Cycles in Secondary Structure Graphs

PETER F. STADLER

Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 40480 665 Fax: **43 1 40480 660 E-Mail: studla@tbi.univie.ac.at

The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
Phone: (505) 984 8800 Fax: (505) 982 0565 E-Mail: stadler@santafe.edu

URL: http://www.tbi.univie.ac.at/~studla

Abstract

Secondary structure graphs are a class of outer-planar graphs that are used in molecular biology as representations of the spatial structure of nucleic acid molecules. The elementary cycles of secondary structure graphs can be characterized in terms of a partial order on a subset of edges. It is proved that the the minimum cycle basis of secondary structure graph is unique.

Keywords

Minimum Cycle Basis — outer-planar graph — RNA secondary structure — elementary cycles

AMS Subject Classification: 05C38 — 92D20

Introduction

The most important classes of biopolymers, namely RNA, DNA, and proteins, are linear chains of monomers. The pairs of monomers that are adjacent on the chain form the backbone T of the molecule. Biopolymers form well-defined three dimensional structures which are of utmost importance for their biological function. These structures can be approximated by the list of so-called contacts, that is, by the set B of all pairs of monomers that are spatial neighbors without being adjacent on the backbone. While this simplification of the 3D shape obviously neglects a wealth of structural details, it captures the type of structural information that can be obtained by a variety of experimental and computational methods, including NMR [16], chemical and enzymatic reactivity [9], and phylogenetic analysis [6].

The *contact structure* of a biopolymer is therefore the graph $\Gamma = (V, E)$ where V is the set of monomers and $E = T \cup B$. It will be convenient to set $V = \{1, 2, \ldots, n\}$. The backbone $T = \{(i, i+1) | 1 \leq i < n\}$ is of course a spanning tree of Γ. Nucleic acids, both RNA and DNA, form a special type of contact structures known as *secondary structures.* A contact structure of a linear polymer is a secondary structure if its contacts satisfy
(i) For each $i \in V$ there is at most one contact $\alpha \in B$ incident with i.
(ii) If $(i, j), (k, l) \in B$ and $i < k < j$ then $i < l < j$.
The contacts in nucleic acids are usually called *bonds* or *base pairs.*

A graph Γ is *outer-planar* [2] if it can be embedded in the plane such that all vertices lie on the boundary of its exterior region. Secondary structures form a special class of outer-planar graphs [13] which generalize Yan's [21] "bamboo shoot graphs". Cycles of planar graphs are considered in [14].

A particular type of cycles, which is commonly termed *loops* in the RNA literature, plays an important role for RNA (and DNA) secondary structures: the energy of a secondary structure can be computed as the sum of energy contributions of the *loops*. Experimental energy parameters are available for the contribution of an individual *loop* as a function of its size, of the type of bonds that are contained in it, and on the monomers (nucleotides) that it is composed of [5, 10, 7, 18]. Based on this energy model one can compute the most stable secondary structure given the sequence of nucleotides in cubic time using a dynamic programming technique [19, 20, 23, 22]. We shall show here that the *loops* form the unique minimum cycle basis of the secondary structure.

Preliminaries

The set $\mathcal{E}$ of all subsets of E forms an m-dimensional vector space over GF(2) with vector addition $X \oplus Y = (X \cup Y) \setminus (X \cap Y)$ and scalar multiplication $1 \cdot X = X$, $0 \cdot X = \emptyset$ for all $X, Y \in \mathcal{E}$. A *cycle* is a subgraph such that any vertex degree is even. We represent a cycle by its edge set C. The set $\mathcal{C}$ of all cycles forms a subspace of $(\mathcal{E}, \oplus, \cdot)$. An *elementary cycle* is a minimal subgraph such that every vertex has degree 2. Suppose T is a spanning tree of Γ. Then for each edge $e \notin T$ there is unique elementary cycle in $T \cup \{e\}$ which is called a *fundamental cycle.* The set of fundamental cycles belonging to a given spanning tree form a basis of the cycle subspace which is called a *fundamental basis.* For details see [3, 15]. The length of a cycle basis is the sum of the lengths of all its cycles. A *minimum cycle basis* is a cycle basis with minimum length.

A contact structure with b contacts has $m = n - 1 + b$ edges. Its cyclomatic number, i.e., the dimension of the *cycle subspace* $\mathcal{C}$, is therefore $b = |B|$, the number of contacts. The backbone T is the most natural spanning tree of Γ. Denote by F_α the fundamental cycle in $T \cup \{\alpha\}$. Thus $\{F_\alpha | \alpha \in B\}$ is a *fundamental basis* of the cycle subspace $\mathcal{C}$. For each subset $U \subseteq B$ we define

$$\overline{U} = \bigoplus\nolimits_{\alpha \in U} F_\alpha \,. \tag{1}$$

All 2^b cycles are of this form. Given a cycle $\overline{U}$ its set of bonds is simply recovered: $U = \overline{U} \cap B$. Of course not all these cycles are elementary.

