LOCAL CONNECTIVITY OF NEUTRAL NETWORKS

CHRISTIAN M. REIDYS

Abstract. This paper studies local connectivity of neutral networks of RNA secondary and pseudoknot structures. A neutral network denotes the set of RNA sequences that fold into a particular structure. It is called locally connected, if in the limit of long sequences, the distance of any two of its sequences scales with their distance in the n-cube. One main result of this paper is that $\lambda_n = n^{-\frac{1}{2}+\delta}$ is the threshold probability for local connectivity for neutral networks, considered as random subgraphs of n-cubes. Furthermore, we analyze local connectivity for finite sequence length and different alphabets. We show that it is closely related to the existence of specific paths within the neutral network. We put our theoretical results into context with folding algorithms into minimum-free energy RNA secondary and pseudoknot structures. Finally, we relate our structural findings with dynamics by discussing the role of local connectivity in the context of neutral evolution.

1. Introduction and background

In this paper, we introduce neutral networks of RNA pseudoknot structures and analyze their local structure. We study these neutral nets, asking the following question: suppose we have two sequences of small Hamming distance contained in a neutral network, when does there exist a short neutral path connecting them? The property guaranteeing the existence of these short neutral paths is called local connectivity and arises in the context of random graph modeling of neutral nets in a natural way. In the process of bridging between the limit of long sequences (the basis for the random graph construction) and finite sequence length, we discover that local connectivity is closely related to the existence of specific paths. We show that locally connected neutral networks of pseudoknot RNA are generic, employing recent combinatorial and probabilistic results [17, 20, 14] and a new random graph theorem proved here.

Date: August 2008.

Key words and phrases. neutral evolution, local connectivity, threshold value, secondary structure, pseudoknot structure, random graph.

1.1. **RNA structures.** In the following we shall provide the basic background and context on RNA secondary and pseudoknot structures, their representation and neutral networks. Without doubt, over the last decade our perspective on RNA in organisms has shifted dramatically [34]. Once considered only an intermediate step between DNA and protein we have at present time an impressive amount of data establishing a variety of RNA functions. Of particular interest for us is one unique RNA feature: its ability to act as genotypic legislative in the form of viruses and viroids and as phenotypic executive in the form of ribosomes, capable of catalytic activity, cleaving other RNA molecules. This is one key feature of Schuster's *RNA world* [36]. For us, RNA represents an ideal conceptual arena in which Motoo Kimura's neutral theory [21, 22] can be developed further. It is clear that more complex genotype-phenotype mappings, for instance into RNA pseudoknot structures, are of key interest in this context.

Let us proceed by reviewing some basic facts on RNA sequences and structures. An RNA (primary) sequence is the sequence of nucleotides **A**, **G**, **U** and **C** and an RNA structure is the helical configuration of an RNA primary sequence, together with the Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairing rules. The central importance of RNA structures lies in the fact that their structures is oftentimes tantamount to their function. However, also sequence specific information (local information) is relevant. For instance, specific, "fixed" nucleotides play a particular role for the folding pathways [6] and stabilization of the tertiary structure of the phenylalanine tRNA structure [2]. We shall identify structures with diagrams [38]. That is, we draw the nucleotide-labels $1,\ldots,n$ in a horizontal line and draw arcs-labels (i,j) in the upper half-plane, if and only if i and j are paired in the structure. We call a diagram k-noncrossing if it does not contain k arcs that mutually cross each other. The length of an arc (i,j) is given by $j-i$ and a stack of length σ is a sequence of "parallel" arcs of the form $((i,j),(i+1,j-1),\ldots,(i+(\sigma-1),j-(\sigma-1)))$. A diagram is called σ-canonical or simply canonical instead of 2-canonical, if it does *not* contain any isolated base pairs and has arc-length ≥ 2. In Figure 1 we display 2- 3- and 4-noncrossing diagrams. While diagrams have a "Raison d'etre" as purely combinatorial objects [4] they offer a very intuitive representation of k-noncrossing structures.

Neutral networks of RNA secondary structures, see Figure 2, have been investigated via:
(a) exhaustive enumeration: the studies [9, 10, 8] employ the algorithm *ViennaRNA* [11] which derives for a given RNA sequence its minimum free energy secondary structure.
(b) structural analysis: considering the embedding of neutral networks into sequence space has led to the Intersection Theorem [27], which guarantees the existence of at least one sequence which is compatible to any two given secondary or pseudoknot structures. This shows that neutral networks

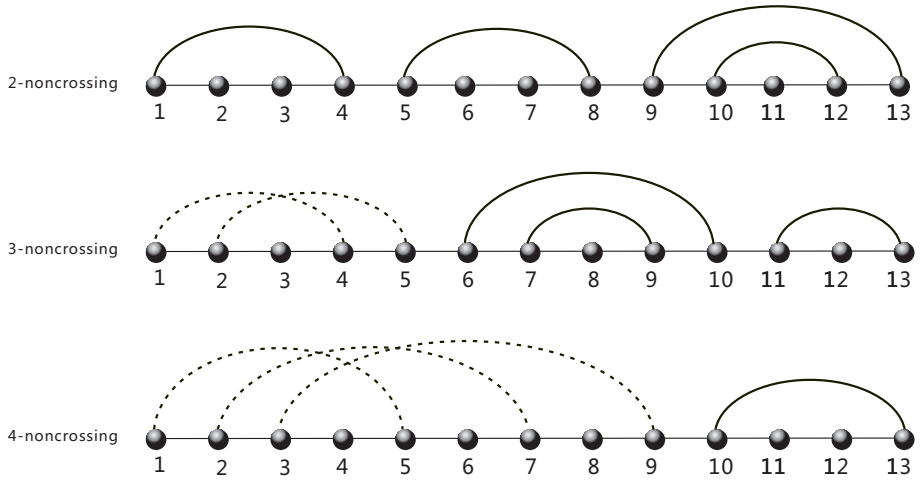

FIGURE 1. k-noncrossing structures: 2- 3- and 4-noncrossing structures (top to bottom). The arcs contained in the maximal set of mutually crossing arcs are dashed.

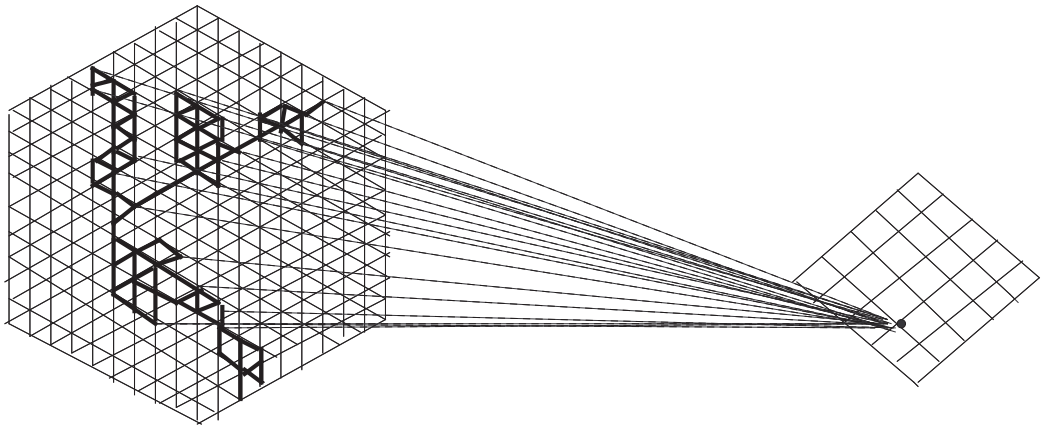

FIGURE 2. Neutral networks: sequence space (left) and shape space (right) represented as lattices. Edges between two sequences are drawn bold if they both map into the given structure. Two key properties of neutral nets are connectivity and percolation. They facilitate neutral evolution.

come "close" in sequence space and has led to exciting experimental work, see, for instance, [35]. (c) random graph modeling: the structure of neutral networks has been studied via random subgraphs of n-cubes [27, 28, 30, 29].

Two important notions that originated from (c) are the concepts of connectivity and density of neutral networks. A neutral network is connected if between any two of its sequences there exists a neutral path connecting them. Neutral paths were investigated by Schuster *et al.* in [37]. Furthermore a neutral network is called ρ-dense if the Hamming ball of radius ρ for an arbitrary sequence has nontrivial intersection with the neutral network. Density is closely related to Schuster's shape space covering conjecture [36, 10]. Of course, by construction, the neutral network depends on the particular concept of structure being employed. In this paper we shall identify structure with the sequence of pairs of nucleotides establishing chemical bonds and any notion of spatial embedding is not considered.

1.2. **Neutral Networks.** In this section we extend the concept of neutral networks from RNA secondary to RNA pseudoknot structures. We furthermore observe that sequence to structure mappings into canonical RNA secondary and pseudoknot structures exhibit two generic properties: (a) there *always* exist neutral networks of exponential size and (b) there *typically* exist exponentially many different structures. Remarkably, (a) and (b) are implied by the combinatorics of the structures themselves. Let us begin by remarking first that, for biophysical reasons, (folding maps produce typically minimum free energy structures) only canonical structures, i.e. structures having *no* isolated base pairs and arc-length ≥ 2 are of relevance, see Figure 3. Second, in the context of k-noncrossing RNA structures, a secondary structure [26, 43, 42, 44, 13] is simply a diagram having only *parallel* arcs and in which all bonds have at least length 2, see Figure 1.

Let us proceed by reviewing Schuster's argument for the existence of neutral networks for secondary structures [12]. Based on some variant of Waterman's basic recursion [42] for enumerating secondary structures over n vertices, Schuster *et al.* proved, using Darboux-type theorems [46], that there are asymptotically

$$1.4848\, n^{-3/2}\, 1.8444^n \tag{1.1}$$

secondary structures with arc-length ≥ 4 and stack-size ≥ 2. Clearly, since there are 4^n sequences over the natural alphabet this proves the existence of neutral networks.

