Bio-Molecular Shapes and Algebraic Structures

Christian Reidys[a,*] and Peter F. Stadler[b,c]

[a]Institut für Molekulare Biotechnologie, Jena, Germany

[b]Institut für Theoretische Chemie, Wien, Austria

[c]Santa Fe Institute, Santa Fe, USA

*Mailing Address:
Institut für Molekulare Biotechnologie
Beutenbergstraße 11, PF 100 813, D-07708 Jena, Germany
Phone: **49 (3641) 65 6458 Fax: **49 (3641) 65 6335
E-Mail: duck@imb-jena.de

Abstract

Shapes of biological macromolecules — RNA, DNA, and proteins — can be represented by abstract algebraic structures provided as suitably coarse resolution is chosen. These abstract structures, for instance partially ordered sets and permutation groups, can be used for deriving new metric distances between biomolecular shapes and for proving surprising theorems on sequence-structure relations.

1. Introduction

One of the central problems in contemporary molecular biology is the comparison of biomolecular structures. The most basic question in this context is *"What do we mean when we say that two structures are the same?"* We argue that even this question is far from being trivial. Clearly, no two biopolymers with different sequence (primary structure) can have identical spatial conformations at atomic resolution.

However, atomic resolution is *not* what is referred to when one says, for instance, that all tRNAs have the same shape. Of course, different levels of precision will be suitable for different problems. In the case of RNA the base pairing patterns provide a natural coarse graining of the structure. In case of proteins the situation is more involved: one might restrict attention to the C^{α} back-bone, or take the orientation of the side chains into consideration, maybe in a simplified form by using the C^{β} atoms in addition to the backbone [28]. A further coarse graining is obtained by embedding the structures into a regular lattice [4, 18]. An even cruder representation, which is closer to the situation for RNA, just recognizes contacts between amino-acids, and structure is represented by the so-called contact matrix [3]. It is this level of description which we will focus on in this paper.

Once identity of structures is defined, one can ask for a distance measure between non-identical structures. Again RNA is a tractable special case. Secondary structures can be represented as linear strings and hence a standard string alignment can be used to define a metric distance measure [12]. A more sophisticated approach converts the secondary structure into a tree, and distance is then defined via "tree editing" [26, 27, 8]. Feasible approaches exist also for the general matrix comparison problem. A heuristic approach termed "double dynamic programming" [30] is widely used for comparing protein structures. Alternatively one can define the similarity between two structures as the size of the largest common substructure allowing for (insertions and deletions). In the case of contact matrices this requires to find the largest common sub-matrix of two matrices when deleting rows and the corresponding columns are the allowed editing operations. An algorithm for this problem was published recently [20].

In this contribution we focus on some intriguing relations between biomolecular shapes and abstract algebraic structures. We will explain how these relations can be used for deriving distance measures between shapes. Furthermore we will see that the algebraic representations of the shapes provide a powerful means of gaining insight into global properties [25] of sequence-structure mappings.

We explicitly caution the reader that we do **not** claim that the algebraic structures presented here are *the* description for biomolecular shapes. We simply discuss some algebraic structures that we have found at least moderately useful for the description of biopolymers. We do not intend to present a consistent theory for the representation of biomolecular shapes, we merely discuss a set of ideas which, in our opinion, deserve further attention.

2. Structures as PO-Sets

2.1. Contact Structures

When restricting the 3D shape of a biopolymer to a contact matrix we obviously omit a wealth of structural information. On the other hand, contact structures capture the type of structural information that can be obtained by a variety of experimental and computational methods, including NMR [31], chemical and enzymatic reactivity [14], and phylogenetic analysis [16].

Definition 1. *A* contact structure *[of a linear polymer] is a vertex-labeled graph* $\Gamma = (V, E)$ *on* n *vertices with an adjacency matrix* A *fulfilling*
(i) $a_{i,i+1} = a_{i+1,i} = 1$ *for* $1 \leq i \leq n-1$.

The edges required by (i) are referred to as the *backbone* of the contact structure; we denote the matrix fulfilling (i) and having no other non-zero entry by B_n. Any other edge is a *contact*. The *contact matrix* C of the structure is $C = A - B_n$. A vertex i connected only to $i-1$ and $i+1$ will be called *isolated* or *unpaired*.

Definition 2. *On the set of contacts* $\mathcal{Q}$ *we define the relation* $\prec$ *by* $(k,l) \prec (i,j)$ *iff* $i < k < l < j$. *We will say that* (k,l) *is* interior *to* (i,j). *A base pair* (k,l) immediately interior *if there is no base pair* (p,q) *such that* $(k,l) \prec (p,q) \prec (i,j)$. *Furthermore we define* $\mathcal{Q}^* = \mathcal{Q} \cup \{*\}$ *and* $(i,j) \prec *$ *for all* $(i,j) \in \mathcal{Q}$.
Finally we define the relation $\preceq$ *by* $a \preceq b$ *if either* $a \prec b$ *or* $a = b$.

The element $*$ does not correspond to a contact. This virtual *root* is introduced for technical convenience [26]. For the convenience of the reader we repeat here the definition of partial orders and PO-sets:

Definition 3. *Let X be a non-empty set and let $\preceq$ be a relation on X such that for all $a, b, c \in X$ holds*

(i) $a \preceq a$;

(ii) $a \preceq b$ *and* $b \preceq a$ *implies* $a = b$;

(iii) $a \preceq b$ *and* $b \preceq c$ *implies* $a \preceq c$.

Furthermore we write $a \prec b$ if $a \preceq b$ and $a \neq b$. The relation $\preceq$ is called a partial order *on X, and X together with the relation $\preceq$ is called a* partially ordered set *or* PO-set.

