta(t_\epsilon) < \epsilon$.

When $\alpha_2 \neq 0$, $\lambda_2 \neq 0$ and $\lambda_2 \neq \lambda_3$, t_ϵ can be approximated to first order by

$$t_\epsilon^1 \simeq \frac{\ln|\alpha_2/\alpha_1| - \ln\epsilon}{\ln|\lambda_1/\lambda_2|}. \tag{7}$$

This approximation turns out to be extremely good in most cases thanks to the exponentially fast suppression of the contributions due to higher-order terms (since $\lambda_i \geqslant \lambda_{i+1}$, $\forall i$). An evaluation of situations where approximation (7) fails can be found in [12]. However, all cases shown in this work are well approximated by (7) within an error of one generation with respect to the exact time to equilibrium implicitly defined through expression (6).

An explicit relationship between the time to equilibrium and the mutation rate can be obtained by expanding Eq. (7) in powers of μ,

$$t_\epsilon^1 = \ln\left(\left|\frac{\alpha_2}{\epsilon\alpha_1}\right|\right)\left[\frac{a}{\mu} + b - c\mu\right] + O(\mu^2), \tag{8}$$

where

$$a = \frac{6l}{(\gamma_1-\gamma_2)}, \quad b = \frac{\gamma_1+\gamma_2-6l}{2(\gamma_1-\gamma_2)}, \quad c = \frac{\gamma_1-\gamma_2}{72l}. \tag{9}$$

Since $c \ll a$, the dependence of the time to equilibrium with the mutation rate follows $t_\epsilon^1 \propto \mu^{-1}$ [12]. For a fixed topology of the genotype network, the mutation rate μ sets the rate at which equilibrium is approached. An additional factor, that we will explore in the following, is the effect of the initial condition on t_ϵ^1, which is implicit in the projections α_1 and α_2, as previously defined.

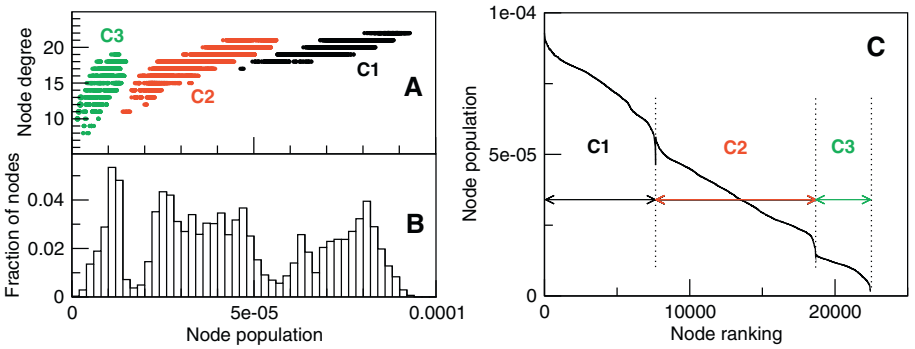

Fig. 2. Equilibrium composition of a neutrally evolving population of RNA sequences: dynamical community indicators. (A) Node degree versus node population for all 22,434 sequences in the network associated to phenotype ..((....))..; three clusters corresponding to three communities identified through the dynamics are observed. (B) Histogram of the population of nodes. (C) Rank plot of the node population. The population in node (or genotype) i is obtained as the ith element of the first eigenvector of the transition matrix **M**.

3. Community division in RNA secondary structure neutral networks

In a previous study [16], all RNA sequences of length $l = 12$ where exhaustively folded *in silico*. The minimum free energy secondary structure of each sequence was predicted through the routine `fold()` from the Program RNAfold included in the Vienna RNA package [17], version 1.5, with energy parameters based on [18]. Subsequently, genotype networks for all possible secondary structures of that length were calculated. That study was directed towards describing the topological properties of RNA genotype networks, but did not address the analysis of how that topology affects the dynamics of populations. In this work we will analyse the hierarchical structure of a large connected network with $m = 22,434$ sequences corresponding to the phenotype ..((....))..[2] with the aim of characterizing the relationship between topology, sequence composition, and population dynamics.