RNA pseudoknot structures [34, 45] exhibit crossing arcs in the diagram representation. They occur in functional RNA (RNAseP [24]), ribosomal RNA [23] and are conserved in the catalytic core of group I introns. Several dynamic programming algorithms have been introduced for pseudoknot prediction [31, 41, 1]. Due to the cross-serial inter-dependencies [33] implied by pseudoknots,

dynamic programming algorithms can *a priori* recognize only restricted types and oftentimes it is nontrivial to formally specify a particular dynamic programming algorithm [32]. Arguably a major drawback of the dynamic programming paradigm [31, 41, 1, 25] is its lack of control of the maximum number of mutually crossing arcs. The maximal crossing number is a key parameter for the molecular complexity and controls linearly the exponential growth factor, see Table 1. The combinatorics of RNA pseudoknot structures has been derived in [17, 18]. Subsequent generalizations to tertiary interactions via a bijection between vacillating tableaux and tangled diagrams can be found in [4]. Again, only σ-canonical pseudoknot structures for $\sigma \geq 2$ are of relevance, see Figure 3. Canonical k-noncrossing structures have been studied in [20] where their asymptotic

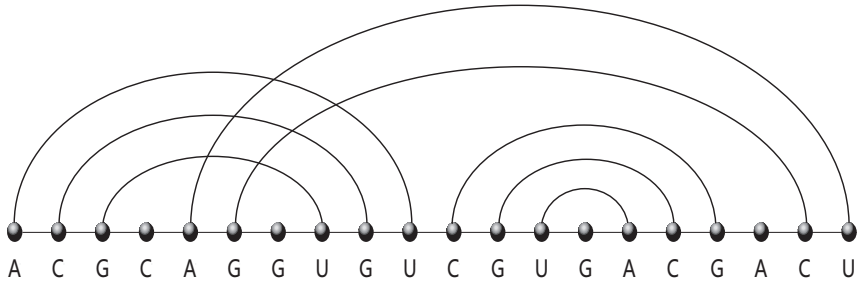

FIGURE 3. Canonical structures: each arc appears in a stack of size at least 2.

numbers are derived, see Table 1. Let $\mathsf{S}_3(n)$ and $\mathsf{S}_4(n)$ denote the numbers of canonical 3- and 4-noncrossing pseudoknot structures. The analogue of Schuster's formula (eq. (1.1)) for arbitrary k-noncrossing RNA pseudoknot structures is proved in [20]. In particular we have

$$\mathsf{S}_3(n) \sim c_3\, n^{-5}\, 2.5881^n \quad \text{and} \quad \mathsf{S}_4(n) \sim c_4\, n^{-\frac{21}{2}}\, 3.0382^n, \tag{1.2}$$

where c_3, c_4 are known constants. Accordingly, there exist exponentially large neutral networks for all mappings into canonical k-noncrossing RNA pseudoknot structures.

Finally let us address (b). The key observation in this context are the central limit theorems for the numbers of arcs of k-noncrossing RNA structures [19, 14]. The central limit theorems imply, that the numbers of arcs of 2- and 3-noncrossing RNA structures are concentrated at $0.31\,n$ and $0.4\,n$, respectively. From this we can conclude that the neutral networks are exponentially small compared to sequence space over the natural alphabet. In other words, the number of different canonical structures grows exponentially. Observations (a) and (b) imply the existence of nontrivial sequence to structure maps, irrespective of details of the underlying energy model.

k	2	3	4	5	6	7	8	9	10
$\sigma = 1$	2.6180	4.7913	6.8541	8.8875	10.9083	12.9226	14.9330	16.9410	18.9472
$\sigma = 2$	1.9680	2.5881	3.0382	3.4138	3.7438	4.0420	4.3162	4.5715	4.8115
$\sigma = 3$	1.7160	2.0477	2.2704	2.4466	2.5955	2.7259	2.8427	2.9490	3.0469
$\sigma = 4$	1.5782	1.7984	1.9410	2.0511	2.1423	2.2209	2.2904	2.3529	2.4100
$\sigma = 5$	1.4899	1.6528	1.7561	1.8347	1.8991	1.9540	2.0022	2.0454	2.0845
$\sigma = 6$	1.4278	1.5563	1.6368	1.6973	1.7466	1.7883	1.8248	1.8573	1.8866
$\sigma = 7$	1.3815	1.4872	1.5528	1.6019	1.6415	1.6750	1.7041	1.7300	1.7533
$\sigma = 8$	1.3454	1.4351	1.4903	1.5314	1.5645	1.5923	1.6165	1.6378	1.6571
$\sigma = 9$	1.3164	1.3941	1.4417	1.4770	1.5054	1.5291	1.5497	1.5679	1.5842
$\sigma = 10$	1.2925	1.3610	1.4028	1.4337	1.4585	1.4792	1.4971	1.5129	1.5270

TABLE 1. The exponential growth rates for pseudoknot RNA [20]: $\sigma = 1$ corresponds to RNA structures with isolated arcs, $\sigma = 2$ are the canonical structures. Increasing σ means to have larger and larger minimum stack-sizes.

1.3. **Modeling neutral networks.** Having established that neutral networks generically exist, how can we understand their structural properties? Of course, there is always exhaustive enumeration. However, sequence to structure maps for RNA of length ≥ 40 are beyond current computational capabilities and for larger sequence lengths all that is available are "local data". We can typically not decide whether two given sequences folding into the phenylalanine tRNA are connected by a neutral path or what their neutral distance is. As a result, the modeling of neutral networks is not just some mathematical exercise.

Our Ansatz is as follows: we consider a fixed RNA structure, $\mathfrak{s}$. Let $C[\mathfrak{s}]$ denote the set of $\mathfrak{s}$-compatible sequences, consisting of all sequences that have at any two paired positions one of the 6 nucleotide pairs $(\mathbf{A},\mathbf{U})$, $(\mathbf{U},\mathbf{A})$, $(\mathbf{G},\mathbf{U})$, $(\mathbf{U},\mathbf{G})$, $(\mathbf{G},\mathbf{C})$, $(\mathbf{C},\mathbf{G})$. We immediately realize that the structure $\mathfrak{s}$ gives rise to a new adjacency relation within $C[\mathfrak{s}]$. Indeed, we can reorganize a sequence $(x_1,\ldots,x_n)$ into the tuple

$$\big((u_1,\ldots,u_{n_u}),(p_1,\ldots,p_{n_p})\big)\,, \tag{1.3}$$

where the u_j denote the unpaired nucleotides and the $p_j = (x_i, x_k)$ all base pairs, respectively, see Figure 4. We can view $v_u = (u_1,\ldots,u_{n_u})$ and $v_p = (p_1,\ldots,p_{n_p})$ as elements of the cubes $Q_4^{n_u}$ and $Q_6^{n_p}$, implying the new adjacency relation for elements of $C[\mathfrak{s}]$. That is, $C[\mathfrak{s}]$ carries the natural graph structure $Q_4^{n_u} \times Q_6^{n_p}$, where "$\times$" denotes the direct product of graphs. We remark that this decomposition is valid whether or not we have crossing arcs. We will discuss the

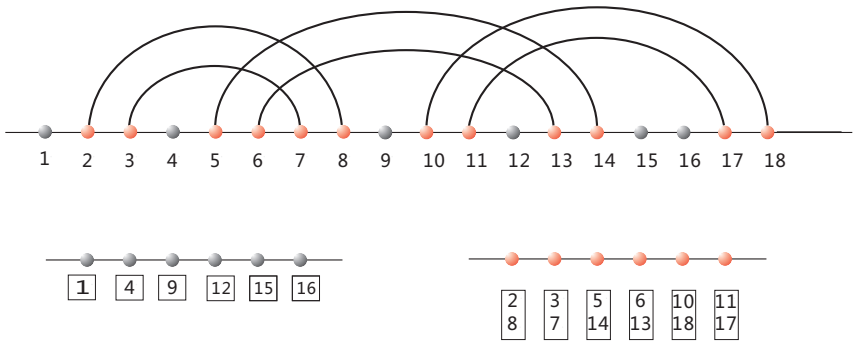

FIGURE 4. Deriving the two subcubes, $Q_4^{n_u}$ and $Q_6^{n_p}$: a structure gives rise to rearrange a compatible sequence into unpaired and paired segment. The former is a sequence over the original alphabet **A**, **U**, **G**, **C** and for the latter we derive a sequence over the alphabet of base pairs, $(\mathbf{A},\mathbf{U})$, $(\mathbf{U},\mathbf{A})$, $(\mathbf{G},\mathbf{U})$, $(\mathbf{U},\mathbf{G})$, $(\mathbf{G},\mathbf{C})$, $(\mathbf{C},\mathbf{G})$.

particular relation of the subcubes $Q_4^{n_u}$ and $Q_6^{n_p}$ with sequence space in detail in Section 3, see Figure 6. Therefore, *a priori*, the neutral network of $\mathfrak{s}$ is contained in its compatible sequences. The next step is to decide whether or not some compatible sequence is contained in the neutral network. The model Ansatz of [27] can be viewed as a mean-field approach and selects the vertices v_u and v_p with independent probability λ_u and λ_p, respectively. The probability λ_u and λ_p is easily measured *locally* via RNA computer folding maps: it coincides with the average fraction of neutral neighbors within the compatible neighbors. Explicitly, λ_u is the percentage of sequences that differ by a neutral mutation in an unpaired position, while λ_p corresponds to the percentage of neutral sequences that are compatible via a base pair mutation (for instance $(\mathbf{A},\mathbf{U}) \mapsto (\mathbf{G},\mathbf{C})$), see Figure 5.

Accordingly, the above construction reduces the random graph analysis of neutral networks to random subgraphs of the subcubes $Q_4^{n_u}$ and $Q_6^{n_p}$. From a conceptual point of view these two cubes "only" differ by the respective alphabet-length. This Ansatz allows to study random neutral networks via random subgraphs of n-cubes.

Maybe the most prominent structural feature of random induced subgraphs of n-cubes is the sudden emergence of a unique giant component at remarkably small vertex-selection probabilities [30].

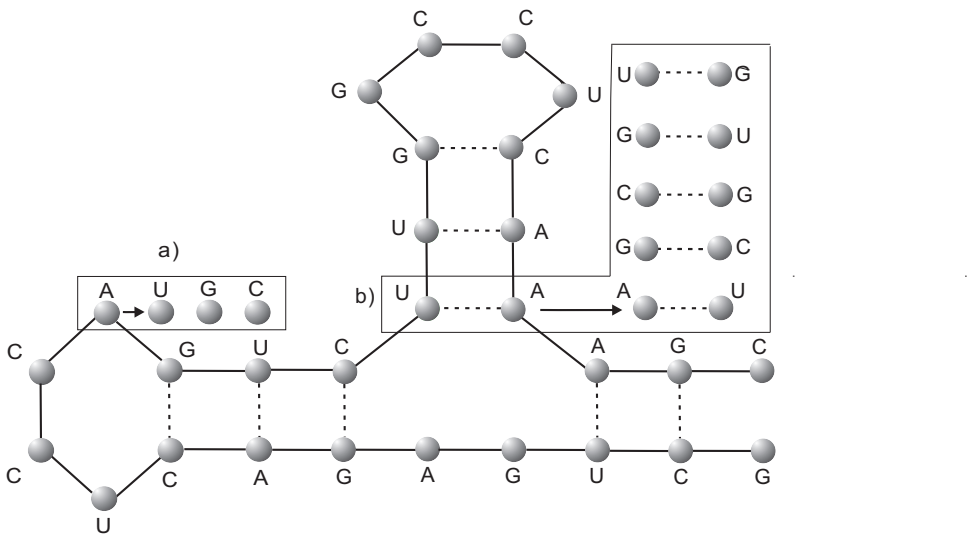

FIGURE 5. Compatible mutations: here we represent a secondary structure as a planar graph. The dashed edges correspond to the arcs in the upper halfplane of its diagram representation. We illustrate the different alphabets for compatible mutations in unpaired (a) and paired (b) positions, respectively.