Theorem 1. *$(\mathcal{Q}^*, \preceq)$ is a PO-set.*

Proof. First we observe that $(i,j) \prec (i,j)$ contradicts the definition of the relation $\prec$ for contacts. Hence $\preceq$ fulfills (i) and (ii) in the above definition. Transitivity, (iii), is obvious. ∎

Definition 4. *Suppose s is a contact structure. The we defined the* contact graph *$G[\mathcal{Q}^*] = (V_{\mathcal{Q}^*}, E_{\mathcal{Q}^*})$ associated with s by*

$$V_{\mathcal{Q}^*} := \mathcal{Q}^* \qquad \textit{and}$$

$$E_{\mathcal{Q}^*} := \{(x, x') \,|\, x, x' \in \mathcal{Q}^* : x \prec x' \textit{ and } x \textit{ is immediately interior to } x'\}.$$

The contact graphs provide a condensed and coarse grained representation: any information about building blocks without contacts is lost at this resolution, hence the contact graph is in general not sufficient for reconstructing the contact matrix. An example is shown in figure 1.

2.2. Secondary Structures

Nucleic acids, RNA and DNA, exhibit a special form of contact structures.

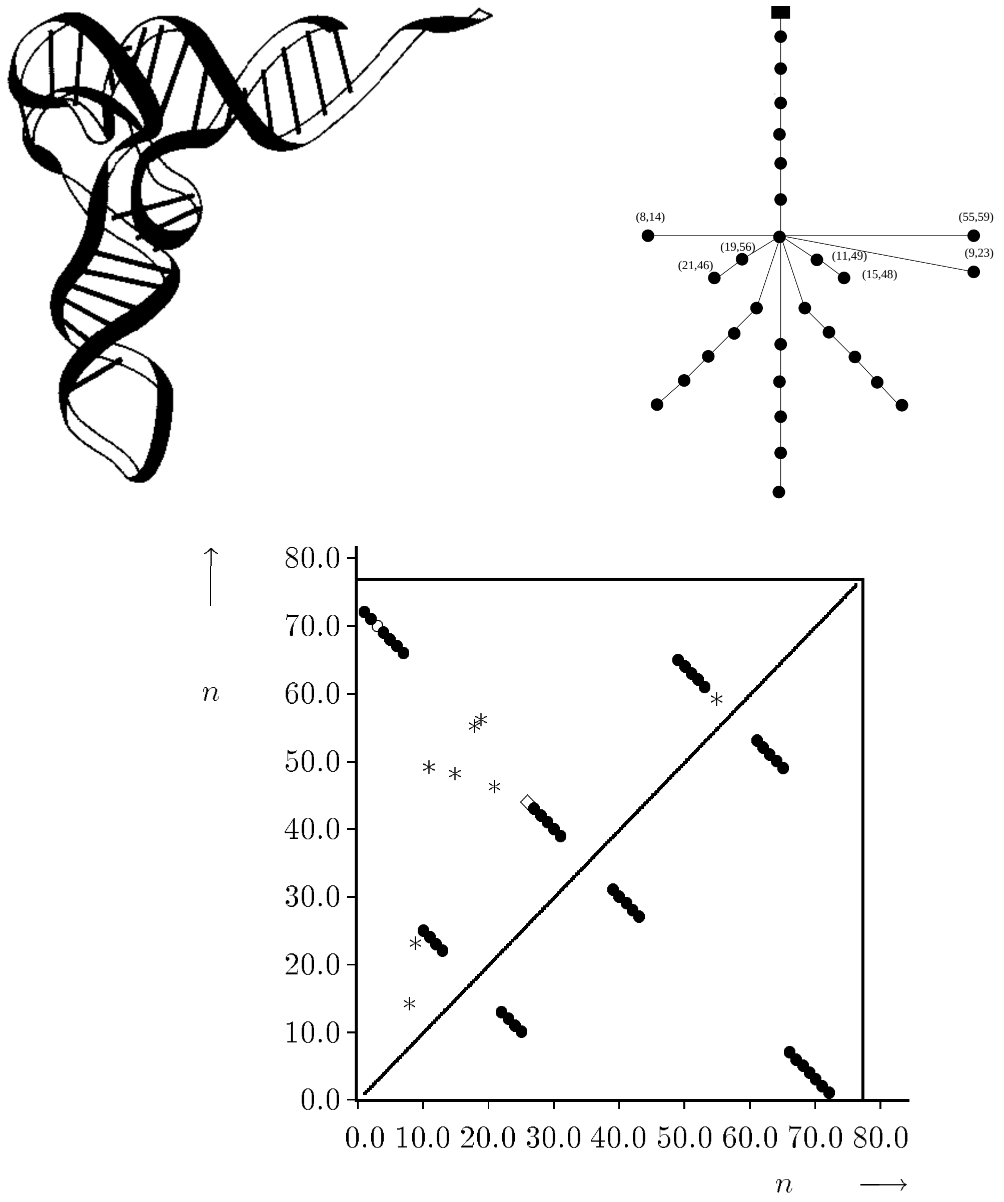

Figure 1: **a)** Three-dimensional structure of the phenylalanin tRNA of yeast.
b) The contact graph representing the PO-set of contacts of tRNA-phe from yeast.
c) Contact map of tRNA-phe.
The upper triangle shows the contact of the 3D-structure: $\bullet$ Watson-Crick base pairs, $\circ$ GU base pairs, $\diamond$ other non-Watson-Crick base pairs in double helical regions, and $*$ tertiary contacts between bases.
The lower triangle shows the secondary structure consisting of both Watson-Crick and GU base pairs, $\bullet$.

Definition 5. [32] *An RNA* secondary structure *is a contact structure of a linear polymer whose adjacency matrix fulfills additionally*

(ii) *For each i there is at most a single $k \neq i-1, i+1$ such that $a_{ik} = 1$;*

(iii) *If $a_{ij} = a_{kl} = 1$ and $i < k < j$ then $i < l < j$.*

In case of nucleic acid (secondary) structures we call an edge (i,k) with $|i-k| \neq 1$ a *bond* or *base pair* rather than a contact. If a contact structure fulfills condition (ii) we will say that it has unique bonds.