3.1. Equilibrium properties

In all the results that will be presented, we work in the limit of infinite population size and thus obtain the dynamical properties of the system by numerically solving the corresponding equations. To evaluate equilibrium properties, as discussed, it is enough to consider the adjacency matrix **C**. The first eigenvector of the transition matrix **M** yields the population of each node at equilibrium, according to (5). Two important properties of genotypes are their degree (k_i, number of neutral neighbors) and the fraction of population they accumulate at equilibrium, $(\vec{u}_1)_i$. We will refer to them as node degree and node population throughout the paper. When these two quantities are represented as a function of each other, we observe the appearance of three disjoint clusters of nodes, Fig. 2(A). From now on, these clusters will be communities 1, 2, and 3 (C1, C2, and C3 for short), ranked in decreasing order with respect to the maximally populated genotype (or node). The separation in three communities is also observed when we represent a histogram of the number of nodes as a function of their population, Fig. 2(B). We obtain a distribution with three well-defined regions of abundance, in agreement with the clusters in Fig. 2(A). Finally, a third indicator of the division in communities appears in a rank-ordering representation of nodes according to their population, Fig. 2(C). There are visible jumps in this curve that coincide with changes of community.

3.2. Time to equilibrium

The previous analysis of equilibrium properties reveals that, regarding dynamical properties, genotype networks seem to organize in communities. In order to further investigate this possibility and the hierarchical organization of the network, we continue our analysis with the study of non-equilibrium properties, and their dependence on quantities characterizing the nodes of the network. A first question is whether the initial condition affects in a significant and meaningful way the time to equilibrium. From a biological viewpoint, this quantity is of interest as well if we consider how an evolving population might find and fix a new phenotype. Previous phenotypes correspond to the exterior of the network we are investigating, and the new phenotype is associated to the current network. The probability to enter the network through a particular node depends on the number of outgoing links it has, that is its outward degree $3l - k_i$. Also, the larger the network the more likely that the phenotype it represents is localized. Nodes that are more connected in the network have lower probability of being the first node visited by an external population. This fact has implications that have been discussed elsewhere [19]. In addition, it is also common that populations enter a new phenotype through a single node, affected by a sort of genomic bottleneck caused by the difficulty to find and fix new phenotypes, and also usually by evolving at not-that-high mutation rates (specifically, in the limit $\mu N \ll 1$) [20].

The time to equilibrium varies approximately twofold depending on the initial distribution of the population. We have compared t^1_ϵ, Eq. (7), in several different

[2] Since we are here using up-to-date energy parameters for RNA folding and do not permit the existence of isolated pairs, the precise ensemble of genotypes that maps onto each secondary structure differs from those obtained in [16]. Statistical and topological properties remain invariant.