Theorem 1.1. *Let Q^n_{2,λ_n} be the random graph consisting of Q^n_2-subgraphs, Γ_n, induced by selecting each Q^n_2-vertex with independent probability $\lambda_n = \frac{1+\epsilon}{n}$, where $\epsilon > 0$. Let $C^{(1)}_n$ denote the largest component in Γ_n. Then we have*

$$\lim_{n\to\infty} \mathbb{P}(\, |C^{(1)}_n| \sim c(\epsilon)\,\frac{1+\epsilon}{n}\,2^n \text{ and } C^{(1)}_n \text{ is unique}) = 1, \quad \text{where } c(\epsilon) > 0. \tag{1.4}$$

Remarkably, Theorem 1.1 remains valid for probabilities $\lambda_n = \frac{1+\chi_n}{n}$, where $o(1) = \chi_n \geq n^{-\frac{1}{3}+\delta}$. Then the largest component grows at the rate

$$|C^{(1)}_n| \sim 2\chi_n\,\frac{1+\epsilon}{n}\,2^n,$$

i.e. we have a "sleeping giant": despite the fact that the probability for a given vertex to be contained in it tends to zero, a unique largest component exists. While the result is originally formulated for binary alphabets, all proofs work verbatim in case of generalized n-cubes. In addition it is shown that the giant component is surprisingly uniformly distributed within the n-cube, see Lemma 5 [30]. Accordingly, only an average of $1+\epsilon$ neutral neighbors is needed in order to assure that a constant fraction of the neutral network is organized in a giant component.

However, sequence to structure maps exhibit significantly higher neutrality degrees [9, 10, 8]. The neutrality degrees found there are close to the random graph connectivity threshold [27, 28],

$$\lambda^* = 1 - \sqrt[\alpha-1]{\alpha^{-1}},$$

where α denotes the alphabet size. At this point isolated sequences suddenly disappear and the neutral networks become almost surely (a.s.) connected. Close inspection of the structure of neutral networks at the connectivity threshold shows that the random graph does *not* undergo major structural changes. "All" that happens is that the isolated vertices suddenly vanish. This gives rise to the question, whether there exists another key threshold probability in the evolution of these random graphs between the emergence of the giant component at $\lambda_n = \frac{1+\epsilon}{n}$ and the connectivity, localized at $\lambda^* = 1 - \sqrt[\alpha-1]{\alpha^{-1}}$.

1.4. **Background and context.** The results of this paper build on the framework and ideas developed in [27, 29]. They are motivated by the extensive studies on sequence to structure maps [9, 10, 8] and recent insights in RNA pseudoknot structures [17, 20]. In particular the growth rate of 2.5881 for 3-noncrossing pseudoknot structures is of key importance for considering sequence to structure maps into pseudoknot RNA structures. In the proof of the connectivity theorem (in difference to the random graph literature where the nonexistence of certain components is established) short paths between sequences are explicitly constructed. In [29] (as a pure random graph result) it is shown that beyond the threshold probability $\lambda_n = n^{-\frac{1}{2}+\delta}$ short paths exist. On the level of random graphs alone our results improve this in two aspects: first our construction improves on the length of the paths by a factor of at least $7/4$ and secondly we prove that the probability in question is exactly the threshold probability. This is possible by proving Lemma 5.2 and using a different approach. We do not "avoid" but "control" the actual correlations between the paths. We use the notion "local connectivity" tossed by Forst *et al.* in a computational study [7] on connectivity of neutral nets. There the objective is to derive a local criterion for testing connectedness.

As already pointed out in the previous section, our interest lies in a property of neutral networks that suddenly emerges long after the giant component already contains almost all vertices [28]. Suppose we have two sequences v and v' contained in a neutral network that are at Hamming distance 2. Then there are exactly two shortest paths starting at v and ending at v'. None of them is necessarily a path contained in the neutral net but in light of the existence of the giant component and exhaustive analysis of sequence to structure maps [9, 10, 8], v and v' are likely to be connected via some neutral path. Therefore it is possible that, in order to neutrally connect v and

v', we have to traverse the entire sequence space! In locally connected neutral networks the above scenario is not likely to appear. There typically exist short neutral paths between elements of small Hamming distance. We will localize the threshold value for local connectivity in Section 2 and in Section 3 we study local connectivity in the context of folding algorithms into minimum free energy secondary and pseudoknot structures. In the process we develop some structural understanding of local connectivity for finite sequence length and its role for neutral evolution. Finally we integrate our results in Section 4.

2. Local connectivity in random subgraphs of n-cubes

Without loss of generality we restrict our analysis in the following to binary n-cubes. All results remain valid for generalized n-cubes and the binary case allows us to formulate the key results avoiding unnecessary notational burden. However, wherever needed, we formulate our findings explicitly for generalized n-cubes, see Corollary 2.2. Let us recall some basics of random induced subgraphs of n-cubes: we select Q_2^n-vertices with independent probability λ_n. Each selection process yields a subset $A \subset Q_2^n$. The set A induces an induced subgraph in Q_2^n in a natural way: $a_1, a_2 \in A$ are adjacent if and only if a_1, a_2 are adjacent in Q_2^n. We have the probability measure

$$\mathbb{P}(A) = \lambda_n^{|A|}\,(1-\lambda_n)^{2^n-|A|}.$$

Suppose Γ_n denotes a random subgraph of Q_2^n. A property $\mathcal{P}$, is a subset of Q_2^n-subgraphs (closed under graph-isomorphisms) and "$\mathcal{P}$ holds a.s." means $\lim_{n\to\infty}\mathbb{P}(\mathcal{P}) = 1$. For instance, for the property "A is connected" for $\lambda > 1/2$ a.s. every random graph is connected and for $\lambda < 1/2$ a.s. none is.

We arrive at the following problem formulation: Given two Q_2^n-vertices v, v' with Hamming distance d–what is their Γ_n-distance? To be precise:

$$(\dagger) \qquad \exists\,\Delta > 0; \quad d_{\Gamma_n}(v,v') \leq \Delta\, d_{Q_2^n}(v,v') \quad \text{a.s., provided } v, v' \text{ are in } \Gamma_n.$$

In order to avoid any confusion: of course, trivially, for any *finite* n such a Δ exists. Since $(\dagger)$ employs the notion "a.s." it is valid for arbitrarily large n. In other words, one fixed Δ has to exist for *any* sequence length larger than some n_0. In the following we will study under which circumstances such a constant Δ exists. We shall prove that for λ_n smaller than $n^\delta/\sqrt{n}$, where $\delta > 0$ is arbitrarily small, there exists a.s. no finite Δ satisfying $(\dagger)$. On the other hand, for λ_n

larger or equal than $n^\delta/\sqrt{n}$, there exists a.s. some finite Δ satisfying (†). The proofs given in Section 5 show how to compute Δ as function of δ.

Theorem 2.1. *Let $0 < \delta \leq \frac{1}{2}$ and v, v' be arbitrary but fixed Q_2^n-vertices, having distance $d_{Q_2}(v, v') = d$, $d \geq 2$, $d \in \mathbb{N}$. Let Γ_n denote the random subgraph of Q_2^n, obtained by independently selecting Q_2^n-vertices with probability λ_n. Suppose v, v' are contained in Γ_n, then the following assertions hold*

(a) *For $\lambda_n < n^{\delta-\frac{1}{2}}$, $\delta > 0$, there exists a.s. no $\Delta > 0$ satisfying*

$$d_{\Gamma_n}(v, v') \leq \Delta \, d_{Q_2^n}(v, v') \; . \tag{2.1}$$

(b) *For $\lambda_n \geq n^{\delta-\frac{1}{2}}$, $\delta > 0$, there exists a.s. some finite $\Delta = \Delta(\delta) > 0$ such that*

$$d_{\Gamma_n}(v, v') \leq \Delta \, d_{Q_2^n}(v, v') \; . \tag{2.2}$$

In view of Theorem 2.1 we shall call a random graph Γ_n locally connected if $\lambda_n \geq n^{\delta-\frac{1}{2}}$ for some $\delta > 0$. We observe that, beyond the critical probability $n^{\delta-\frac{1}{2}}$, a.s. a Q_2^n-ball of radius d centered at v, transforms into a Γ_n-ball of radius $\Delta \, d$. This means that if Γ_n is locally connected it can be viewed as a Δ-dilated n-cube.

Theorem 2.1 proves that the giant component undergoes significant structural changes besides just growing in size. At its emergence, around $\lambda_n = \frac{1+\chi_n}{n}$, where $\chi_n = o(1)$ it is according to [30] a.s. of size $2\chi_n \frac{1+\chi_n}{n} 2^n$. At this stage the giant component is of limited usefulness for evolutionary optimization: for λ_n close to $\frac{6}{n}$ an argument given by Balister *et al.* [3] can be used in order to prove that typical distances are of the order n. Consequently, the entire sequence space has to be traversed in order to connect two sequences of distance 2. According to Theorem 2.1, structural change occurs for probabilities around $1/\sqrt{n}$. Suddenly, if two sequences contained in the neutral network have small Hamming distance, then there exists a short neutral path with high probability.

We next study Δ for constant probabilities $0 < \lambda \leq 1$. Revisiting the proof of Theorem 2.1 we can immediately conclude

Corollary 2.2. *Let $b, d \in \mathbb{N}$, $b, d \geq 2$, v, v' be arbitrary but fixed Q_b^n-vertices, having distance $d_{Q_b}(v, v') = d$ and $n' = n - (d-1)$. Suppose we select Q_b^n-vertices with the probability $0 < \lambda < 1$. Then there exists a Γ_n-path connecting v and v' of length exactly $2 + d$ with probability at least*

$$\sigma_{\lambda,d}^{[b]}(n) = 1 - \exp\left(-\frac{(b-1)\, n' \, \lambda^{2+(d-1)}}{4}\right), \tag{2.3}$$

provided v, v' are contained in Γ_n.

We remark that Corollary 2.2 "almost" implies the connectivity theorem for random subgraphs of n-cubes. In order to recover the connectivity theorem we only need to observe that at the threshold any Γ_n-vertex has arbitrarily large *finite* degree. This allows us to employ Corollary 2.2 "in parallel" for each of those vertices.

Unfortunately, for all practical purposes n is *never* sufficiently large. To make this precise let, for instance, $n = 70$, $d = 2$ and $\lambda = 0.5$. According to Corollary 2.2, a Γ_n-path exists in the binary n-cube with probability

$$\sigma^{[2]}_{\frac{1}{2},2}(70) = 1 - \exp\left(-69 \cdot 0.5^3/4\right) \approx 0.8843. \tag{2.4}$$

Therefore, although in the limit of long sequences we are a.s. guaranteed to find such a Γ_n-path of length 4, *de facto*, for binary alphabets there is still more than 11% chance of failure. But what about the natural alphabet? Here we immediately obtain

$$\sigma^{[4]}_{\frac{1}{2},2}(70) = 1 - \exp\left(-3 \cdot 69 \cdot 0.5^3/4\right) \approx 0.9984. \tag{2.5}$$

This illustrates a key difference between binary and quaternary alphabets: sequence space over the natural alphabet does guarantee significantly higher probability of finding the short neutral paths. This gives rise to the question whether or not there exists an alternative to larger alphabet sizes? To this end we observe that Theorem 2.1 implies a second corollary, which is a consequence of considering slightly longer paths.