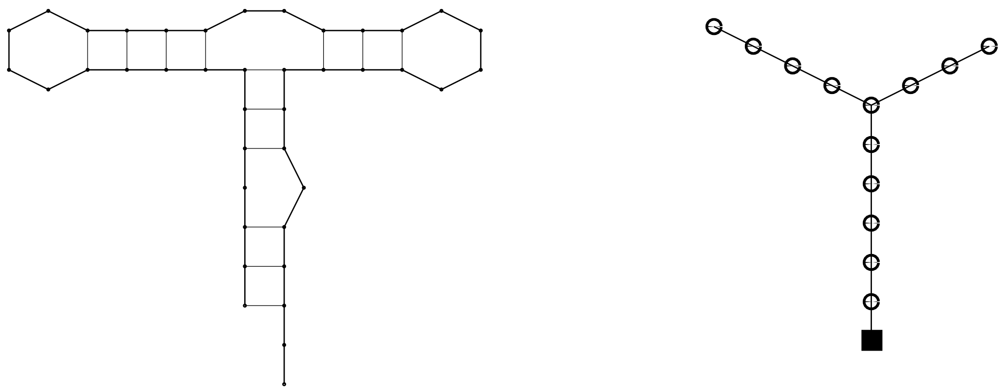

Figure 2: A secondary structure and its contact graph. By Lemma 1 it is a tree.

Lemma 1. *The contact graph of a secondary structure is a tree.*

Proof. We take a partition of $\mathcal{Q}$ as follows:
$\mathcal{Q} = \bigcup \Lambda_i$ where $\Lambda_0 := \{(*)\}$ and $\Lambda_i := \{x \in \mathcal{Q} \,|\, \forall x' \in \Lambda_{i-1} : x' \prec x\}$.
Writing the elements of Λ_i in columns we obtain a graph connecting the pairs x, x' for which holds $x \prec x'$. The resulting graph is in fact a *tree* since from the column corresponding to Λ_i there can be *some* $x' \in \Lambda_{i+1}$ such that $x \prec x'$, but there is at most one $x \in \Lambda_i$ connected to a fixed $y \in \Lambda_{i+1}$. ∎

The above result is intuitively obvious, and has been tacitly used for example in [26, 27, 8]. An example is shown in figure 2. The converse is not true in general. In fact the PO-set of the pseudo-knot in figure 3 is also a tree.

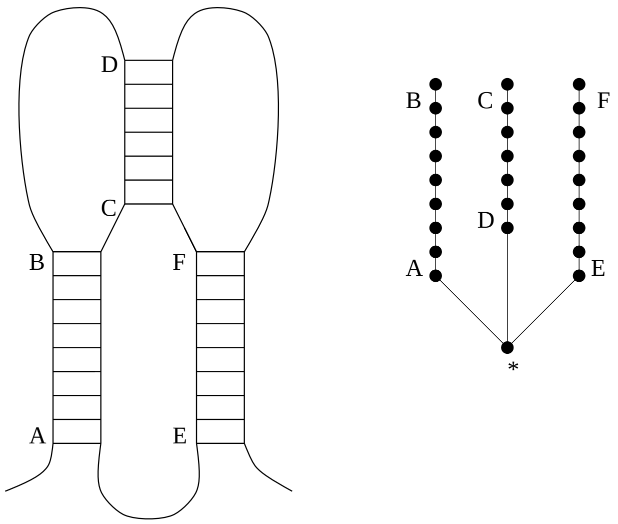

Figure 3: The PO-set of a pseudoknot is also a tree.

In the case of secondary structures the tree can be interpreted as planar. In fact, there is a second partial order of $\mathcal{Q}^*$ defined by $(i,j)\dot{\preceq}(k,l)$ if $i \leq k$ and $j \leq l$. For a secondary structure we have the following

Lemma 2. *Let $\mathcal{Q}^*$ be the contact set of a secondary structure. Then for any two contacts $p \neq q \in \mathcal{Q}^*$ are comparable either with respect to $\prec$ or with respect to $\dot{\prec}$.*

Proof. This is an immediate consequence of the "no-knot" condition (iii). ∎

In other words, given to contacts, we know that either one is interior to the other or one is to the right of the other. Of course this is not true for more general contact structures. As a consequence of lemma `rl-0` there is a well defined order $<$ on $\mathcal{Q}^*$ given by $x < y$ is $x \prec y$ or $x \dot{\prec} y$; this is crucial for the tree editing algorithms[29].

Remark. The construction of the corresponding tree in the above proposition does not lead to a one-to-one correspondence between secondary structures and trees. For this it is necessary to take additionally into account the explicit positions of the base pairs or equivalently the unpaired bases [26, 27, 8]. A related representation can be found in [12].

2.3. Loops

A contact structure can be decomposed into "*loops*":

Definition 6. *A loop of a contact structure is a subset $L \subseteq V$ such that the induced subgraph $\Gamma[L]$ is a circle.*

Loops are a means of describing the structure. In fact, the contact matrix A can be recovered if all loops and the back-bone is given. Let $\Gamma^* = \bigcup_{L<\Gamma} \Gamma[L]$, i.e., the graph obtained by superimposing all loops. That is, Γ^* is obtained from Γ by removing all edges that are not contained in a (non-trivial) loop. The connected components of Γ^* correspond to the components of the contact structure.

Loops play a special role for the secondary structures of nucleic acids. In this special we call an edge (i, k) with $|i - k| \neq 1$ a *bond* or *base pair* rather than a contact. A crucial feature of secondary structures is that there is a one-to-on correspondence between loops and base pairs: For each base pair (i, j) there a (uniquely defined) loop, $L_{(i,j)}$, consisting of the pair (i, j) itself, all base pairs immediately interior to (i, j) and all unpaired strands connecting these base pairs. Consequently, any two loops have at most two vertices in common. Loops of size $|L| = 4$ are referred to as stacks; they are the smallest non-trivial loops in "real" nucleic acids.