Remark. In nature there are not only linear RNA and DNA molecules but also nucleic acids with a circular backbone. All results of the presents contribution hold for this case as well. To see this we simply treat the additional backbone edge $\omega = (n, 1)$ as a contact, i.e., we set $\omega \in B$. Since the numbering in a circular nucleic structure can always be chosen such that n and 1 are both not part of a contact (as a result of steric constraints on the molecule [19]) we can interpret the graph Γ also as a contact structure with open backbone $T \setminus \{\omega\}$ and contact set $B \cup \{\omega\}$.

Elementary Cycles in Secondary Structures

In this section we shall characterize the set of elementary cycle of secondary structure graphs in terms of a partial order $\prec$ on the set of bonds B.

Definition. Let $(i, j), (k, l) \in B$. Then $(k, l) \prec (i, j)$ iff $i < k < l < j$. We will say that (k, l) is interior to (i, j). A bond (k, l) is immediately interior to (i, j) if there is no base pair (p, q) such that $(k, l) \prec (p, q) \prec (i, j)$. Two bonds are *$\prec$-comparable* if $\alpha \prec \beta$, $\beta \prec \alpha$ or $\alpha = \beta$. We write $\alpha \preceq \beta$ if either $\alpha \prec \beta$ or $\alpha = \beta$.

It is easy to see that $\prec$ is a partial order on B [12]. In the same paper it is shown that a contact structure is a secondary structure if any two bonds (i, j) and (k, l) satisfy either one of the following three conditions: $i, j < k, l$, or $k, l < i, j$, or (i, j) and (k, l) are $\prec$-comparable. (In the circular backbone case we have $\alpha \preceq \omega$ for all $\alpha \in B$.)

Lemma 1. If $\overline{Q}$ is an elementary cycle then Q has a unique upper bound, i.e., there is a unique element $\mu \in Q$ such that $\alpha \prec \mu$ for all $\alpha \in Q \setminus \{\mu\}$.

Proof. Let U be the set of all upper bounds in Q and let $W_v = \{\alpha \in Q | \alpha \preceq v\}$ for each $v \in U$. Obviously $Q = \emptyset$ does not give rise to a cycle, thus U is non-empty.
Claim: $\{U_v | v \in U\}$ is a partition of Q. The elements of U are by definition not $\prec$-comparable. In a secondary structure graph they can therefore be ordered from left to right: Set $(v_k^b, v_k^e) = v_k \in U$. Then $v_1^b < v_1^e < v_2^b < v_2^e < \ldots < v_k^b < v_k^e$. If $\alpha \in W_v$ then $v^b \le \alpha^b < \alpha^e \le v^e$ and hence $W_v \cap W_{v'} = \emptyset$ whenever $v \neq v'$. Clearly $\bigcup_{v \in U} W_v = Q$, and the claim is proved.
Therefore $\overline{Q} = \bigoplus_{v \in U} \overline{W_v}$. The edge sets $\overline{W_v}$ cannot contain common vertices in a secondary structure graph, thus $\overline{Q}$ is disconnected if U contains more than one element. ∎

Remark. The subgraphs spanned by the subsets W_v with $Q = B$ are known as the *components* of a secondary structure. The *components* are 2-connected while a secondary itself is in general only 1-connected. In the following the term component will however have is usual graph theoretic meaning.

Lemma 2. Let $S = \{\alpha, \beta, \gamma\}$ be a three-element subset of $Q \subseteq B$, such that β is immediately interior of α and γ is immediately interior of β with respect to Q. Then $\overline{Q}$ is disconnected.

Proof. It is obvious that $\overline{S}$ consists of two disjoint elementary cycles, with α and β in one component and γ in the other component. Now consider a $\delta \in Q \setminus S$. Then $\overline{S \cup \{\delta\}}$ is one of the following: If (i) δ is not $\prec$-comparable with any (and thus all) of the bonds in S then δ is part of an elementary cycle that contains no element of S, i.e., $\overline{S \cup \{\delta\}}$ is the disjoint union of three elementary cycles. If (ii) $\alpha \prec \delta$ then δ and α are contained in one component and β and γ are in another component of $\overline{S \cup \{\delta\}}$. If (iii) $\delta \prec \gamma$ then α and β remain in the same component and both δ and γ are contained in another component. Repeating this argument shows that α, β, and γ can never be combined in the same component of $\overline{Q}$. ∎

Lemma 3. If Q has a unique upper bound and there are no three bonds in Q satisfying $\alpha \prec \beta \prec \gamma$ then $\overline{Q}$ is an elementary cycle.