Corollary 2.3. *Let v, v' be arbitrary but fixed Q_2^n-vertices, having distance $d_{Q_2}(v, v') = d$, $d \geq 2$ and $n' = n - (d-1)$. Suppose we select Q_2^n-vertices with the probability $0 < \lambda < 1$. Then there exists a Γ_n-path connecting v and v' of length exactly $4 + d$ with probability at least*

$$\tau_{\lambda,d}(n) = 1 - \exp\left(-\left[\frac{2}{\lambda^2\left[\frac{n'-2}{n'-1}\right]n'} + \frac{2(2+\lambda^2)}{n'(n'-1)\lambda^{4+(d-1)}}\right]^{-1}\right), \tag{2.6}$$

provided v, v' are contained in Γ_n.

Indeed, Corollary 2.3 represents a significant improvement over Corollary 2.2: for sufficiently large n we have

$$\tau_{\lambda,d}(n) = 1 - \exp\left(-\left[\frac{2}{\lambda^2\left[\frac{n'-2}{n'-1}\right]n'} + \frac{2(2+\lambda^2)}{(n'(n'-1)\lambda^{3+d}}\right]^{-1}\right) \sim 1 - \exp\left(-\frac{\lambda^2 n'}{2}\right). \tag{2.7}$$

That is, the actual distance between v and v' does not factor in. We remark that we compute the *exact* correlation terms in Corollary 2.2 and Corollary 2.3. In this sense we obtain best possible results. The only possible improvement could result from revisiting Janson's inequality itself, see Lemma 5.1. As for the natural alphabet, we observe that in order to obtain an improvement over $\sigma^{[4]}_{\frac{1}{2},2}(n)$, Corollary 2.3 requires a sequence length of $n > 51$, see Figure 12 in Section 4 and the discussion therein. For shorter sequence length the longer alphabet size is the only way to guarantee reliably local connectivity. We remark that the improvement is "only" by a constant factor despite the fact that $O(n^2)$ path were considered. It confirms the intuition that the critical stage for finding the neutral path is checking the immediate v and v' neighbors. Not surprisingly, additional improvement by increasing the length of the neutral path beyond $4+d$ is only marginal and as a result, if short paths exist, they can be found quickly.

3. From random graphs to RNA

In this section we put the theory to a test and provide some data on local connectivity for RNA secondary and pseudoknot structures. Since it is not the scope of this paper to provide exhaustive computer data, we restrict ourselves to a few paradigmatic studies. Suppose we are given a structure $\mathfrak{s}$ and sequence v, contained in its neutral network. One main objective is to make local connectivity measurable for finite sequence length. Local connectivity refers to the two n-cubes $Q_4^{n_u}$ and $Q_6^{n_p}$, induced by rearranging a sequence of the neutral network into its unpaired and paired segments, see eq. (1.3) and Figure 4. Accordingly, there are two types of compatible neighbors in sequence space: u- and p-neighbors: a u-neighbor has Hamming distance one and differs exactly by a point mutation at an unpaired position. Analogously a p-neighbor differs by a compatible base pair-mutation, see Figures 5. A p-neighbor has either Hamming distance one $((\mathbf{G},\mathbf{C}) \mapsto (\mathbf{G},\mathbf{U})))$ or Hamming distance two $((\mathbf{G},\mathbf{C}) \mapsto (\mathbf{C},\mathbf{G})))$. We call a u- or a p-neighbor, y, a compatible neighbor. If y is contained in the neutral network we refer to y as a neutral neighbor. It is therefore natural to consider the compatible- and neutral distance, denoted by $C(v,v')$ and $N(v,v')$. These are the minimum length of a $C[\mathfrak{s}]$-path and path in the neutral network between v and v', respectively.

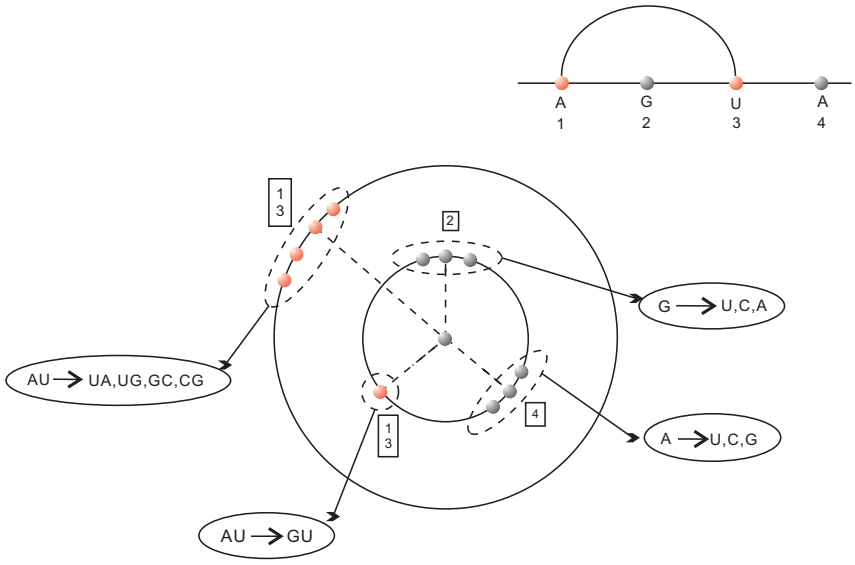

FIGURE 6. Compatible neighbors in sequence space: diagram representation of an RNA structure (upper right) and its induced compatible neighbors in sequence space (lower left). Note that each base pair gives rise to 5 compatible neighbors exactly one of which is in Hammimg distance one.

Let $C_2 = |\{v' \mid C(v,v') = 2\}|$. We proceed by defining a measure of local connectivity:

$$\partial_{\mathfrak{s}}(v) = |\{v' \mid C(v,v') = 2,\ N(v,v') \leq 4\}|\, C_2^{-1}. \tag{3.1}$$

We call $\partial_{\mathfrak{s}}(v)$ the degree of local connectivity of $\mathfrak{s}$ at v. In other words, $\partial_{\mathfrak{s}}(v)$ is the fraction of the compatible distance two neighbors of v, that are connected via Corollary 2.2-paths. Eq. (3.1) can be viewed as a "testable" criterion for local connectivity of RNA structures. Corollary 2.2 implies that for neutral networks, modeled as random graphs above the threshold, a.s. all neutral vertices at compatible distance two are locally connected. Accordingly, we then have, in the limit of long sequences

$$\partial_{\mathfrak{s}}(v) \sim |\{v' \mid v' \text{ is neutral and } C(v',v) = 2\}|\, C_2^{-1}. \tag{3.2}$$

That is, the degree of local connectivity simply coincides then with the binomially distributed random variable counting the fraction of neutral sequences at compatible distance two. In particular, $\partial_{\mathfrak{s}}(v)$ is in this case independent of v. In the following we shall, by abuse of notation, refer to the degree of local connectivity and the absolute number of locally connected sequences simply as local connectivity.

3.1. **Local connectivity for RNA secondary structures.** We study in the following local connectivity of the phenylalanine tRNA denoted by $\mathfrak{s}_{\text{phe}}$, see Figure 7, whose natural sequence, v_{phe}, is given by

ACCACGCUUAAGACACCUAGCUUGUGUCCUGGAGGUCUAAAAGUCAGACCGCGAGAGGGUUGACUCGAUUUAGGCG

For our analysis of RNA secondary structures we use the folding algorithm *ViennaRNA* [11].

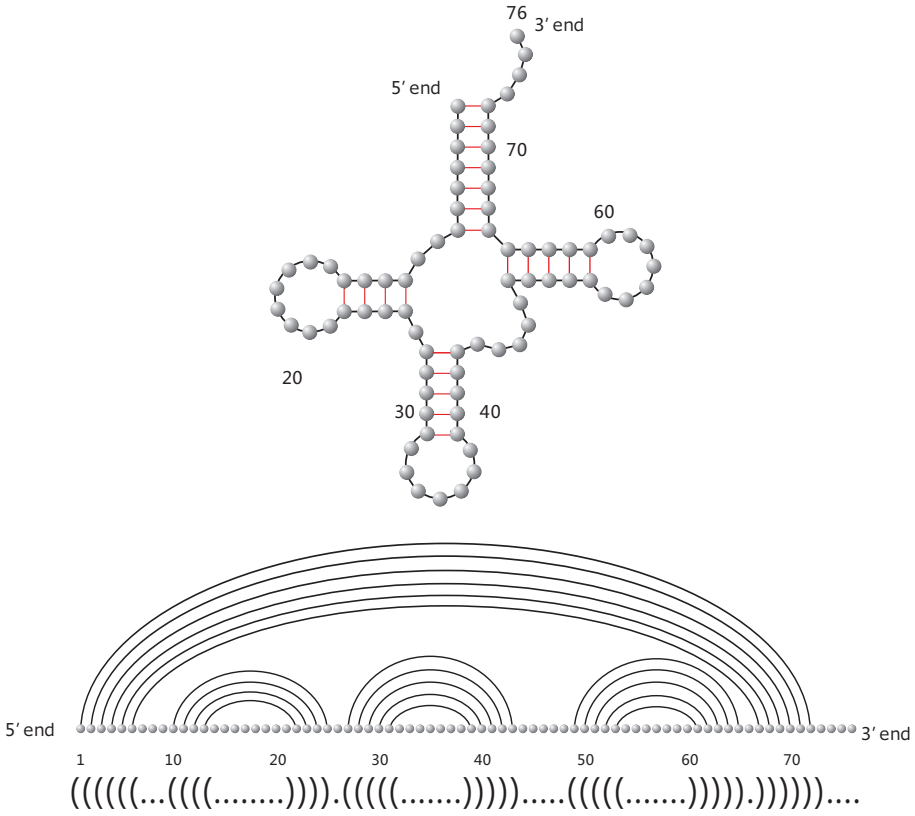

FIGURE 7. The phenylalanine tRNA secondary structure: represented as a planar graph (top), diagram (middle) and dot/bracket-sequence (bottom).

While local connectivity of random neutral networks is *a priori* an isotropic property, i.e. it does not depend on the choice of the base sequence, folding algorithms into minimum free energy RNA secondary structures are based on sequence specific loop-energies. As a result, we can expect systematic deviations from the random graph model. However, it is not *a priori* evident, how these deviations affect local connectivity.

We begin by generating via *Inversefold* of the ViennaRNA-package 240 sequences of the neutral network. Then we compute for each sequence, the number of locally connected sequences. These data are presented on the left hand side of Figure 8 in terms of a frequency histogram. Complementing the above, local connectivity is studied by inductively walking from the natural sequence $v_{\rm phe}$ with steps of compatible distance two and with strictly increasing the distance to the starting point. The corresponding data are obtained from 240 neutral sequences along such paths and shown on the right hand side of Figure 8.