The free energy ΔG of a secondary structure is a linear function of (sequence dependent) energies for the loops; experimental energy parameters are available for the contribution of an individual loop as functions of its size, of the type of its delimiting base-pairs, and partly of the sequence of the unpaired strands. [10, 15]. Based on this energy model the folding problem can be solved by dynamic programming [32, 33, 37, 36]. Knots, pseudo-knots and triple-helices are considered as parts of the tertiary structure. There are two good reasons for these restrictions, i.e., for conditions (ii) and (iii) in the definition of secondary structures: firstly, they are necessary in order to allow for efficient folding algorithms, and secondly, no reliable experimental data are available for the free energy contributions of pseudo-knots or triple helices.

2.4. Coarse Graining

Representations of secondary structures in full resolution make it often difficult to focus on the major structural features of RNA molecules since they are often overloaded with irrelevant details. Coarse-grained tree representations were invented previously to solve this problem [26, 27]. They are based on plausible *ad hoc* assumptions. A more systematic procedure has been proposed in ref.[8]: Vertices of degree 2 are omitted, and their first descendent with degree not equal to 2 is assigned a weight corresponding to the number of omitted vertices. Using this procedure on obtains so-called homeomorphic-ally irreducible trees with weighted vertices. These weight can now be used to parameterize a coarse graining: vertices with small weights, i.e., small substructures, are simply deleted from the representation of the structure, which then reflects only structural elements larger than a certain threshold. This procedure can of course be generalized to arbitrary PO-sets.

An alternative coarse graining starts with the loop-decomposition. A simplified graph is obtained by replacing a loop by a simple vertex, and by joining two such vertices by an edge if and only their corresponding loops have at at least one vertex in common.

3. Permutations

3.1. Contact Structures

RNA secondary structures consist (in essence) of a single type of contacts, namely the Watson-Crick and the GU base pairs. In addition, spatial RNA structures contain a number of additional contacts which are usually considered as tertiary interactions. It makes perfect sense therefore to distinguish different classes of contacts, see figure 3. In general we might consider structures that are composed of r different kinds of contacts, and hence

$$\mathcal{Q} = \bigcup_{j=1}^{r} \mathcal{Q}_j .$$

This motivates the following

Definition 7. *Let $\mathcal{Q}$ be the set of contacts of a structure Γ and S_n be the symmetric group in n letters. Suppose $\mathcal{Q}$ can naturally be r-partitioned. Then we set*

$$\pi(s) \in S_n^r, \quad \pi(s) := \prod_{1 \le j \le r} \prod_{k=1}^{|\mathcal{Q}_j|} (i_k, j_k)_j \ ,$$

where $(i_k, j_k)_j$ is the k-th contact contained in $\mathcal{Q}_j$ interpreted as a transposition.

Recall that a contact structure s has unique bonds if (ii) in def. 5 is fulfilled.

Lemma 3. *Let $\mathcal{S}$ be a shape space consisting of contact structures with corresponding partition $(\mathcal{Q}_{1\le j \le r})$. Suppose a structure s has* unique *bonds in the corresponding elements of the partition $\mathcal{Q}_j$. Then $\pi : \mathcal{S} \hookrightarrow S_n^r$ is an embedding.*

Proof. Obvious. ∎

In more down-to-earth terms this lemma simply says that the permutation π uniquely defines the contact structure if condition (ii) of def. 5 holds for each class of contacts.

The mapping from sequences to their structures,

$$f : \mathcal{Q}_\kappa^n \to \mathcal{S}$$

which assigns a shape to each sequence in the sequence space[1] $\mathcal{Q}_\kappa^n$ is at the heart of theoretical molecular biology. The global properties of this mapping have been subject to detailed investigations in the case of RNA secondary structures [9, 25], and to some extent for lattice models of proteins [4]. An important question about this mapping is the structure of the preimages $f^{-1}(s)$, i.e., the geometry of the sets of sequences that fold into a common shape s. The mapping π imposes restrictions on the graph-theoretical structure of the preimage $f^{-1}(s)$ in sequence space [25, 24] and hence encapsulates a first set conditions in the context of inverse folding. More details will be discussed in the following section.

3.2. Secondary Structures

Magarshak and coworkers [22, 17] have proposed to represent RNA sequences as vectors of complex numbers using the correspondence

$$(\text{C,G, U/T, A}) \to (-1, 1, -i, i).$$

[1] The sequence space Q_κ^n consists of all sequences of length n over an alphabet with κ letters. It can be viewed as a graph by connecting sequences which differ in a single position by an edge. This induces the Hamming distance between sequences, whence these graphs are sometimes called Hamming graphs [7].

Furthermore they assign a signed permutation matrix to each secondary structure by means of

$$s_{ij} = \begin{cases} -1 & \text{if} \qquad\qquad (i,j) \in \mathcal{Q} \\ 1 & \text{if} \quad i=j \,\wedge\, i \,\text{is an unpaired base} \\ 0 & \qquad\qquad\quad \text{otherwise} \end{cases}$$

The most important consequences of this definition are the following:

(i) $S = S^{-1}$.

(ii) If a sequence x is compatible with a secondary structure S, then $Sx = x$.

(iii) For any two secondary structures S_1 and S_2 there is a transfer matrix $T_{12} := S_2 S_1^{-1} = S_2 S_1$. Transfer matrices are orthogonal, $T^{-1} = T^{+}$.

Unfortunately this formalism does not allow for GU pairs, and hence is restricted to a non-realistic definition of compatibility between sequences and structures. An extension of this formalism using quarternions instead of complex numbers is described in [21].

Remark. The definition of the transfer matrices gives rise to a metric distance measure between secondary structures. Let $\|\,.\,\|$ be an arbitrary matrix norm. Then

$$d(S_1, S_2) := \|T_{12}\|$$

is a metric on the set of all secondary structures of given length.