Proof. Let μ be the unique upper bound of Q. Of course, $\overline{\{\mu\}} = F_\mu$ is an elementary cycle. Since no two other bonds are $\prec$-comparable, these bonds (denoted by β_i, $i = 1, \ldots, k$) are arranged from left to right along the part $T \cap F_\mu$ of the back-bone: $\beta_i^b < \beta_i^e < \beta_{i+1}^b < \beta_{i+1}^e$ where $\beta_i = (\beta_i^b, \beta_i^e)$. The cycle $B_1 = \overline{F_\mu} \oplus F_{\beta_1}$ is obviously elementary and and the bonds $\beta_2, \ldots, \beta_k$ are arranged from left to right along the rest of the backbone. Repeating the argument for $B_{i+1} = B_i \oplus F_{\beta_{i+1}}$ shows that $B_k = \overline{Q}$ is indeed an elementary cycle. ∎

Summarizing the above three lemmata we have

Theorem 1. $\overline{Q}$ is an elementary cycle in a secondary structure if and only if Q has a unique upper bound and there are no three bonds in Q satisfying $\alpha \prec \beta \prec \gamma$.

Remark. The number of elementary cycles may vary between b (if no two bonds are $\prec$-comparable) and 2^{b-1} if $\overline{B}$, and hence each of the 2^{b-1} subsets of B that contain the maximal element of B, is an elementary cycle.

The Minimal Cycle Basis of a Secondary Structure is Unique

Horton proposed an algorithm for finding a minimum cycle basis in polynomial time which is based on the following

Proposition [8]. If C is a member of a minimum cycle base then for all for all vertices x in C there is an edge $(y,z) \in C$ such that $C = P(x,y) \oplus P(x,z) \oplus (y,z)$ where $P(a,b)$ denotes a shortest path from a to b.

Remark. The converse is not true in general.

Definition. An elementary cycle C is a *simple cycle* if $(p,q) \in E$ holds if and only if $(p,q) \in C$.

In other words, an elementary cycle is simple if and only if it is an induced subgraph of Γ. Simple cycles play an important role in ring perception algorithms for chemical structure storage and retrieval systems [4].

Corollary 1. If C is a member of minimal cycle base, then C is simple.

Proof. If $|C| = 3$ there is nothing to show. Hence assume $|C| \geq 4$. Suppose that C is a member of a minimal cycle basis and such an edge (p,q) exists. Set $x = p$ in Horton's theorem. The edge (p,q) divides C in two half-cycle. Let $(y,z) \in C$. The vertices y and z lie in the same half-cycle and (p,q) is of course shorter than either one of the half-cycles, each of which contains at least two edges. Thus the other half-cycle, which does not contain (y,z), cannot be contained in a shortest path from x to y or z (in E), and therefore $P(x,y) \oplus P(x,z) \oplus (x,y) \neq C$. ∎

Lemma 4. Let Y_α denote the set of bonds that are immediately interior to $\alpha \in B$. A cycle in a secondary structure is simple if and only if it is of the form

$$\overline{C_\alpha} = \left[\bigoplus_{\beta \in Y_\alpha} F_\beta \right] \oplus F_\alpha \, . \tag{2}$$

Proof. It is obvious that $\overline{C_\alpha}$ is simple cycle. In order to prove the converse, consider an elementary cycle $\overline{Q_\alpha}$ with upper bound α. Suppose that $\overline{Q_\alpha}$ contains a bond β that is not immediately interior to α (w.r.t. B). Then there is another bond, $\gamma \in B$, which is part of the subgraph H of Γ induced by $\overline{Q_\alpha}$, i.e., $\overline{Q_\alpha}$ in not simple. A bond δ immediately interior to α that is not contained in Q_α leads to the same contradiction. ∎

Theorem 2. The simple cycles $\overline{C_\alpha}$, $\alpha \in B$, form the unique minimum cycle basis of the secondary structure $\Gamma = (V,E)$.

Proof. The cycles defined in equ.(2) are the only possible candidates for members of minimum cycle bases as a consequence of lemma 4. Since a minimum cycle base always exists and there are exactly b such cycles we conclude that $\{\overline{C_\alpha}|\alpha \in B\}$ is the unique minimum cycle base of Γ. ∎

Remark. The cycles $\overline{C_\alpha}$ are exactly the "loops" on which the energy model for RNA is based.