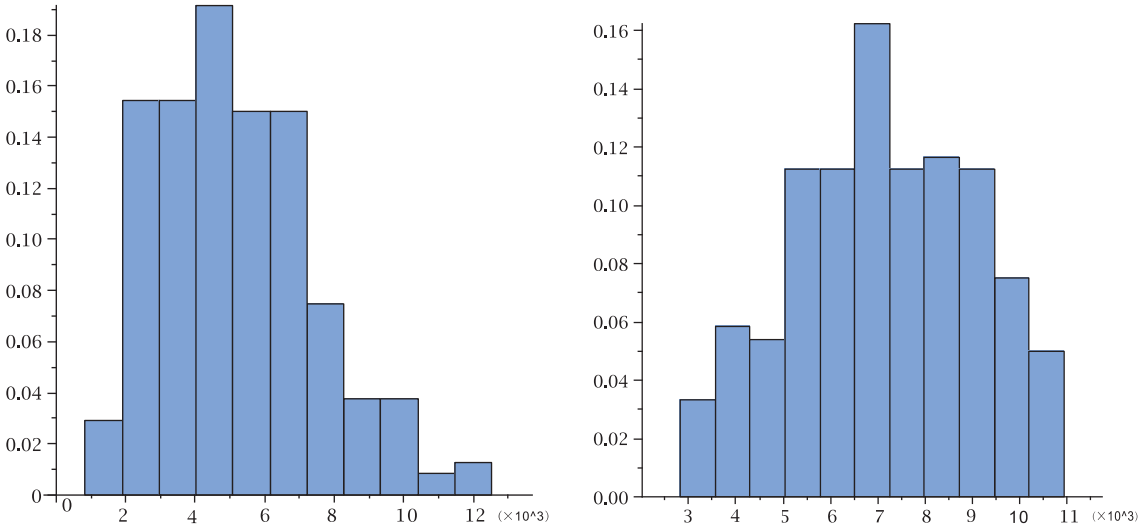

FIGURE 8. Local connectivities of the phenylalanine tRNA: we display the frequency distribution of local connectivities of 240 randomly sampled neutral sequences of the phenylalanine tRNA (left hand side). The x-axis shows the number of locally connected sequences and the y-axis shows the frequencies of the latter. The right hand side shows the frequency distribution of the local connectivities of 240 sequences sampled along a neutral path originated at the natural sequence.

We observe that the degrees of locally connectivity are very close to one. This fact suggests that the local connectivity of RNA secondary structures is in accordance with the predictions of random graph theory. The data displayed in Figure 8 exhibit distinct sequence dependencies, i.e. deviations from the random graph model. Inversefold-sequences exhibit an average of ≈ 5216 and a variance of $\approx 5.1021 \times 10^6$ locally connected sequences, while sequences of paths originated at $v_{\rm phe}$ have the significantly larger average local connectivity of ≈ 7111 with a variance of $\approx 3.7291 \times 10^6$.

In order to provide further context we randomly select ten 3-canonical RNA secondary structures of length 76 and generate via Inversefold 240 random sequences of their respective neutral networks. Again we analyze local connectivity observing an average of ≈ 4035 locally connected sequences with a standard deviation of $\approx 4.2559 \times 10^6$.

3.2. **Local connectivity for RNA pseudoknot structures.** Next we analyze local connectivity of RNA pseudoknot structures. To this end we use the algorithm cross [15], which computes the minimum free energy 3-noncrossing 4-canonical structure [15]. Cross is a hybrid algorithm employing branch and bound, as well as dynamic programming routines. It constructs the minimum free energy 3-noncrossing, 4-canonical structure via a sequence of substructures called shadows. The latter are obtained via motifs, i.e. 3-noncrossing diagrams with stack-size *exactly* 4, whose cores (see [20] for details) are nonnesting. A shadow of a motif is a structure obtained by extending some motif-stacks from top to bottom. Note that a given motif has in general many shadows. Clearly, any k-noncrossing diagram has a unique core, which is obtained by identifying its stacks by single arcs. It is straightforward to show [15], that each k-noncrossing σ-canonical RNA structure can be constructed inductively via motifs and shadows. Accordingly, cross is capable of searching all 3-noncrossing, canonical pseudoknot structures. As for our local connectivity study via cross, we choose the UTR pseudoknot of the mouse hepatitis virus (of length 56), see Figure 9. Its natural sequence is given by

(3.3) CUCUCUAUCAGAAUGGAUGUCUUGCUGUCAUAACAGAUAGAGAAGGUUGUGGCAGA.

In Figure 10 we display the frequency distribution of the local connectivities of 240 sequences sampled along a neutral path originated at the natural sequence. The particular construction of these paths is completely analogous to that of Section 3.1. The data show that local connectivity is also a reality for pseudoknot RNA. We find a mean of ≈ 7547 and variance of $\approx 1.9791 \times 10^6$ locally connected sequences. Our data confirm the hypothesis that many important features observed for genotype phenotype maps into RNA secondary structures also hold for RNA pseudoknot structures.

4. Discussion

We have studied local connectivity of neutral networks of RNA secondary and pseudoknot structures using the random graph model of [27]. In the process we showed that local connectivity implies the existence of "short paths" between neutral sequences. Local connectivity reflects a

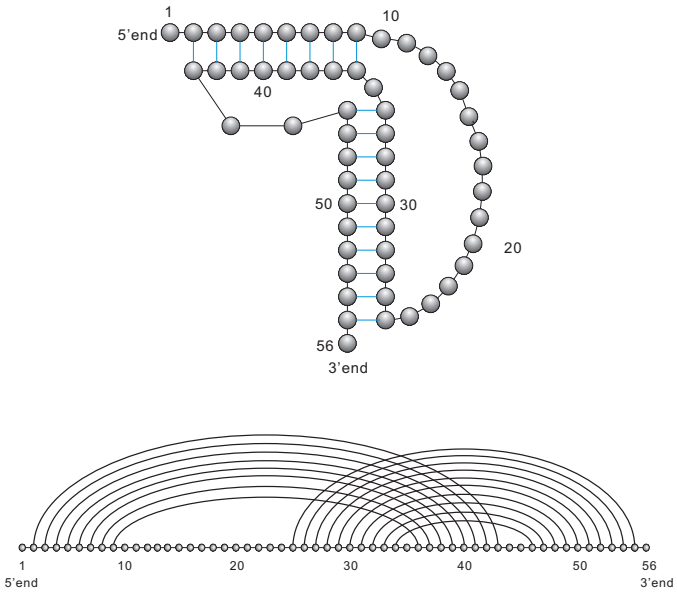

FIGURE 9. The UTR pseudoknot structure of the mouse hepatitis virus: planar graph- (top), diagram- (middle) and dot/bracket-representation. The algorithm `cross` confirms this particular structure for the natural sequence given in eq. (3.3).

structural relation between the neutral network and the sequence space. One may draw the analogy between the properties "Γ_n is Δ-connected at v" and "f has a derivative at v". Along these lines, local connectivity can be paraphrased as the "derivative of a graph". In more graph theoretic terms, local connectivity implies that a graph is locally not "tree-like". In fact, the key idea for proving Theorem 2.1 is to construct specific trees of arbitrary polynomial size, eventually limiting the number of ways for deriving certain splits.

Since local connectivity is a monotone graph property (i.e. once Γ_n is locally connected increasing the probability λ_n is not changing local connectedness), there exists a threshold value, localized via Theorem 2.1. This means that suddenly neutral networks become a.s. locally connected. If locally connected, neutral networks can be viewed as Δ-dilated n-cubes. Therefore, a small Hamming distance for two sequences on the neutral network implies the existence of a short neutral path connecting them.

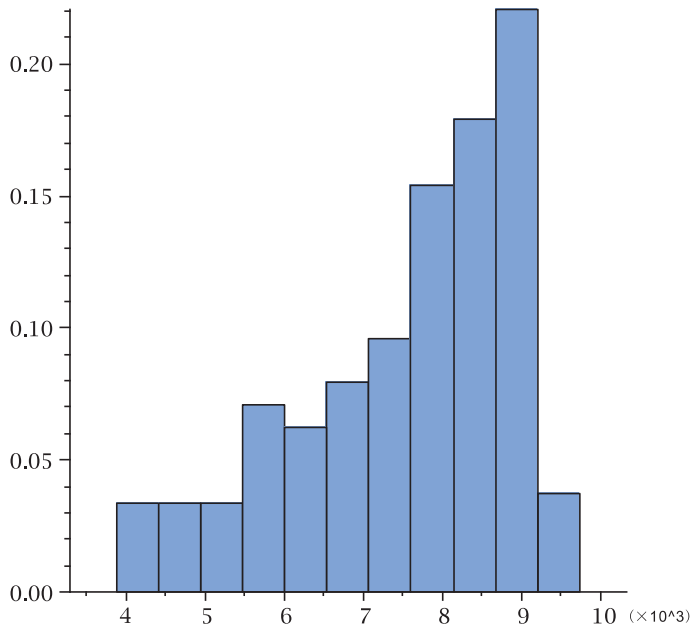

FIGURE 10. Local connectivity of the UTR pseudoknot of the mouse hepatitis virus via `cross`: we display the frequency distribution of the local connectivities of 240 sequences sampled along a neutral path originated at the natural sequence, given in eq. (3.3).

We have studied the immediate question arising in the context of an "almost surely"-formulation: the likeliness of local connectivity for *finite* sequence length n. In this context we have analyzed in Corollary 2.2 and Corollary 2.3 alternative path-classes. Of course it is of interest to find a neutral path for two sequences in the neutral network within the boundary of some shortest path in sequence space. The key difference between Corollary 2.2 and Corollary 2.3 is that the path-class of the former tests $O(n)$ and the latter tests $O(n^2)$ paths, respectively. We find in this context, that the natural alphabet plays a particular role: it does not imply the existence of shorter paths but it significantly lowers the probability of failure to find one.

Let us take a closer look: set $\nu(\lambda)$ to be the maximal sequence length such that

$$\sigma^{[4]}_{\lambda,2}(n) \geq \tau_{\lambda,2}(n). \tag{4.1}$$

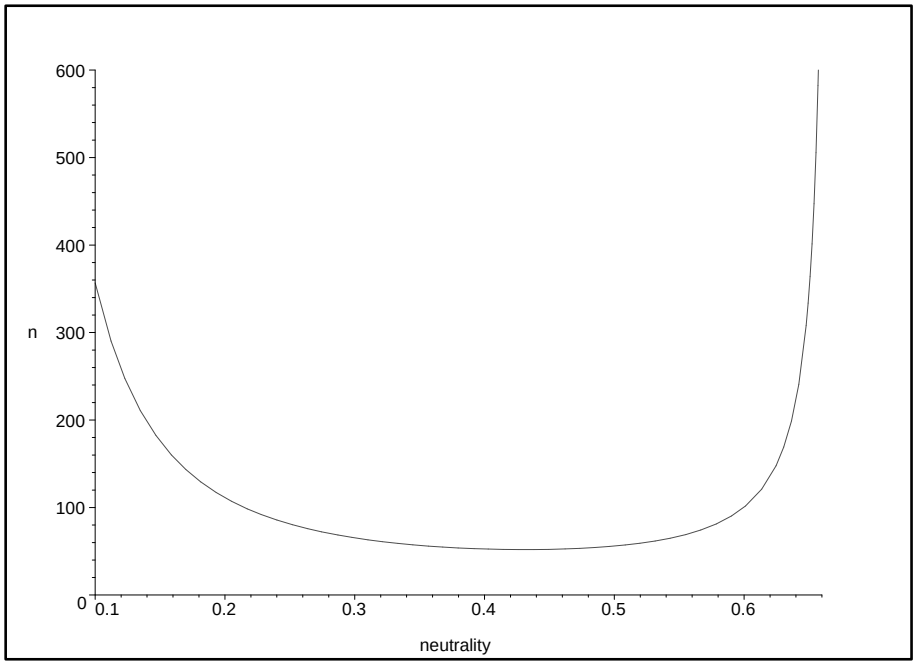

FIGURE 11. The role of the natural alphabet for local connectivity in random neutral networks: we display the curve $\nu(\lambda)$, i.e. the smallest sequence length at which local connectivity is improved by longer neutral paths. On the x-axis we have λ from 0 to 1 and the y-axis displays the sequence length $\nu(\lambda)$. The minimum of $\nu(\lambda)$ is located at $\mu^* \approx 0.4311$, where $\nu(\mu^*) = 51$ holds.