Let us now consider the unsigned permutations defined in the previous section. In the case of a single class of contacts the above definition reduces to

$$\pi : \mathcal{S} \to S_n,\;\; \pi(s) = \prod_{i=1}^{n_p(s)} (x_i, x_i'),$$

where (x_i, x_i') denotes a base pair.

Lemma 4. *Let s be a secondary structure. Then $\pi(s)$ is an involution, i.e. $\pi(s)\pi(s) = id$, where id denotes the identity permutation.*

Proof. This follows directly from the definition of π and the fact that all transpositions are disjoint as a consequence of property (ii) of secondary structures. ∎

3.3. Dihedral Groups and Involutions

Any two involutions $\imath, \imath'$ form a *dihedral group* $D_m := \langle (\imath, \imath') \rangle$. A basic result is the following

Lemma 5. *The dihedral group D_m is a semi-direct product*

$$D_m \cong \langle \imath \rangle \ltimes \langle \imath\imath' \rangle := \langle \imath \rangle \ltimes C_{\imath\imath'},$$

where $C_{\imath\imath'}$ is a cyclic subgroup of index 2, and hence normal, if and only if $\imath \neq \imath$.

Proof. See [19]. ∎

This lemma suggest to consider the following mapping in the case of secondary structures (where $\pi(s)$ is an involution):

$$\jmath : \mathcal{S} \times \mathcal{S} \to \{D_m < S_n\}, \ \jmath(s, s') = \langle \pi(s), \pi(s') \rangle$$

This construction has turned out to be useful for a variety of applications to RNA secondary structures which will be discussed in the following paragraphs.

3.3.1. Neutral Networks in Sequence Space

Following [24] to a RNA secondary structure s corresponds the *set of compatible sequences*, $\mathbf{C}[s]$. It consists of the sequences that *could* fold into s. As argued in [24] we can introduce a natural graph structure on this set by observing that its vertex set is isomorphic to the vertex set of the direct product of the two sequence spaces. Hence we construct the graph

$$\mathcal{C}[s] := \mathcal{Q}_{\kappa}^{n_u} \times \mathcal{Q}_{\beta}^{n_p} \, .$$

where κ is the number of different monomers, n_u is the number of unpaired positions in s, β is the number of different types of base pairs, and n_p is the number of base pairs in s. Now we define the *neutral network* $\Gamma[s]$ of a shape s as the induced subgraph of $f^{-1}(s)$ in $\mathcal{C}[s]$. It contains all sequences that fold into the structure s, and we consider two such sequences as neighbors if they differ either in a single unpaired base or in the type of single base pair.

Neutral networks are a very important feature, probably even *the* crucial property of sequence-structure mappings. For results on RNA see [25]. Computational studies are restricted to RNA since it is the only case where a reliable folding algorithm is available. Furthermore, an exhaustive search for all neutral networks, or an analysis of the components of a given neutral network, is feasible only for very small chain length [11]. We have developed an approach based on random graphs which circumvents this problem [24]. In the simplest case a vertex of $\mathcal{C}[s]$ is chosen to be part of given neutral network at random with probability λ. Using the dihedral group representation we can prove that to each pair of secondary structures s, s', there exists a sequence x that is compatible to both s and s'. This implies in turn that any two neutral networks come close to each other in sequence space.

This result has been extended by Jacqueline Weber [34] in the following form For any pair of secondary structures (s_1, s_2) there exists a pair of *complementary* sequences $(x, \bar{x})$ such that x is compatible to s_1, and $\bar{x}$ is compatible to s_2. This explains why we can find even very short RNA sequences that fulfill the rigid structural requirements posed by Qβ replicase for both the plus and the minus-strand [1]. The dihedral-group representation is also useful for explicitly constructing such pairs of sequences.

3.3.2. Transition between Neutral Networks

Let us now briefly consider a simple model for the evolution of an asexually replicating species. Suppose a finite population starts initially on the network $\Gamma[s]$. Then we ask for the probability of a transition to another network $\Gamma[s']$ which consists of fitter sequences. This may be the central question of evolutive optimization. It turns out that the topology of the so called intersection set $I[s, s']$ consisting of sequences that are compatible with *both* structures s and s' is of particular importance. In fact, the operation of $\langle \pi(s)\pi(s')\rangle$, and in particular the corresponding orbit decomposition, is closely related to the topology of the intersection set $I[s, s']$. For more details we have to refer to a forthcoming paper [35].

3.3.3. Metrics for Secondary Structures

The dihedral group representation can be used to obtain a new metric on the set of secondary structures. It is related to the transition probability between two neutral networks. Furthermore it might be useful for the extension of the concept of the error threshold in Eigen's quasispecies model [5, 6] to the level of phenotypes [13]. The goal in this context is to construct an analogue of the Hamming metric that is suitable for the transitions between structures [23]. The construction of this metric will be discussed in section 5.

4. Graphs, Groups, and Distances

4.1. Some Background

Definition 8. *A set G together with a binary operation $\circ : G \times G \to G$ is a* group *if for all $x, y, z \in G$ holds*

(i) $x \circ y \in G$ (closure)

(ii) $(x \circ y) \circ z = x \circ (y \circ z)$ (associativity)

(iii) *There is an $r \in G$ such that $x \circ e = e$ for all $x \in G$* (neutral element)

(iv) *For all $x \in G$ there is a x^{-1} such that $x \circ x^{-1} = e$* (inverse element)

In the following we drop the symbol $\circ$ and write simply xy for $x \circ y$.

Definition 9. *Let G be a group. A function $|.| : G \to \mathbb{R}_{0+}$ is called a* length function *on G if*

(i) $|x| = 0 \iff x = e$.

