Journal of Knot Theory and Its Ramifications, Vol. 3 No. 3 (Special Issue, 1994) 233–245

GRAPH INVARIANTS AND THE TOPOLOGY OF RNA FOLDING

LOUIS H. KAUFFMAN
Department of Mathematics, Statistics and Computer Science (M/C 249)
University of Illinois at Chicago, 851 South Morgan Street
Chicago, Illinois 60607-7045, USA

YURI MAGARSHAK
Biomathematical Sciences Department, Mount Sinai School of Medicine
City University of New York, New York 10029, USA

Received 3 February 1994
Revised 23 May 1994

ABSTRACT

A new general method is described for obtaining ambient isotopy or regular isotopy invariants of even valence rigid vertex graphs embedded in three-dimensional space. The paper concentrates on the case of 4-valent vertices and defines an RNA vertex in analogy to the structure of a folded molecule. Examples are given to show how these methods can discriminate graph embeddings that are indistinguishable via Vassiliev invariants. Applications to molecular folding are discussed.

Introduction

This paper describes a new general method for obtaining ambient isotopy or regular isotopy invariants of even valence rigid vertex graphs embedded in three-dimensional space. We concentrate on the case of 4-valent vertices and define an RNA vertex in analogy to the structure of a folded protein molecule.

Section 1 describes the general setting of the invariants. These invariants are obtained by substituting tangles for the vertices of the rigid vertex graph that is being analyzed. Proposition 1 details how to obtain topological information about the graph embedding from such substitutions. A special case of interest is the substitution of just a crossing or its reverse at a 4-valent vertex. This gives the formalism for the Vassiliev invariants. Section 1 gives an example of two graph embeddings that can not be discriminated by any Vassiliev invariant, but are discriminated when the substitution of a smoothing of the vertex is included in the definition of the invariant. Section 1 defines a vertex structure that we term an RNA vertex. The RNA vertex is an abstraction of the structure of a site of consecutive molecular bonds in a folded chain . The mathematical structure of this RNA vertex lends itself to abstract analysis. In Section 2 we discuss the motivation for the definition of the RNA vertex in terms of the structure of foldings of the RNA molecule, and we delineate possible applications.

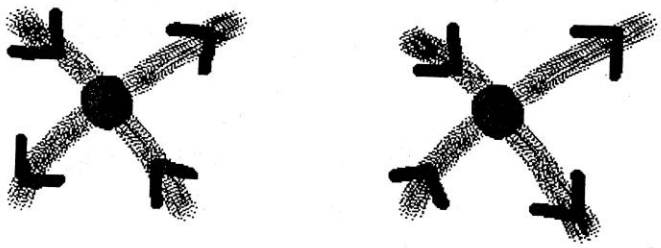
1. Invariants of Embedded Graphs in 3-Space

The invariants discussed in this paper use little more than the structure of a rigid vertex and the existence of invariants of knots and links in 3-space. We generalize the invariants discussed in [6] and [8] to a large class of invariants of rigid vertex graphs with arbitrary numbers of lines impinging on a vertex. Because of our interest in applications involving 4-valent vertices, this discussion will first concentrate on 4-valent vertices. The full generalization is discussed at the end of this section.

We begin by recalling the notion of a rigid vertex embeddings of a 4-valent graph in 3-space [6]. The 4-valent vertex can be regarded as a rigid disk with flexible strings attached at four given points. Thus a cyclic order is assigned to this vertex, and not all permutations of the strings are allowed in a deformation of it. The topological properties of such a vertex are summarized in the list of generalized Reidemeister moves for rigid vertex graphs. These moves are illustrated in Figure 1.

Here we include the first Reidemeister move in the list of allowed operations so that these moves generalize ambient isotopy. Moves I, II and III are the Reidemeister moves for link diagrams. Moves IV and V are the extra moves for the graphs. The illustrations in Figure 1 do not include all cases of the moves. Other cases can be obtained by changing crossings in these figures in the obvious way. The discussion can be easily broadened to include regular isotopy or ambient isotopy of framed links and graphs.