In particular, if eq. (4.1) holds for any n, we have $\nu(\lambda) = \infty$. Solving eq. (4.1) for n we derive

$$\nu(\lambda) = \begin{cases} \lfloor \frac{6\lambda^3 - 7\lambda^2 - 6}{\lambda^2 (3\lambda - 2)} \rfloor & \text{for } 0 < \lambda < \frac{2}{3} \\ \infty & \text{for } \lambda = 0 \text{ and } \frac{2}{3} \leq \lambda \leq 1, \end{cases} \tag{4.2}$$

which we display in Figure 11. $\nu(\lambda)$ exhibits some intriguing features: for $0.4 \leq \lambda \leq 0.5$, $\nu(\lambda)$ obtains its smallest values and exactly at $\lambda = \frac{2}{3}$ we pass from finite to infinite sequence length. Beyond neutrality degrees of $\lambda = \frac{2}{3}$ we observe that $\sigma^{[4]}_{\lambda,2}(n)$ cannot be exceeded by $\tau_{\lambda,2}(n)$ for any n. In other words, for sufficiently high neutrality degrees there is no need to consider longer neutral paths.

For all neutrality degrees and sequences over the natural alphabet, local connectivity is guaranteed by the short paths of Corollary 2.2. Our analysis shows that, within a certain range of neutrality,

considering longer paths marginally increases the probability of finding one, see Figure 12. If one

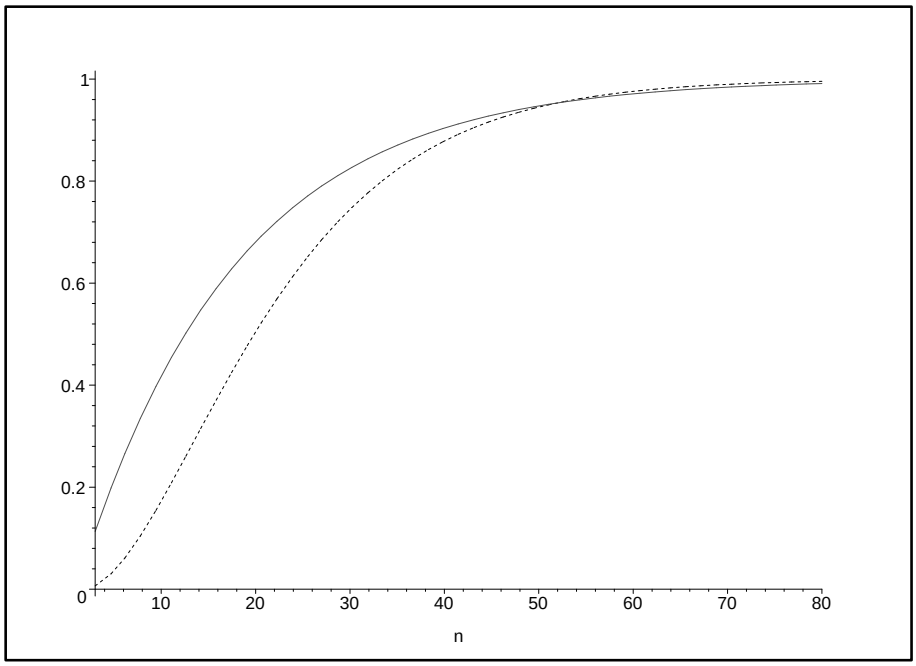

FIGURE 12. The critical neutrality $\lambda^\sharp$: we display the success probabilities $\sigma^{[4]}_{\lambda^\sharp,2}(n)$ (solid) and $\tau_{\lambda^\sharp,2}(n)$ (dashed) as functions in n. Both curves intersect at ≈ 51.3, i.e. we have $\nu(\lambda^\sharp) = 51$ which is the minimum of $\nu(\lambda)$, see in Figure 11. We observe, that for sequences of length > 51 longer neutral paths provide a marginal improvement in local connectivity.

believes in evolution to have started with short sequences the results suggest that the natural alphabet has indeed a key role by assuring local connectivity for short sequence length. At its minimum at $\lambda^\sharp \approx 0.4311$ we have $\nu(\lambda^\sharp) = 51$, see Figure 12 for details. For $\lambda^\sharp \approx 0.4311$ the natural alphabet remains vital for sequence lengths ≤ 51, since longer paths cannot improve local connectivity in this parameter range.

Furthermore our results show that local connectivity of neutral networks becomes more and more likely with increasing sequence length. We may paraphrase the situation by saying that evolving larger and larger molecules facilitates evolutionary search.

We have also shown that local connectivity is a reality for folding algorithms into minimum free energy structures. First, in the case of RNA secondary structures, we find local connectivity degrees of almost one, indicating that already for $n = 76$ there are neutral networks in which almost all sequences are connected via the path of Corollary 2.2. Second, we observe specific deviations from the isotropic model: for the phenylalanine tRNA displayed in Figure 7, the sequences related to the natural sequence exhibit a significantly higher local connectivities than randomly sampled sequences of its neutral network, see Figure 8. Furthermore we find that the local connectivity of the phenylalanine tRNA is significantly higher than that of random structures.

For RNA pseudoknot structures we have at present time no Inversefold algorithm. Therefore we cannot produce data on random sequences of their neutral networks. Further difficulty arises from the fact that currently there exists no polynomial time generation of random RNA pseudoknot structures (with uniform probability). This prohibits any comparative analysis with random RNA pseudoknot structures. However our neutral path data indicate, that we have for the UTR pseudoknot of the mouse hepatitis virus locally connected neutral networks. Interestingly, the far more complex class of pseudoknot structures appears to have neutral networks very similar to those of secondary structures.

The fact that the observed local connectivity degrees are so close to one implies that for the tRNA and the UTR pseudoknot structure the short paths of Corollary 2.2 connect neutral sequences in compatible distance two. However, there are for the tRNA and the UTR pseudoknot structure sequences in compatible distance two, that are not connected by Corollary 2.2-paths. For instance, for the tRNA we have

ACCACGCUUAAGACAUCUAGCUUGUGUCCUGGGGGUCUAAAAGUCAGACCGCGAGAGGGUUGACUCGAUUUAGGCG

ACCACGCUUAAGACACCUAGCUUGUGUCCUGGAGGUCUAAAAGUCAGACCGCGAGAGGGUUGACUCGAUUUAGGCG

and for the UTR pseudoknot structure, such a pair of sequences is given by

CUCUCUAUCAGAAUGGAUGUCUUGCUGUCAUAACAGAUAGAGAAGGUUGUGGCAGA

CUCUCUAUCAGAAUGGAUGUCUUUCUGUCAUAACAGAUAGAGAAGGUUGUGGCAGC.

Let us finally discuss local connectivity in the context of neutral evolution. For this purpose we consider erroneously replicating RNA strings, evolving on a neutral network. Suppose we have a disc with infinite radius, spinning at high speed (the speed being the mutation rate). Suppose further we have some entities (with mass) located close to the center of the disc. Instantly, these

entities will be catapulted "away" from each other and there is only a chance of coming mutually close at infinity. This thought-experiment illustrates what it means to be locally disconnected. A locally disconnected neutral network is locally a tree. The implication for an evolving population is the following: the only relation between two elements in the population comes from the fact that they had a common ancestor some generations ago. Consequently, the mutational force pulls the population apart. Only the dynamics itself, that is the particular replication scheme and the finite life time of individuals produce clusters within the population. It is clear that without local connectivity the population will inevitably split in the course of evolution into isolated individuals. It could not preserve any type of local, i.e. sequence specific information. A particular instance of local connectivity in neutral evolution appears in the work of Derrida and Peliti [5] who studied neutral evolution in the extreme case where the neutral network coincides with sequence space itself. The latter is obviously locally connected and the authors take this into consideration by relating the branching process with random walks on n-cubes.

5. Proof of Theorem 2.1

The first result in this section is Janson's inequality [16]. It is the key tool for proving Theorem 2.1. Intuitively, Janson's inequality can be viewed as a large deviation result in the presence of correlation.

Lemma 5.1. *Let R be a random subset of $[n] = \{1, \dots, n\}$, obtained by selecting each element $v \in [n]$ independently with probability λ. Let $S_1, \dots, S_s$ be subsets of $[n]$ and X be the r.v. counting the number of S_i, for which $S_i \subset R$ holds. Let furthermore*

$$\Omega = \sum_{(i,j);\, S_i \cap S_j \neq \varnothing} \mathbb{P}(S_i \cup S_j \subset R), \tag{5.1}$$

where the sum is taken over all ordered pairs (i, j). Then, for any $\gamma > 0$, we have

$$\mathbb{P}(X \leq (1-\gamma)\mathbb{E}[X]) \leq e^{-\frac{\gamma^2 \mathbb{E}[X]}{2+2\Omega/\mathbb{E}[X]}}. \tag{5.2}$$

Let us explain how we can employ Lemma 5.1. The S_i will be specific paths connecting v and v'. R will be the vertex set of a random induced subgraph of Q_2^n. Clearly, $S_i \subset R$ means that the path is contained in the random graph Γ_n. Similarly, $\mathbb{P}(S_i \cup S_j \subset R)$ means that both, S_i and S_j are contained in Γ_n. We finally remark that for the indicator r.v. of the event $S_i \subset R$, denoted by X_{S_i}, $\mathbb{E}[X_{S_i}] = \mathbb{P}(S_i \subset R)$ and $\mathbb{E}[X_{S_i} X_{S_j}] = \mathbb{P}(S_i \cup S_j \subset R)$ hold.

Our next lemma is instrumental for the proof of Theorem 2.1 and provides an upper bound on the number of Q_2^n-paths between two given Q_2^n-vertices.

Lemma 5.2. *Let $d \in \mathbb{N}$, $d \geq 2$ and let v, v' be two Q_2^n-vertices where $d(v,v') = d$. Then any Q_2^n-path from v to v' has length $2\ell + d$ and there are at most*

$$\binom{2\ell+d}{\ell+d}\binom{\ell+d}{\ell} n^\ell\, \ell!\, d! \tag{5.3}$$

Q_2^n-paths from v to v' of length $2\ell + d$.