(ii) $|x| = |x^{-1}|$ *for all $x \in G$.*

(iii) $|xy| \le |x| + |y|$ *for all $x, y \in G$.*

Lemma 6. *$|.|$ is a length function on G if and only if $d(x, y) = |xy^{-1}|$ is a metric. For a given metric on G, $|x| = d(x, e)$ is a length function.*

Proof. It is obvious that (i) corresponds to the axiom $d(x, y) = 0 \iff x = y$ for metrics. Item (ii) is equivalent to symmetry: $d(x, y) = |xy^{-1}| = |yx^{-1}| = d(y, x)$ and (iii) is a consequence of the triangle inequality and symmetry. $d(x, y) = |xy^{-1}| \le d(x, e) + d(e, y) = |x| + |y^{-1}| = |x| + |y|$. ∎

Definition 10. *A subset $\Omega \subseteq G$ is called a set of* generators *of G if for all $x \in G$ there is a sequence $g_1, g_2, \ldots, g_n \in \Omega$ such that $x = g_n g_{n-1} \ldots g_2 g_1$. A set of generators Ω will be called* proper *if it fulfills*

(i) $e \notin \Omega$

(ii) $x \in \Omega \iff x^{-1} \in \Omega$

Definition 11. *A length function* $|.|$ *on* G *is called* consistent *with the set of generators* Ω *if*

(iv) $\quad |x| = 1 \quad$ *if* $\quad x \in \Omega$.

The set of all proper length functions on G *that are consistent with* Ω *will be denoted by* $\mathcal{L}(G, \Omega)$.

Definition 12. *Let* G *be a group, and let* Ω *be a proper set of generators of* G. *The graph* Γ *with vertex set* $V\Gamma = G$ *and edge set* $E\Gamma = \{(x, y) | xy^{-1} \in \Omega\}$ *is called a* Cayley *graph of* G. *It will be denoted by* $\Gamma(G, \Omega)$. *Furthermore, we will say that two elements* $x, y \in G$ *are* neighbors *(of each other) if the is a* $g \in \Omega$ *such that* $y = gx$.

Definition 13. *Let* Γ *be a graph*[2] *with vertex set* $V\Gamma$ *and edge set* $E\Gamma$. *A* path *of length* ℓ *on* Γ *is a sequence of vertices* $(x_1, x_2, \ldots, x_\ell)$ *such that* $(x_i, x_{i+1}) \in E\Gamma$ *for* $1 \le i < \ell$. *The (canonical)* distance $d_\Gamma(x, y)$ *of two arbitrary vertices in* $V\Gamma$ *is the length* ℓ *of the shortest path* $(x = x_1, x_2, \ldots, x_\ell = y)$ *connecting* x *and* y. *If such a path does not exist, we define* $d(x, y) = \infty$.

Definition 14. *A finite Graph* Γ *is* connected *if* $d(x, y) < \infty$ *for all* $x, y \in V\Gamma$. *The maximum distance*

$$\mathrm{diam}\Gamma = \max_{x,y \in V\Gamma} d(x, y)$$

is called the diameter *of the graph* Γ.

Theorem 2. *Let* $\Gamma(G, \Omega)$ *be a Cayley graph with canonical metric* $d_\Gamma(\,.\,,\,.\,)$, *and let* $\|\,.\,\|$ *be the extremal proper length function on* G *that is consistent with* Ω, *i.e.,*

$$\|x\| = \max_{|\,.\,| \in \mathcal{L}(G,\Omega)} |x| \qquad \text{for all } x \in G.$$

Then $d_\Gamma(x, y) = \|xy^{-1}\|$.

[2] Following Harrary's terminology we use the word *graph* only if there are not multiple edges or loops; more general "graphs" are referred to as *pseudo-graphs*[2].

Proof. By Lemma 1 the canonical metric is equivalent to some proper length function $|\,.\,|$. It is obvious that $|xy^{-1}|$ is simply the minimum number of generator necessary to construct xy^{-1} for all x, y. Thus $|\,.\,|$ must be consistent with the set of generators Ω. Now consider $|x|$. Let $g_i \in \Omega$ and $x = g_1 g_2 \dots g_n$. Then property (iii) of the length functions implies $|x| \leq \sum_{i=1}^{n} |g_i|$, and by property (iv) we obtain $|x| \leq n$, where n can of course by chosen to be the minimum number of generators necessary to construct x. It remains to show that the minimum number of generators $\|x\|$ is in fact a length function in $\mathcal{L}(G, \Omega)$; this is obvious since $\|x\| = d_\Gamma(x, e)$ is derived from a metric and fulfills (iv) by construction. ∎

Remark. Whenever we can write structures as elements of a group G we can define a canonical distance between them as $d(x, y) = \ell(xy^{-1})$ where ℓ is some length function on G.

Theorem 3. *The set T of all* transpositions *generates S_n. The length function associated with this set of generators is*

$$\ell(\pi) = n - \Theta(\pi), \qquad \pi \in S_n,$$

where $\Theta(\pi)$ is the number of cycles into which π decomposes.

Proof. This result is well known. We give a proof here for illustrative purposes.
(i) We show first that the minimum number of transpositions is in fact a length function on S_n. We proceed by induction on $\ell(y)$. Assume y is a transposition then we have $\ell(x) - 1 \leq \ell(x\,y) \leq \ell(x) + 1$. $\ell(x\,y) \leq \ell(x) + 1$ holds by definition and the assumption $\ell(x\,y) < \ell(x) - 1$ leads together with the first inequality to the contradiction $\ell(x) \leq \ell(x\,y) + 1 < \ell(x)$. Finally we assume the inequality holds for $\ell(y) = k - 1$. Then for any element y' with $\ell(y') = k$ there exists a transposition τ' such that $\ell(y'\tau') = k - 1$. Applying the induction hypothesis we obtain $\ell(x\,y'\tau') \leq \ell(x) + \ell(y') - 1$ whence $\ell(x\,y'\tau') + 1 \leq \ell(x) + \ell(y')$. It remains to observe $\ell(x\,y') \leq \ell(x\,y'\tau') + 1$ and the claim follows by the induction principle.
(ii) Each permutation can be written uniquely in a product of pairwise disjoint

cycles i.e. $\pi = \prod_{j=1}^{k} Z_j^{k_j}$ where k_j is the number of cycles of length j. This representation results from the action of the cyclic group $\langle\pi\rangle$ on the set of all positions of the string. Each cycle Z_j of length j can be written uniquely as a product of $j-1$ transpositions. Therefore we obtain

$$\ell(\pi) \leq \sum_j k_j\, j - \sum_j k_j$$

and it remains to show equality since $\sum_j k_j\, j - \sum_j k_j = n - \Theta(\pi)$. It is straightforward to prove by induction on j that every cycle Z_j requires at least $j-1$ different transpositions for its representation and the theorem is proved. ∎