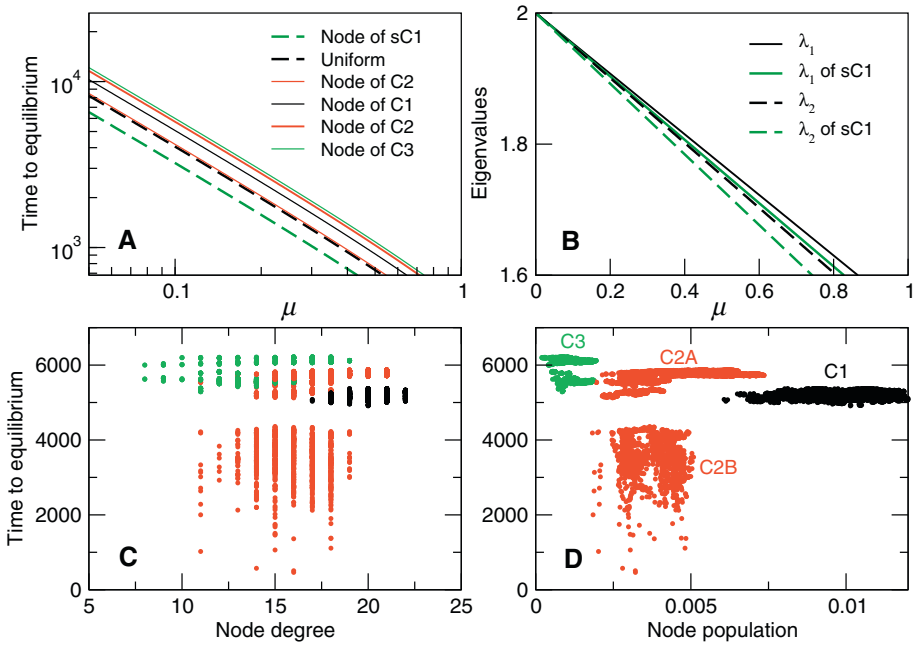

Fig. 3. Dependence of time to equilibrium on the initial condition. (A) Time to equilibrium as a function of the mutation rate μ when the evolutionary process takes place only in isolated community 1 (sC1, green dashed curve) or in the whole network (remaining curves) for a sample of different initial conditions. Time to equilibrium is shorter in the former case up to a factor 2. Different situations are detailed in the main text and in the legend. We observe a significant variation among nodes of C2, two examples are given. (B) Comparison between the first and second eigenvalues λ_1 and λ_2 of the whole network and the isolated C1 as a function of μ. (C) Time to equilibrium versus node degree when the population initially occupies a single node i, as a function of the degree k_i of that node. (D) Time to equilibrium versus equilibrium population $(\vec{u}_1)_i$. C2 is clearly split in two subcommunities. The coloring of each node in (C) and (D) follows the division presented in Fig. 2(A). Parameters are $\epsilon = 0.0001$ in (A-D) and $\mu = 0.1$ in (C) and (D). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

situations, in particular when all the initial population is concentrated in a single node of communities 1, 2, or 3, or when it spreads uniformly over the whole network. Further, we have also considered whether C1, which is the subset of the network accumulating most of the sequence population (see Fig. 2), could be a good representative of the dynamics we observe over the whole network. To this end, we have simulated the evolution towards equilibrium in a population spreading only on C1 (with 7460 nodes) and compared with the previous cases. All these results are summarized in Fig. 3(A). The slowest dynamics corresponds to initial conditions starting in C3, pointing out at its relative isolation within the network, while an initial condition uniformly distribution over the whole network leads to the fastest dynamics. The comparison between the subnetwork C1 isolated from the rest (sC1) and the whole network reveals that dynamics is faster in the former, though both are quite comparable. Differences arise from a variation in the two largest eigenvalues, as shown in Fig. 3(B).

Next, we have represented the time to equilibrium t_ϵ^1 as a function of the degree k_i of each node i, fixing the mutation rate $\mu = 0.1$ and repeating the calculation above for all nodes in the whole network. The relation between these two quantities offers little discriminatory power regarding community structure, since only two major communities can be resolved, see Fig. 3(C). On the contrary, a plot of the time to equilibrium as a function of the population at equilibrium (now relating two dynamical quantities) reveals a rich structure in communities and hints at a possible hierarchical organization, Fig. 3(D). C2 is now clearly separated into two independent communities (labeled C2A and C2B) and further subdivisions of C3 and C2A can be hypothesized.

Fig. 4 compares the overall properties of the four communities C1, C2A, C2B, and C3 clearly detected in Fig. 3(D). It is interesting that three of these quantities appear correlated (the fraction of the total population in each community, the average — per node — population, and the average degree), while two others are weakly dependent on each other (size of the community and the average time to equilibrium) and apparently uncorrelated to the former three.

3.3. Topological communities

Our results up to now indicate that populations evolving on genotype networks organize in communities that are revealed through dynamical quantities, notably the time to achieve equilibrium when a population enters the network from a single node and the final population that same node (or genotype) succeeds at attracting at equilibrium. Now we wish to compare these dynamical indicators of community structure with two other methods to detect communities based solely on network topology [21,22].