It is interesting to note that the fundamental basis $\{F_\alpha|\alpha \in B\}$ has a simple relation with the minimum cycle basis. Adding $F_\alpha \oplus \overline{C_\alpha}$ to equ.(2) we find

$$\begin{aligned} F_\alpha &= \overline{C_\alpha} \oplus \left[\bigoplus_{\beta\in Y_\alpha} F_\beta\right] = \overline{C_\alpha} \oplus \left[\bigoplus_{\beta\in Y_\alpha} \overline{C_\beta} \oplus \left[\bigoplus_{\gamma\in Y_\beta} F_\gamma\right]\right] \\ &= \overline{C_\alpha} \oplus \left[\bigoplus_{\beta\in Y_\alpha} \overline{C_\beta} \oplus \left[\bigoplus_{\gamma\in Y_\beta} \overline{C_\gamma} \oplus \left[\bigoplus_{\delta\in Y_\gamma} F_\delta\right]\right]\right] = \ldots = \bigoplus_{\beta\in\{\alpha\}\cup W_\alpha} \overline{C_\beta} \end{aligned} \tag{3}$$

The final equality is established by observing that the expansion eventually stops if $Y_\psi = \emptyset$ and hence $F_\psi = \overline{C_\psi}$. Clearly, the nested sums contain each bond in $W_\alpha \cup \{\alpha\}$ exactly once.

The length of the minimum cycle basis $\{\overline{C_\alpha}|\alpha \in B\}$ can be computed explicitly. Let p be the number of 2-connected components, q the number of of edges that are not contained in any cycle. Each bond except the upper bound of a 2-connected component appears in exactly two cycles $\overline{C_\alpha}$. Each edge of T appears once if and only if it is contained in a cycle. Thus $\ell = n - 1 - q + 2b - p$.

Concluding Remarks

Condition (ii) of the definition of secondary structures is violated in many functionally important RNA molecules. These molecules form so-called *pseudo-knots*. Except for *ad hoc* assumptions for the most simple class of so-called *H-type pseudo-knots* [1] there is no energy model for these generalized classes of structures. It is unlikely that these more general contact structures have a unique minimal cycle basis. Recently Vismara [17] published an efficient algorithm for finding the set $\mathcal{C}_R$ of *relevant cycles*, that is, the union of all minimal cycle bases, see also [11]. The set $\mathcal{C}_R$ seem to form a natural extension of the energy model of nucleic acid contact structures.

References

[1] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.*, 18:3035–3044, 1990.

[2] G. Chartrand and F. Harary. Planar permutation graphs. *Ann. Inst. Henri Poincarè B*, 3:433–438, 1967.

[3] W.-K. Chen. On vector spaces associated with a graph. *SIAM J. Appl. Math.*, 20:525–529, 1971.

[4] G. M. Downs, V. J. Gillet, J. D. Holliday, and M. F. Lynch. Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.*, 29:172–187, 1989.

[5] 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.

[6] R. R. Gutell. Evolutionary characteristics of RNA: Inferring higher-order structure from patterns of sequence variation. *Curr. Opin. Struct. Biol*, 3:313–322, 1993.

[7] L. He, R. Kierzek, J. SantaLucia, A. E. Walter, and D. H. Turner. Nearest-neighbor parameters for GU mismatches. *Biochemistry*, 30, 1991.

[8] J. D. Horton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput.*, 16:359–366, 1987.

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

[10] 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.

[11] M. Plotkin. Mathematical basis of ring-finding algorithms in CIDS. *J. Chem. Doc.*, 11:60–63, 1971.

[12] C. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Comp. & Chem.*, 20:85–94, 1996.

[13] P. F. Stadler and C. Haslinger. RNA structures with pseudo-knots: Graph-theoretical and combinatorial properties. *Bull. Math. Biol.*, 1997. submitted, Santa Fe Institute Preprint 97-03-030.

[14] M. M. Sysło. An efficient cycle vector space algorithm for listing all cycles of a planar graph. *SIAM J. Comput.*, 10:797–808, 1981.

[15] K. Thulasiraman and M. N. S. Swamy. *Graphs: Theory and Algorithms.* J. Wiley & Sons, New York, 1992.

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

[17] P. Vismara. Union of all the minimum cycle bases of a graph. *Electronic J. Comb.*, 4:#R9 (15 pages), 1997.

[18] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of rna folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

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

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

[21] L. Yan. A family of special outerplanar graphs with only one triangle satisfying the cycle basis interpolation property. *Discr. Math.*, 143:293–297, 1995.

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

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