4.2. A Pseudo-Metric for Contact Structures

Lemma 7. *Let* $\xi = (\xi_1, \ldots, \xi_r)$, $\eta = (\eta_1, \ldots, \eta_r)$, $\zeta = (\zeta_1, \ldots, \zeta_r) \in S_n^r$. *Then*

(i) $L(\zeta) := \sum_{j=1}^{r} \ell(\zeta_j)$ *is a length function on* S_n^r *whenever* ℓ *is length function on* S_n;

(ii) $D(\xi,\eta) := \sum_{j=1}^{r} d(\xi_j,\eta_j)$ *is a metric on* S_n^r *whenever* d *is a metric on* S_n.

(iii) D *is the metric associated with the length function* L *if* d *is the metric associated with* ℓ *on* S_n.

Proof. This follows directly from the properties of the direct product. ∎

If the mapping $\pi : \mathcal{S} \to S_n^r;\ s \mapsto \pi(s)$ is injective, then D is metric for the structures, defined by $D(\pi(s), \pi(s'))$. Otherwise we have only a pseudo-metric, since $\pi(s) = \pi(s')$ does not in general imply $s = s'$.

The representation of a contact structure as an element of S_n^r defined in the previous section has made use explicitly of a representation in terms of transpositions. Hence the appropriate choice for ℓ is the length function defined by the transpositions.

Corollary 1. *Let s and s' be two secondary structures of length n, and let $\pi(s)$ and $\pi(s')$ be their representations as permutations. Then*

$$d(s,s') := \ell(\pi(s)\pi(s')^{-1}) = n - \Theta(\pi(s)\pi(s')^{-1})$$

is a metric distance.

Proof. We know that the mapping $\pi : \mathcal{S} \to S_n$ is injective c.p. lemma 3, and that ℓ is a length function on S_n. ∎

5. Structures as Subgroups

5.1. Secondary Structures and Dihedral Groups

We are now in the position to present the metric based on the mapping $\jmath : \mathcal{S}\times\mathcal{S} \to D_m$, which we have promised in subsection 3.3.3.

Theorem 4. *For any pair $s, s' \in \mathcal{S}$ we construct the cyclic group $C_{ss'} := \langle\pi(s)\pi(s')\rangle$ which operates on the set of sequence positions $\{x_1, ..., x_n\}$. Let $\Psi(C_{ss'})$ denote the number of orbits induced by this operation. Then the mapping $(s, s') \to n - \Psi(C_{ss'})$ is a metric on the set of secondary structures $\mathcal{S}$.*

Proof. We have to show that $\Psi(C_{ss'}) = \Theta(\pi(s)\pi(s')^{-1})$ for all $s, s' \in \mathcal{S}$ which is exactly the content of corollary 1. ∎

5.2. Secondary Structures as Subgroups

In this section we propose another possibility of representing the base pairing information of a secondary structure.

Definition 15. *Let s be a secondary structure with base pairs*

$$\mathcal{Q} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)\}.$$

Let $T(s) = \{(x_i, y_i) \in S_n \,|\, i = 1, \ldots, p\}$ be set of transpositions corresponding to the base pair. Then $S(s) = \langle T(s)\rangle$, the permutation group generated by $T(s)$ is the (permutation) group of the the secondary structure.

For a finite group G we denote by $\Sigma(G)$ the set of all subgroups.

Lemma 8. *The mapping $T : \mathcal{S} \to \Sigma(S_n)$ an embedding.*

Proof. Since each base is contained in at most one base pair, the transpositions belonging to one structure are disjoint, and hence commute. Obviously now, two different structures have different base pairs and hence induce different permutation groups. ∎

Definition 16. *Let G be a finite group. For any two subgroups S and S' of G define*

$$\psi(S, S') := \ln\,[\,SS' \,:\, S \cap S'\,]\;.$$

The following proposition shows that $\psi(\,,\,)$ serves as a metric on the set of subgroups in general. In particular we have then a new matrix on the set of secondary structures.

Theorem 5. *Let G be a finite group. Then $\psi : \Sigma(G) \times \Sigma(G) \to \mathbb{R}$ is a metric on $\Sigma(G)$.*

Proof. (i) Symmetry is trivial. (ii) Clearly $[\,SS' \,:\, S \cap S'\,] \geq 1$, and this expression can be 1 only if $S = S'$. (iii) We will show that

$$[\,SS'' \,:\, S \cap S''\,]\;[\,S''S' \,:\, S'' \cap S'\,] \geq [\,SS' \,:\, S \cap S'\,]$$

This is equivalent to

$$\frac{|S|\,|S''|}{|S \cap S''|^2}\,\frac{|S'|\,|S''|}{|S' \cap S''|^2} \geq \frac{|S'|\,|S|}{|S' \cap S|^2}$$
$$|S''|\,|S \cap S'| \geq |S'' \cap S|\,|S'' \cap S'|$$

Since $S \cap S' \cap S''$ is a subgroup of S, S', and S'', we may rewrite this as

$$|S \cap S' \cap S''|\,|S''(S' \cap S)| \geq |(S \cap S'')(S' \cap S'')||(S \cap S'') \cap (S' \cap S'')|$$
$$|S''(S' \cap S)| \geq |(S \cap S'')(S' \cap S'')|$$

The latter inequality is always true since both $S \cap S''$ and $S' \cap S''$ are subgroups of S'' and hence their product is still contained in S''. ∎

6. Discussion

We have seen that biomolecular shapes and abstract algebraic structures are linked by a wealth of sometimes surprising relationships. Some of them are probably accidental, others might eventually lead to new insights into the architecture and dynamics of biopolymers. Having presented our approach we are left with far more open questions than results. We conclude this paper by list of some research questions along the lines of this contribution.