The modularity Q of a given network is quantified as the fraction of edges that fall within the groups specified by a

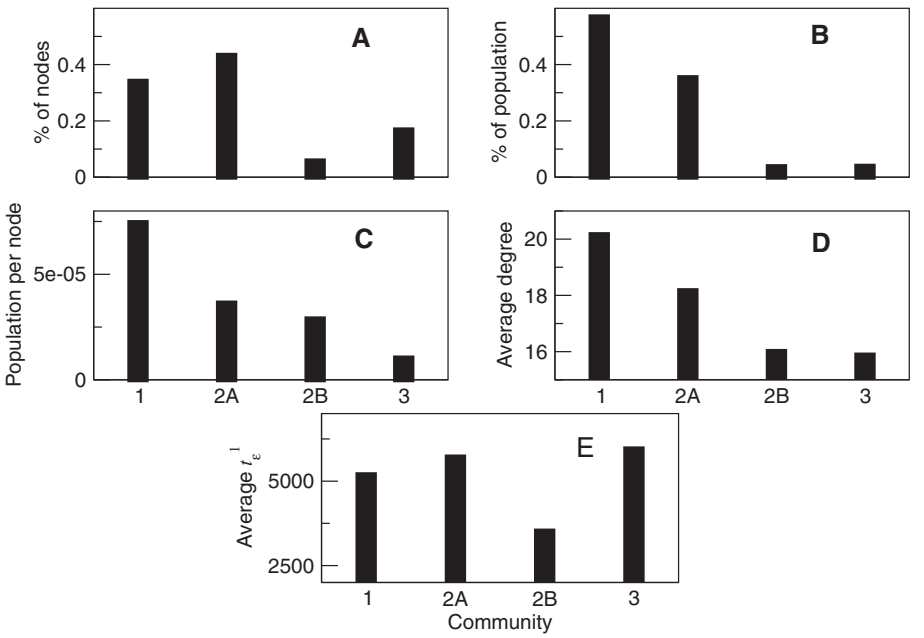

Fig. 4. Overall properties of the communities detected in Fig. 3(D). While C1 is formed by only 34% of the nodes (A), it attracts 57% of the population at equilibrium (B) and reaches the largest average population per node at the equilibrium (C). In fact, its largest average degree per node (which correlates with higher folding stability) makes C1 more robust under mutations and an attractor for the population (D). The time to equilibrium depends in a complex manner on the location of nodes within communities, as well as on the particular topology of the network (E). C3 behaves quite differently, since it is a relatively large community (A) but appears as isolated (E) and demonstrates little attracting power (B,C), due among others to its low average degree (D).

particular division in communities minus the expected fraction should those edges be distributed at random. Therefore, communities are defined as sets of nodes sharing more links among them than with nodes outside the community. Formally, modularity is defined as

$$Q = \sum_{v=1}^{c} (e_{vv} - a_v^2), \tag{10}$$

where e_{vv} is the fraction of edges with both end vertices in the same community v and a_i is the fraction of ends of edges that are attached to vertices in community v (with origin in nodes outside v). The sum runs over the different c communities in the partition tested. Optimal divisions into communities are obtained by maximizing the value of Q [23,24].

We have implemented two different methods to detect topological communities in our genotype network. First, we have used the stochastic block model inference method in [25].[3] It is based on a nested generative model that uses a hierarchical characterization of the entire network at different scales and allows to perform a correct statistical inference and a proper detection of its modular structure. This method permits to establish *a priori* the number of communities in the network, so we have fixed it to 4 in order to compare the result with the dynamical communities in Fig. 5(A). The obtained modularity is $Q = 0.66$, which is not optimal (since we fix the number of communities) but

[3] See also http://graph-tool.skewed.de/static/doc/community.html for an algorithm that implements the method.

