GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES

CHRISTIAN REIDYS*, †, PETER F. STADLER*, ‡ and PETER SCHUSTER*, †, §, ¶
*Santa Fe Institute, Santa Fe, NM 87501, U.S.A.
†Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.
‡Institut für Theoretische Chemie der Universität Wien, A-1090 Wien, Austria
§Institut für Molekulare Biotechnologie, D-07708 Jena, Germany
(E.mail: pks@tbi.univie.ac.at)

Random graph theory is used to model and analyse the relationships between sequences and secondary structures of RNA molecules, which are understood as mappings from sequence space into shape space. These maps are non-invertible since there are always many orders of magnitude more sequences than structures. Sequences folding into identical structures form neutral networks. A neutral network is embedded in the set of sequences that are compatible with the given structure. Networks are modeled as graphs and constructed by random choice of vertices from the space of compatible sequences. The theory characterizes neutral networks by the mean fraction of neutral neighbors ($\lambda$). The networks are connected and percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value ($\lambda > \lambda^*$). Below threshold ($\lambda < \lambda^*$), the networks are partitioned into a largest “giant” component and several smaller components. Structures are classified as “common” or “rare” according to the sizes of their pre-images, i.e. according to the fractions of sequences folding into them. The neutral networks of any pair of two different common structures almost touch each other, and, as expressed by the conjecture of shape space coveting sequences folding into almost all common structures, can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared to data obtained by folding large samples of RNA sequences. Differences are explained in terms of specific features of RNA molecular structures. © 1997 Society for Mathematical Biology

NOMENCLATURE

\[ v[G] \] Vertex set of graph $G$
\[ e[G] \] Edge set of graph $G$

1Dedicated to Professor Manfred Eigen.
2Author to whom all correspondence should be addressed.
\(\omega(X)\) Cardinality of \(X\) as a set
\(\delta_v\) Vertex degree in a corresponding graph \(G\)
\(\mathcal{G}_a^n\) Generalized hypercube
\(\hat{X}\) \(X\) is a random variable
\(\mathbb{E}[\hat{X}]\) Expectation value of the random variable \(\hat{X}\)
\(\mathbb{V}[\hat{X}]\) Variance of \(\hat{X}\)
\(\mathbb{E}[\hat{X}]_r\) \(r\)th factorial moment of \(\hat{X}\)
\(\mu_{n,\lambda}, \mu_n\) Measure \(\mu_n(\Gamma_n) = \lambda \omega(\mathcal{V}[\Gamma])(1 - \lambda)\omega(\mathcal{V}[\Gamma]) - \omega(\mathcal{V}[\Gamma])\)
\(\Omega_n\) Probability space \((\{\mathcal{F}_n\}, \mathcal{W}_n, \mathcal{A})\)
\(X_{n,k}\) Number of vertices in a random graph \(\Gamma_n\) having degree \(k\)
\(k_n(\Gamma_n) = \omega(\{u \in \mathcal{V}[\Gamma_n] | \mathcal{V}(u) \cap \mathcal{V}[\Gamma_n] = \emptyset\})\), i.e., the number of isolated vertices in a random graph \(\Gamma_n\)
\(Z_n(\Gamma_n) = \omega(\{u \in \mathcal{V}[\mathcal{G}_n] | u \notin \mathcal{V}[\Gamma_n]\})\), i.e. the number of vertices in \(\mathcal{G}_a^n\) that are at least of distance 2 w.r.t. a random graph \(\Gamma_n\)
\(M_{n,v',v}(\Gamma_n)\) Set of paths \((\mathcal{V}(v_1) | \mathcal{V}(v_1) \in \Pi(\Gamma_n))\)
\(\mathcal{Y}_{n,v',v}(\Gamma_n) = \omega(M_{n,v',v}(\Gamma_n))\) for \(v, v' \in \mathcal{V}[\Gamma_n]\) and 0 otherwise
\(\Lambda_{n,k}\) Random variable that is 1 if all pairs \(v, v' \in \mathcal{V}[\Gamma_n]\) with \(d(v, v') < k\) occur in a path of \(\Gamma_n\) and 0 otherwise
\(\mathcal{V}_{G,V}\) Set of adjacent vertices w.r.t. a vertex set \(V \subset \mathcal{V}[G]\) in a graph \(G\)
\(\mathcal{B}_r(v)\) \(\mathcal{V}[\mathcal{B}_r(v)] = \{v' \in \mathcal{V}[\mathcal{G}_a^n] | d(v', v) \leq r\}\), the “ball” with radius \(r\) and center \(v\).
\(n\) Chain length
\(n_u, n_p\) Number of unpaired and paired bases of a certain secondary structure
\(\gamma_n\) \((\alpha - 1)n\), i.e. the vertex degree of \(\mathcal{S}_\alpha^n\)
\(s\) RNA secondary structure in \(n\) vertices
\(\Pi(s) = \{(i, k) | a_{i,k} = 1, k \neq i - 1, i + 1\}\), i.e. the set of contacts of the secondary structure \(s\)
\(\mathcal{S}_n\) Shape space, in particular, the space of RNA secondary structures in \(n\) vertices
\(\mathcal{G}[s]\) Graph of compatible sequences with respect to \(s\)
\(\mathcal{C}[s]\) \(\mathcal{V}[\mathcal{G}[s]]\), the set of compatible sequences
\(\mathcal{S}_n\) Permutation group of \(n\) letters
\(D_m\) Dihedral group of order \(2m\)
\(\Phi_x\) \(G_1 \times G_2[\mathcal{Y} \in \mathcal{V}[G_2] | (x, y) \in \mathcal{V}[\Gamma]]\), the fiber in \(x\)
\(\Phi_y\) \(G_1 \times G_2[\mathcal{X} \in \mathcal{V}[G_1] | (x, y) \in \mathcal{V}[\Gamma]]\), the fiber in \(y\)
\(\Gamma_n^A[s]\) Random induced subgraph of \(\mathcal{G}_a^n \times \mathcal{G}_b^n\) according to model A
\(\Gamma_n^B[s]\) Random induced subgraph of \(\mathcal{G}_a^n \times \mathcal{G}_b^n\) according to model B
\(\text{dist}(\Gamma_1,\Gamma_2)\) Minimum Hamming distance between the graph \(\Gamma_1\) and \(\Gamma_2\) considered as subgraphs of \(\mathcal{G}_a^n\).