Proof. W.l.o.g. we can assume $v = (0,\ldots,0)$ and $v' = (x_i)_i$, where $x_i = 1$ for $1 \leq i \leq d$ and $x_i = 0$, otherwise. Each path of length m induces the family of steps $(\epsilon_s)_{1\leq s\leq m}$, where $\epsilon_s \in \{e_j \mid 1 \leq j \leq n\}$. Since each path ends at v', we have for fixed $1 \leq i \leq n$

$$\sum_{\{\epsilon_s \mid \epsilon_s = e_i\}} \epsilon_s = \begin{cases} 1 & \text{for } \ 1 \leq i \leq d \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

Hence the families induced by these paths contain necessarily the set $\{e_1,\ldots,e_d\}$. Let $(\epsilon'_s)_{1\leq s\leq m'}$ be the family obtained from $(\epsilon_s)_{1\leq s\leq m}$ by removing the steps $e_1,\ldots,e_d$, at the smallest index at which they occur. Then $(\epsilon'_s)_{1\leq s\leq m'}$ represents a cycle starting and ending at v. Furthermore, we have for all i; $\sum_{\{\epsilon'_s \mid \epsilon'_s = e_i\}} \epsilon'_s = 0$, i.e. all steps must come in up-step/down-step pairs. As a result we derive $m = 2\ell + d$ and there are exactly ℓ steps of the form e_j that can be freely chosen (free up-steps). We proceed by counting the number of $(2\ell + d)$-tuples $(\epsilon_s)_{1\leq s\leq 2\ell+d}$. There are exactly $\binom{2\ell+d}{\ell+d}$ ways to select the $(\ell + d)$ indices for the up-steps within the set of all $2\ell + d$ indices. Furthermore there are at most $\binom{\ell+d}{\ell}$ ways to select the positions for the ℓ up-steps and n^ℓ ways to choose the free up-steps themselves (once their positions are fixed). Since a free up-step is paired with a unique down-step reversing it, the ℓ free up-steps determine all ℓ down-steps. Clearly, there are at most $\ell!$ ways to assign the down steps to their ℓ indices. Finally, there are at most $d!$ ways to assign the fixed up-steps and the lemma follows. □

Proof of Theorem 2.1. Suppose $d = d(v,v')$ and $\Delta > 0$ are fixed. Let $Z = Z(d,\Delta)$ be the r.v. counting the paths of length $\leq \Delta\, d$ from v to v'. According to Lemma 5.2, we have

$$\mathbb{E}[Z] \leq \sum_{2\ell+d\leq \Delta\, d} \binom{2\ell+d}{\ell+d}\binom{\ell+d}{\ell} n^\ell\, \ell!\, d!\, \lambda_n^{2\ell+d-1}. \tag{5.5}$$

Since $\lambda_n < n^{\delta-\frac{1}{2}}$ for any $\delta > 0$, we obtain

(5.6)
$$\sum_{2\ell+d\leq\Delta\, d} \binom{2\ell+d}{\ell+d}\binom{\ell+d}{\ell}\, n^{\ell}\,\ell!\,d!\,\lambda_n^{2\ell+d-1} \leq \sum_{2\ell+d\leq\Delta\, d} \binom{2\ell+d}{\ell+d}\binom{\ell+d}{\ell}\,\ell!\,d!\; n^{\delta\,2\ell}\left[\frac{1}{n^{\frac{1}{2}-\delta}}\right]^{d-1}.$$

For given $d \geq 2$ and Δ, the quantity ℓ is bounded and choosing δ sufficiently small, we derive the upper bound

$$\mathbb{E}[Z] \leq O(n^{-\mu}) \quad \text{for some} \quad \mu > 0, \tag{5.7}$$

proving assertion (a). To prove (b) we consider a specific subset of paths, $\mathbb{A}_\sigma$, where σ is some permutation of $d-1$ elements. The $\mathbb{A}_\sigma$-elements are called α-paths and given by the following data:

(I) some family $(e_{j_1},\ldots,e_{j_\ell})$, where $d-1 < j_i \leq n$ and $|\{j_i \mid 1\leq i\leq \ell\}| = \ell$.

(II) the fixed family $(e_{\sigma(1)},\ldots,e_{\sigma(d-1)})$

(III) the family $(e_{j_\ell},\ldots,e_{j_1})$, i.e. the mirror image of the family chosen in (I).

Let X_α be the indicator r.v. for the event "α is a path in Γ_n". Clearly, $A = \sum_{\alpha\in\mathbb{A}_\sigma} X_\alpha$ is the r.v. counting the number of α-paths contained in Γ_n. Let $n' = n-(d-1)$. By construction of α-paths and linearity of expectation we observe

$$\mathbb{E}[A] = \ell!\binom{n'}{\ell}\lambda_n^{2\ell+(d-1)} = (n')_\ell\,\lambda_n^{2\ell+(d-1)}, \tag{5.8}$$

where $(n)_\ell = n(n-1)\cdots(n-(\ell-1))$. Since $\lambda_n \geq n^{-\frac{1}{2}+\delta}$ for some $0<\delta<\frac{1}{2}$

$$\mathbb{E}[A] \;\geq\; \left[\frac{(n'-\ell)}{n}\right]^{\ell} n^{2\ell\delta}\,\left[n^{-\frac{1}{2}+\delta}\right]^{d-1}. \tag{5.9}$$

The idea is now to use Janson's inequality (Lemma 5.1) in order to show that a.s. at least one α-path is contained in Γ_n. For this purpose we estimate the correlation between the indicator r.v. X_α and $X_{\alpha'}$. The key term we have to analyze is

$$\Omega = \sum_{\alpha\in\mathbb{A}_\sigma}\sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ \alpha'\cap\alpha\neq\varnothing}} \mathbb{E}[X_\alpha X_{\alpha'}].$$

Let $u_s = v + (\sum_{i=1}^{s} e_{j_i})$, where $s\leq\ell$. Since the sequence given in (III) represents the mirror-image of the sequence $(e_{j_1},\ldots,e_{j_\ell})$ we inspect

$$|\alpha\cap\alpha'| = 2\,|\{u_s\in\alpha\cap\alpha' \mid 1\leq s\leq\ell\}| + \begin{cases} d-1 & \text{if } u_\ell\in\alpha\cap\alpha' \\ 0 & \text{otherwise.}\end{cases} \tag{5.10}$$

Indeed, only if α and α' intersect at u_ℓ, the subsequent $(d-1)$ steps of (II) coincide. In view of eq. (5.10), we distinguish the cases

$$\text{(i)}\ \ u_\ell \notin \alpha \cap \alpha' \quad \text{and} \quad \text{(ii)}\ \ u_\ell \in \alpha \cap \alpha'\ . \tag{5.11}$$

Case (i): in this case we have $|\alpha \cap \alpha'| = 2h$, where $1 \leq h \leq \ell - 1$. For fixed h, there are exactly $\binom{\ell-1}{h}$ ways to select the h vertices where α and α' intersect. For each such selection, there at most $h!\,(n'-h)_{\ell-h}$ paths α', whence

$$|\{\alpha' \mid |\alpha' \cap \alpha| = 2h\}| \leq \binom{\ell-1}{h} h!\,(n'-h)_{\ell-h}. \tag{5.12}$$

The probability for choosing a correlated α'-path is given by $\lambda_n^{2[2\ell+(d-1)]-2h}$ and we compute

$$\begin{aligned}
\sum_{\alpha\in\mathbb{A}_\sigma} \sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ u_\ell\notin\alpha'\cap\alpha\neq\varnothing}} \mathbb{E}[X_\alpha X_{\alpha'}] &= \mathbb{E}[A] \sum_{h=1}^{\ell-1} |\{\alpha' \mid |\alpha'\cap\alpha| = 2h\}| \lambda_n^{[2\ell+(d-1)]-2h} \\
&\leq \mathbb{E}[A] \sum_{h=1}^{\ell-1} h! \binom{\ell-1}{h} (n'-h)_{\ell-h} \lambda_n^{[2\ell+(d-1)]-2h} \\
&= \mathbb{E}[A]^2 \sum_{h=1}^{\ell-1} h! \binom{\ell-1}{h} (n')_h^{-1}\, \lambda_n^{-2h} \\
&\leq \mathbb{E}[A]^2 \sum_{h=1}^{\ell-1} h! \binom{\ell-1}{h} \frac{n^h}{(n')_h}\, n^{-2h\delta},
\end{aligned}$$

where the last inequality is implied by $\lambda_n \geq n^{-\frac{1}{2}+\delta}$. We have for sufficiently large n

$$\sum_{h=1}^{\ell-1} h! \binom{\ell-1}{h} \frac{n^h}{(n')_h}\, n^{-2h\delta} = \underbrace{(\ell-1)\,\frac{n}{n'}\, n^{-2\delta}}_{h=1} + \underbrace{O\left(n^{-4\delta}\right)}_{h>1}. \tag{5.13}$$

Consequently, in case of (i), we can give the following upper bound :

$$\sum_{\alpha\in\mathbb{A}_\sigma} \sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ u_\ell\notin\alpha'\cap\alpha\neq\varnothing}} \mathbb{E}[X_\alpha X_{\alpha'}] \leq \left[(\ell-1)\,\frac{n}{n'}\, n^{-2\delta} + O(n^{-4\delta})\right] \mathbb{E}[A]^2. \tag{5.14}$$

Case (ii): the key observation is that for fixed α, there are at most $\ell!$ paths α' that intersect α at least in u_ℓ. Each of these appears with probability at most 1 whence

$$\sum_{\alpha\in\mathbb{A}_\sigma} \sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ u_\ell\in\alpha'\cap\alpha\neq\varnothing}} \mathbb{E}[X_\alpha X_{\alpha'}] \leq \ell!\, \mathbb{E}[A]. \tag{5.15}$$

Using eq. (5.14) and eq. (5.15), we arrive at

$$\Omega \;\leq\; \left(\underbrace{(\ell-1)\,\frac{n}{n'}\,n^{-2\delta}+O(n^{-4\delta})}_{\text{(i)}} + \underbrace{\frac{\ell!}{\mathbb{E}[A]}}_{\text{(ii)}} \right) \mathbb{E}[A]^2. \tag{5.16}$$

According to Lemma 5.1, we have $\mathbb{P}(A \leq (1-\gamma)\mathbb{E}[A]) \leq e^{-\frac{\gamma^2\mathbb{E}[A]}{2+2\Omega/\mathbb{E}[A]}}$, i.e.

$$\mathbb{P}(A \leq (1-\gamma)\mathbb{E}[A]) \leq \exp\left[-\frac{\gamma^2}{2/\mathbb{E}[A]+2\left((\ell-1)\,\frac{n}{n'}\,n^{-2\delta}+O(n^{-4\delta})+\frac{\ell!}{\mathbb{E}[A]}\right)} \right]. \tag{5.17}$$