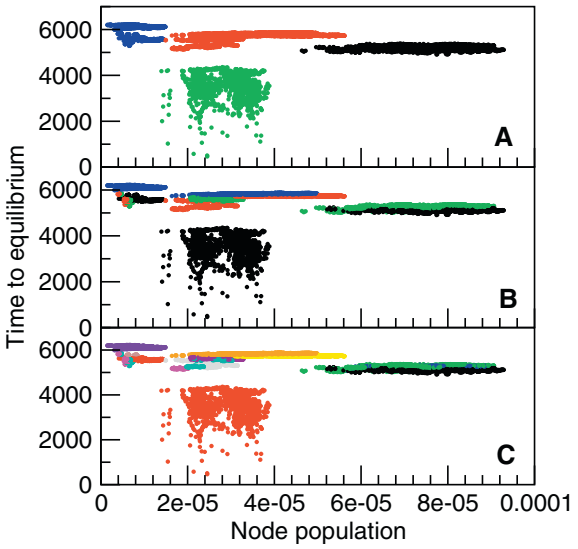

Fig. 5. Network communities obtained with different methodologies. The three panels repeat the values plotted in Fig. 3(D) and color the nodes according to the community they are assigned to with each method. (A) The functional dependence between the time to equilibrium and node population retrieves four communities (for comparison, this is a repetition of Fig. 2(D) with four colors). This partition yields a modularity value $Q = 0.46$. (B) Communities obtained with a stochastic block model inference method fixing the number of communities to 4. Here, $Q = 0.66$. (C) Community structure detected through a modularity optimization method. In this case, the number of communities is not fixed, so modularity reaches a higher value, $Q = 0.74$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

takes a value higher than in the dynamical case (with $Q = 0.46$). However, the stochastic block model method mixes nodes from all four communities, as shown by the distribution of different colors in Fig. 5(B). Second, we use a Potts model approach [26] which optimizes Q and freely selects the corresponding number of communities. In this case we obtain a division in 12 communities and an optimal modularity $Q = 0.74$. As with the previous topological method, we observe a high mixing of all 12 communities in Fig. 5(C) in comparison to the dynamical result shown in Fig. 5(A).

3.4. Dynamical communities and genotype composition

Methods to optimize topological communities yield divisions that differ substantially from those obtained through population dynamics. In order to disentangle the reasons that concentrate populations in certain regions (as C1) while others, though significantly large are comparatively depleted in population (as C3), we have investigated the existence of compositional differences among the four communities in Fig. 5(A). To this end, we need here to reconsider the system that we began with, RNA sequences, and analyse differences and similarities between the composition, in terms of nucleotides, of sequences in each community.

As we illustrated in Fig. 1, not all positions along the RNA sequence admit mutations with the same probability. In the phenotype of the network that we have analysed, corresponding to the secondary structure `..((....))..`, there are four positions that are invariant in all network genotypes. These are those forming pairs, which occupy positions 3, 4, 9, and 10, with nucleotides `C`, `C`, `G`, and `G`, respectively. In order for the phenotype to be maintained, there are some restrictions on the nucleotides occupying positions 2 and 11: they have to be occupied by nucleotides unable to pair; otherwise, a third pair would be stably formed, leading to a different phenotype. This condition excludes six combinations (`A-U`, `U-A`, `G-U`, `U-G`, `G-C`, and `C-G`) for positions 2 and 11. The remaining positions (1, 5, 6, 7, 8, and 12) can be occupied by any possible nucleotide, since they do not participate in structural changes — therefore the phenotype is maintained.

A compositional analysis of genotypes in the network reveals that it is precisely the limitations in the possible pairs at positions 2 and 11 that separates nodes into major communities. Fig. 6 illustrates the relationship between communities and sequence composition. Thanks to this analysis we have realized that three of the communities can be further divided into subcommunities attending to their composition, and that non-trivial dynamical relationships appear. We have calculated the Hamming distance, defined as the number of differences in the composition of two sequences, between all possible pairs in the network. It turns out that some communities are two mutational steps away, thus requiring an intermediate group of genotypes for a population to move from one group to another. The most remarkable example relates communities 2A and 2B, which can only be linked through nodes at communities 1 or 3. Further, if going through C3 the jump can only take place through sequences in community 3.3 going to 2A.2, since all other possibilities are again two mutations away, and thus an intermediate (sub) community is again required. Also, a similar decoupling occurs among the subcommunities of 2A: 2A.2 is two mutations away from 2A.1, but intermediate sequences reside in different communities. Therefore, if staying in 2A the only possible path has at least three mutations: 2A.2 → 2A.4 → 2A.3 → 2A.1.