- Are there any relation between the three distance measures defined for the secondary structures? Is any of them related to the more common tree-editing metric for RNA?
- Is there any hope for extending or altering any of the above concepts in order to incorporate variable sizes of structures?
- Is there a way of characterizing secondary structures in terms relations defined on the set of contacts? As a suitable partial order sufficient, or do we need stronger structures?
- Is there a simple characterization of contact structures with unique contacts in terms of their contact graphs?
- What if contact are not unique as in the case of proteins? Can we still find a embedding into group, maybe larger than S_n^r?.
- Is there a framework in which Magarshak's ideas can be extended such as to allow for a more general logic of base pairing?

This list is far from being exhaustive.

Acknowledgments

We are grateful for stimulating discussions with Jacqueline Weber, Christian Forst, and Peter Schuster. Discussions with the participants of the 4th International Workshop on *Open Problems in Computational Molecular Biology* in Telluride prompted us to prepare this report.

References

[1] C. Biebricher, M. Eigen, and W. C. Gardiner, jr. Quantitative analysis of selection and mutation in self-replicating RNA. In L. Peliti, editor, *Biologically Inspired Physics*, pages 317–337, New York, 1991. Plenum Press.

[2] F. Buckley and F. Harrary. *Distances in Graphs.* Addison-Wesley, Reading, Ma., 1990.

[3] H. S. Chan and K. A. Dill. Compact polymers. *Macromolecules*, 22:4559–4573, 1989.

[4] H. S. Chan and K. A. Dill. "Sequence space soup" of proteins and copolymers. *J. Chem. Phys.*, 95:3775–3787, 1991.

[5] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 10:465–523, 1971.

[6] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv.Chem.Phys.*, 75:149 – 263, 1989.

[7] M. Eigen and P. Schuster. *The Hypercycle: a principle of natural self-organization.* Springer-Verlag, Berlin, 1979.

[8] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[9] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys.Rev.E*, 47:2083 – 2099, 1993.

[10] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, 1986.

[11] W. Grüner. *Evolutionary Optimization on RNA folding landscapes.* PhD thesis, University of Vienna, 1994.

[12] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucleic acids research*, 12:67–74, 1984.

[13] M. A. Huynen, P. F. Stadler, and W. Fontana. Evolution of rna and the Neutral Theory. Submitted to Nature, 1994.

[14] J. A. Jaeger, J. SantaLucia, and I. Tinoco. Determination of RNA structure and thermodynamics. *Annu.Rev.Biochem.*, 62:255–287, 1983.

[15] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc.Natl.Acad.Sci., USA*, 86:7706–7710, 1989.

[16] B. James, G. Olsen, and N. Pace. Phylogenetic comparative analysis of rna secondary structure. *Meth.Enzymol.*, 180:227–239, 1989.

[17] A. Kister, Y. Magarshak, and J. Malinsky. The theoretical analysis of the process of RNA molecule self-assembly. *BioSystems*, 30:31–48, 1993.

[18] A. Kolinski and J. Skolnick. Monte Carlo simulations of protein folding. I. Lattice and interaction scheme. *Proteins*, 18:338–352, 1994.

[19] H. Kurzweil. *Endliche Gruppen.* Springer-Verlag, Berlin, Heidelberg, 1977.

[20] A. M. Lesk. Boolean programming formulation of some pattern matching problems in molecular biology. *J.Chem.Soc.Faraday.Trans.*, 89:2603–2607, 1993.

[21] Y. Magarshak. Quarternion representation of RNA sequences and tertiary structures. *BioSystems*, 30:21–29, 1993.

[22] Y. Magarshak and C. J. Benham. An algebraic representation of RNA secondary structure. *J. Biomol. Struct. & Dyn.*, 10:465 – 488, 1992.

[23] C. Reidys and C. Forst. Replication on neutral networks in rna induced by rna secondary structures. Preprint, 1994.

[24] C. Reidys, P. Schuster, and P. F. Stadler. Generic properties of combinatory maps and application on rna secondary structures. Preprint, 1994.

[25] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.Lond.B*, 255:279–284, 1994.

[26] B. A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *CABIOS*, 4:387–393, 1988.

[27] B. A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[28] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force — An approach to the knowledge-based prediction of local structures in globular proteins. *J.Mol.Biol.*, 213:859–883, 1990.

[29] K. Tai. The tree-to-tree correction problem. *J. ACM*, 26:422–433, 1979.

[30] W. R. Taylor and C. A. Orengo. Protein structure alignment. *J.Mol.Biol.*, 208:1–22, 1989.

[31] G. Varani and I. Tinoco. RNA structure and NMR spectroscopy. *Quart. Rev. Biophys.*, 24:479–532, 1991.

[32] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv.Math. Suppl. Studies*, 1:167 – 212, 1978.

[33] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math.Biosc.*, 42:257–266, 1978.

[34] J. Weber. Appliction of the intersection theorem on pairs of complementary sequences. Personal Communication, 1994.

[35] J. Weber, C. Reidys, and P. Schuster. Transitions between neutral networks. Preprint, 1995.

[36] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull.Math.Biol.*, 46:591–621, 1984.

[37] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl.Acids Res.*, 9:133–148, 1981.

Table of Contents