In view of $\mathbb{E}[A] \geq \left[\frac{(n'-\ell)}{n}\right]^{\ell} n^{2\ell\delta} \left[n^{-\frac{1}{2}+\delta}\right]^{d-1}$, we observe, for sufficiently large ℓ,

$$\left[\frac{\gamma^2}{2/\mathbb{E}[A]+2\left((\ell-1)\,\frac{n}{n'}\,n^{-2\delta}+O(n^{-4\delta})+\frac{\ell!}{\mathbb{E}[A]}\right)} \right] = O(n^{2\delta}). \tag{5.18}$$

Setting $\gamma = 1$, eq. (5.17) becomes

$$\mathbb{P}(A=0) \leq e^{-c'n^{2\delta}} \qquad \text{for some } c'>0\ . \tag{5.19}$$

Since an α-path has length $2\ell+d$, eq. (5.19) proves (b) and the proof of the theorem is complete. □

Proof of Corollary 2.2. The expected number of α-paths is according to Theorem 2.1

$$\mathbb{E}[A] = (b-1)\,(n-(d-1))\,\lambda^{2+(d-1)} = (b-1)\,n'\,\lambda^{d+1}.$$

For $\ell = 1$ we have only type (ii) correlation, which is given via eq. (5.15). In this case any two correlated paths necessarily coincide, whence

$$\sum_{\alpha\in\mathbb{A}_\sigma} \sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ u_\ell\in\alpha'\cap\alpha\neq\varnothing}} \mathbb{E}[X_\alpha X_{\alpha'}] = (b-1)\,n'\,\lambda^{d+1}.$$

Consequently, eq. (5.17) becomes

$$\mathbb{P}(A=0) \leq \exp\left[-\frac{\mathbb{E}[A]}{4}\right] = \exp\left[-(b-1)\,(n-(d-1))\,\lambda^{d+1}/4\right],$$

whence the corollary. □

Proof of Corollary 2.3. We compute $\mathbb{E}[A] = n'(n'-1)\lambda^{4+(d-1)}$,

$$\sum_{\alpha\in\mathbb{A}_\sigma}\sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ u_2\notin\alpha'\cap\alpha\neq\varnothing}}\mathbb{E}[X_\alpha X_{\alpha'}] = \mathbb{E}[A]\,(n'-2)\lambda^{(d+3)-2} = \mathbb{E}[A]^2\,(n')^{-1}\underbrace{\left[\frac{n'-2}{n'-1}\right]}_{\mu(n')}\lambda^{-2}$$

$$\sum_{\alpha\in\mathbb{A}_\sigma}\sum_{\substack{\alpha'\in\mathbb{A}_\sigma;\\ u_2\in\alpha'\cap\alpha\neq\varnothing}}\mathbb{E}[X_\alpha X_{\alpha'}] = (1+\lambda^2)\,\mathbb{E}[A]\ .$$

Indeed, the second equality results from the following alternative: α' either intersects α at u_1 in which case they coincide or not. In the latter case we obtain the factor λ^2 since then u_1' and its mirror image for the down-step have also to be selected and there is exactly one choice to select u_1'. Therefore Lemma 5.1 implies

$$\mathbb{P}(A=0) \leq \exp\left[-\frac{1}{2/\mathbb{E}[A] + 2\,\{\,n'^{-1}\,\mu(n')\,\lambda^{-2} + (1+\lambda^2)/\mathbb{E}[A]\}}\right] \tag{5.20}$$

and Corollary 2.3 follows. □

Acknowledgments. We are grateful to Hillary S. Han, Emma Y. Jin, Fenix W.D. Huang, Jing Qin and Rita R. Wang for their help. Special thanks to Ivo L. Hofacker for providing the program for randomly generating canonical secondary structures and to Linda Y.M. Li for her help obtaining the RNA data. This work was supported by the 973 Project, the PCSIRT Project of the Ministry of Education, the Ministry of Science and Technology, and the National Science Foundation of China.

References

[1] Akutsu, T, 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, Discrete Appl. Math., 104, 45-62.

[2] Batey. R. T., Rambo. R. P., Doudna. J. A., 1999. Tertiary Motifs in RNA Structure and Folding Angew. Chem. Int. Ed., 38, 2326-2343.

[3] Balister P.N., Bollobás B., Frieze A.M., 2000. The first passage time diameter of the cube.

[4] Chen, W.Y.C., Qin. J., Reidys, C.M., 2008. Crossings and Nestings of tangled-diagrams, E.J. Combin. 15(86).

[5] Derrida, B., Peliti, L., 1999. Evolution in a flat fitness landscape, Bull. Math. Biol., 53, 355-382.

[6] Flamm, C., Fontana W., Hofacker, I.L., Schuster, P., 2000. RNA Folding at Elementary Step Resolution, RNA, 6, 325-338.

[7] Goebel, U., Forst, C.V., 2000. RNA Pathfinder–Global Properties of Neutral Networks, Zeitschrift fuer physikalische Chemie, 216.

[8] Goebel, U., 2000. Neutral Networks of Minimum Free Energy Structures, PhD-Thesis, University of Vienna.

[9] Grüner, W., Giegerich, R., Strothmann, D., Reidys, C.M., Weber, J., Hofacker, I.L., Stadler, P.F., Schuster, P., 1996. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration I. Neutral Networks, Chemical Monthly", 127, 355-374.

[10] Grüner, W., Giegerich, R., Strothmann, D., Reidys, C.M., Weber, J., Hofacker, I.L., Stadler, P.F., Schuster, P., 1996. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration II. Structures of Neutral Networks and Shape Space Covering, Chemical Monthly, 127, 375-389.

[11] Hofacker. I.L., Fontana. W., Stadler. P.F., Bonhoeffer. L.S., Tacker. M., Schuster. P., 1994. Fast Folding and Comparison of RNA Secondary Structures. Monatsh. Chem. 125: 167-188.

[12] Hofacker. I.L., Schuster. P., Stadler. P.F., 1998. Combinatorics of RNA Secondary Structures, Discr. Appl. Math., 88, 207-237.

[13] Howell. J.A., Smith. T.F., Waterman. M.S., 1980. Computation of generating functions for biological molecules. SIAM J. Appl. Math. 39, 119-133.

[14] Huang, W.D., Reidys, C.M., 2008. Statistics of canonical RNA Pseudoknot Structures, J. Theoret. Biol., doi:10.1016/j.jtbi.2008.04.002.

[15] Huang, W.D., Peng, W.D., Reidys, C.M., 2008. Folding RNA pseudoknot structures, in preparation.

[16] Janson S., 1990. Poisson approximation for large deviations, Random Structures and Algorithms 1, 221-229.

[17] Jin. E.Y., Qin. J., Reidys. C.M., 2008. Combinatorics of RNA structures with pseudoknots. Bull. Math. Biol. **70**(2008), 45-67.

[18] Jin. E.Y., Reidys. C.M, 2008. Asymptotic enumeration of RNA structures with pseudoknots. Bull. Math. Biol., 70: 951-970.

[19] Jin. E.Y. Reidys. C.M., 2008. Central and Local Limit Theorems for RNA Structures J. Theoret. Biol. **250**, 547-559.

[20] Jin. E.Y., Reidys, C.M. 2008, RNA-Lego: Combinatorial Design Of Pseudoknot RNA, Adv. Appl. Math., in press.

[21] Kimura, M., 1968. Evolutionary rate at the molecular level, Nature, 217, 624-626.

[22] Kimura, M., 1983. The Neutral Theory of Molecular Evolution, Cambridge Univ. Press, Cambridge, UK.

[23] Konings. D.A.M., Gutell. R.R., 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. RNA 1, 559-574.

[24] Loria. A., Pan. T., 1996. Domain structure of the ribozyme from eubacterial ribonuclease P. RNA 2, 551-563.

[25] Lyngso, R., Pedersen, C., 1996. Pseudoknots in RNA secondary structures, Physics of Biological Systems: From Molecules to Species, Springer, Berlin, Heidelberg, New York.

[26] R.C. Penner and M.S. Waterman, *Spaces of RNA secondary structures* Adv. Math. **101**(1993), 31-49.

[27] Reidys, C.M, Stadler, P.F., Schuster, P.K., 1997 Generic Properties of Combinatory Maps and Neutral Networks of RNA Secondary Structures, Bull. Math. Biol., 59(2), 339-397.

[28] Reidys, C.M., 1997. Random induced subgraphs of generalized n-cubes, Advances in Applied Mathematics, 19, 360-377.

[29] Reidys, C.M., 2002. Distance in Random induced subgraphs of generalized n-cubes, Combinatorics, Probability and Computing, 11, 599-605.

[30] Reidys, C.M., 2008. Large components in random induced subgraphs of n-cubes, Discrete Mathematics, accepted, arXiv:0704.2868

[31] Rivas. E., Eddy. S., 1999. A Dynamic Programming Algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 285, 2053-2068.

[32] Rivas, E., Eddy, S.R., 2000. The language of RNA: a formal grammar that includes pseudoknots Bioinformatics **16**(4), 334-340.
[33] Searls, D.B., 2002. The language of genes. Nature bf 420, 211-217.
[34] 2005. Mapping RNA Form and Function. Science, 309(2), No.5740.
[35] Schultes, E.A., Bartels, P. B., 2000. One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds, Science, 289, 448-452.
[36] Schuster, P., 2002. A testable genotype-phenotype map: Modeling evolution of RNA molecules, Michael Laessig and Angelo Valeriani, editors, Springer, 56-83.
[37] Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L., 1994. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures, Proceedings: Biological Sciences,255(1344), 279-284.
[38] Haslinger, C., Stadler, P.F., 1999. RNA Structures with Pseudo-Knots, Bull.Math.Biol., 61, 437-467.
[39] Tacker, M., Stadler, P.F., Bauer, E.B., Hofacker, I.L., Schuster, P., 1996. Algorithm Independent Properties of RNA Secondary Structure Predictions, Eur.Biophy.J., 25, 115-130.
[40] Tuerk. C., MacDougal. S., Gold. L., 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. Proc. Natl. Acad. Sci. USA, 89:6988-6992.
[41] Uemura, Y., Hasegawa, A., Kobayashi, S., Yokomori, T., 1999. Tree adjoining grammars for RNA structure prediction, Theoret. Comput. Sci., 210, 277-303.
[42] Waterman. M.S., 1978. Secondary structure of single - stranded nucleic acids. Adv. Math.I (suppl.) 1, 167-212.
[43] Waterman. M.S., 1979. Combinatorics of RNA hairpins and cloverleafs. Stud. Appl. Math. 60, 91-96.
[44] Waterman. M.S., Schmitt, W.R., 1994. Linear trees and RNA secondary structure. Discr. Appl. Math 51, 317-323.
[45] Westhof. E., Jaeger. L., 1992. RNA pseudoknots. Current Opinion Struct. Biol. 2, 327-333.
[46] R. Wong and M. Wyman, *The method of Darboux* J. Approx. Theory. **10** (1974), 159-171.

Center for Combinatorics, Nankai University, Tianjin 300071, P.R. China, Phone: 86-22-2350-21**, Fax: 86-22-2350-9272

E-mail address: duck@santafe.edu