The process of division in communities attending to sequence composition could be in principle iterated to unveil a complete hierarchical structure in genotype networks. However, not all divisions are equally meaningful. For instance, the fact that most unpaired nucleotides inside loops of sizes 3 and 4 can take any value without disturbing the secondary structure makes divisions at this level

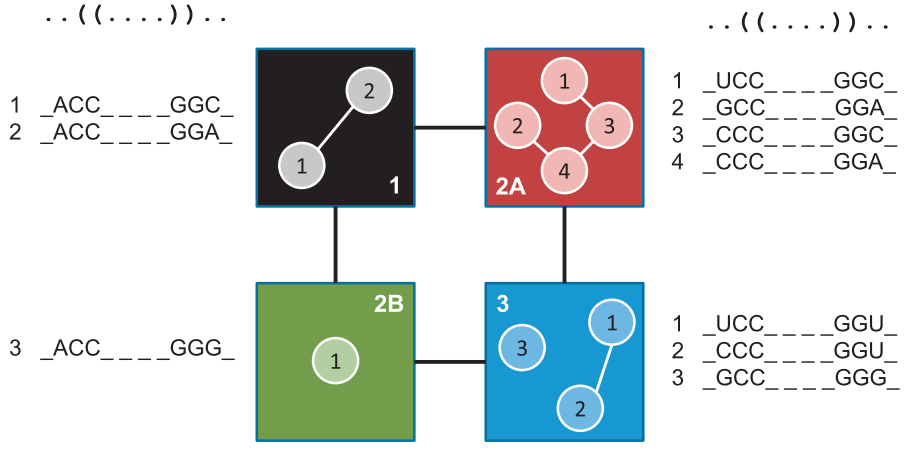

Fig. 6. Genotype composition and community structure. Sequences in each of the four dynamical communities belong to different compositional groups, as specified. Colors as in Fig. 5(A). Compositional differences and Hamming distances indicate that a further community structure can be identified, in agreement with the existence of additional divisions, first inferred from Fig. 3(D) and here indicated as numbered circles inside major communities. Links (black between major communities, white between subcommunities) join compositional groups that are at a Hamming distance of 1.

uninteresting regarding population dynamics. From what has been discussed in this section, compositional restrictions in sequence positions that affect RNA secondary structure generates communities that are dynamically detected by evolving populations: these communities are *de facto* separated by valleys, or bottlenecks, in the space of genomes, a situation that difficults the exchange of population among groups.

4. Conclusions

We have presented an analysis of the structure of genotype networks based on the dynamical properties of populations of replicators conditioned to evolve on that network. By means of an RNA secondary structure neutral network, it has been shown that sequence populations are able to detect a hierarchical community structure that reflects compositional differences among communities and the concomitant existence of restrictions to population exchanges. The dynamical communities detected cannot be recovered through existing measures of modularity, which analyse the topological (thus static) structure of networks.

The network we have chosen to illustrate the methodology is not special in any way. Actually, other secondary structures fulfilling stronger symmetry conditions probably lead to even clearer community patterns, as might be inferred from previous studies which have also shown clustering similar to that in Fig. 2(A) in other RNA networks [16]. The phenomenology here described is thus generic. Actually, longer sequences leading to larger networks will probably have a richer hierarchical structure, since, as we have seen, structural elements of RNA folded configurations play an essential role in separating communities.

We believe that the detection of communities through dynamical methods can be not only extended to other genotype networks, but very likely to any other system whose dynamics is constrained by the topology of a complex network. A variety of network community detection methods based on meaningful underlying dynamical processes have been proposed in different contexts [27–29], and all of them share the feature that detected communities can be very different from those based solely on network connectivity. The identification of relevant dynamical indicators of community structure (as were here time to equilibrium or population of nodes) may be system-dependent, and thus remains at present as an open problem worth pursuing.

Acknowledgements

The authors are indebted to Dr. Carlos Lugo for the generation of the RNA genotype network used in this work. This study has been supported by project FIS2011-27569 from the Spanish Ministry of Economy and Competitivity.

References

[1] Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet 2007;8:610–8.
[2] Stich M, Lázaro E, Manrubia SC. Phenotypic effect of mutations in evolving populations of RNA molecules. BMC Evol Biol 2010;10:46.
[3] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–6.
[4] Maynard Smith J. Natural selection and the concept of a protein space. Nature 1970;225:563–4.
[5] Wagner A. The origins of evolutionary innovations. Oxford University Press; 2011.
[6] Manrubia SC, Cuesta JA. Neutral networks of genotypes: evolution behind the curtain. ARBOR Ciencia Pensamiento y Cultura 2010;746:1051–64.
[7] Ancel LW, Fontana W. Plasticity, evolvability and modularity in RNA. J Exp Zool (Mol Dev Evol) 2000;288:242–83.
[8] Fontana W. Modelling 'evo-devo' with RNA. BioEssays 2002;24:1164–77.
[9] Schuster P. Prediction of RNA secondary structures: from theory to models and real molecules. Rep Prog Phys 2006;69:1419–77.
[10] Schuster P, Fontana W, Stadler PF, Hofacker IL. From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc London B 1994;255:279–84.
[11] Anderson PC, Mecozzi S. Unusually short RNA sequences: design of a 13-mer RNA that selectively binds and recognizes theophylline. J Am Chem Soc 2005;127:5290–1.
[12] Aguirre J, Buldú JM, Manrubia SC. Evolutionary dynamics on networks of selectively neutral genotypes: effects of topology and sequence stability. Phys Rev E 2009;80:066112.
[13] van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 1999;96:9716–20.
[14] Bornberg-Bauer E, Chan HS. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc Natl Acad Sci USA 1999;96:10689–94.
[15] Aguirre J, Papo D, Buldú JM. Successful strategies for competing networks. Nat Phys 2013;9:230–4.
[16] Aguirre J, Buldú JM, Stich M, Manrubia SC. Topological structure of the space of phenotypes: the case of RNA neutral networks. PLos ONE 2011;6:e26324.
[17] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, et al. Fast folding and comparison of RNA secondary structures. Monatsh Chem 1994;125:167–88.
[18] Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999;288:911–40.
[19] Manrubia S, Cuesta JA. Evolution on neutral networks accelerates the ticking rate of the molecular clock. J R Soc Interface 2015;12:20141010.
[20] Koelle K, Cobey S, Grenfell B, Pascual M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science 2006;314:1898–903.
[21] Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. J Stat Mech 2005:P09008.
[22] Porter MA, Onnela J-P, Mucha PJ. Communities in networks. Not Am Math Soc 2009;56:1082–97.
[23] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004;69:026113.
[24] Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA 2006;103:8577–696.
[25] Peixoto T. Hierarchical block structures and high-resolution model selection in large networks. Phys Rev X 2014;4:011047.
[26] Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E 2006;74:016110.
[27] Arenas A, Díaz-Guilera A, Pérez-Vicente C. Synchronization reveals topological scales in complex networks. Phys Rev Lett 2006;96:114102.
[28] Prada-Gracia D, Gómez-Gardeñes J, Echenique P, Falo F. Exploring the free energy landscape: from dynamics to networks and back. PLoS Comput Biol 2009;5:e1000415.
[29] Gómez-Gardeñes J, Zamora-López G, Moreno Y, Arenas A. From modular to centralized organization of synchronization in functional areas of the cat cerebral cortex. PLoS ONE 2010;5:e0012313.