In this paper we assume two possible orientations for the 4-valent rigid vertex, as shown below.



alternate **standard**

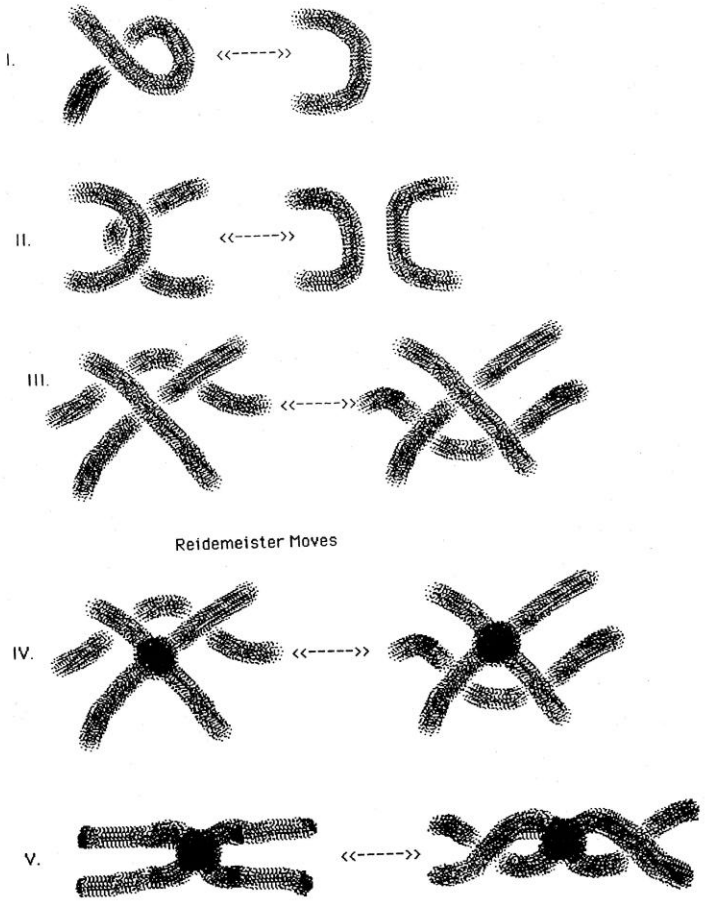


Fig. 1. Rigid Vertex Graph Moves

These types of vertex will be called *standard* and *alternate*. In both types of vertex there are two ingoing lines and two outgoing lines. In the standard vertex the two ingoing lines are adjacent as are the outgoing lines. In the alternate vertex the ingoing and outgoing lines alternate adjacency.

Our invariants of rigid vertex graphs are based on topological information in the set of knots and links obtained by eliminating all the rigid vertices – replacing each vertex by a tangle that respects the orientations of the vertex. Recall that a tangle is a piece of link diagram with free ends indicated in a given pattern so that the configuration can be put in a box with the diagram inside except for the free ends emanating from the box. Two tangles are *equivalent* if there is an ambient isotopy taking one tangle to the other that restricts to the box and leaves the boundary of the box and the strings emanating from it fixed.

Proposition 1. Let G be an embedding of a rigid vertex graph in 3-dimensional space. Let $T(G)$ be a link that is obtained from G by replacing each vertex in G by a tangle (possibly using a different tangle for each vertex of G). Let the vertices of G be labeled from the set $\{1, 2, \ldots, n\}$ and let the tangle replacing vertex i be denoted by T_i. Suppose that G is rigid vertex isotopic to G'. Label the vertices of G' from $\{1, 2, \ldots, n\}$ so that the vertex i of G goes to the vertex i of G' under this isotopy. Then $T(G)$ is ambient isotopic to $T(G')$ where $T(G')$ is obtained from G' by replacing the i-th vertex by T_i. Thus the ambient isotopy class of the link $T(G)$ is an invariant of the rigid vertex isotopy class of the graph G.

Proof. The rigidity of the vertex means that during the rigid vertex isotopy we can replace each vertex by a box that is moved rigidly in 3-space. If the contents of this box consist in a tangle, then the tangle is carried along by this isotopy. This results in an ambient isotopy of the link $T(G)$. Since each tangle is inside its rigidly moved box, it is not changed by the isotopy except for the translation to its new position. This means that image of $T(G)$ under the induced isotopy is of the form $T(G')$, as described in the statement of this proposition. This completes the proof. □

Discussion. In order to use this proposition, it is useful to have a standard procedure for inserting the set of tangles $\{T_i | i = 1, 2, \ldots, n\}$ into the graph G. This is easily accomplished in many cases. For example, in the case of standard vertices we can insert the tangles $[n/2]$, where n is an integer – as shown below. The tangle $[n/2]$ consists in two strands with $|n|$ half twists of sign(n).

$[n/2] : n$ half-twists

For a given standard vertex there is no ambiguity about the insertion of $[n/2]$ at that vertex. Note that these insertions at standard vertices include the smoothing ($[0]$), and the two resolutions ($[+1/2]$ and $[-1/2]$.

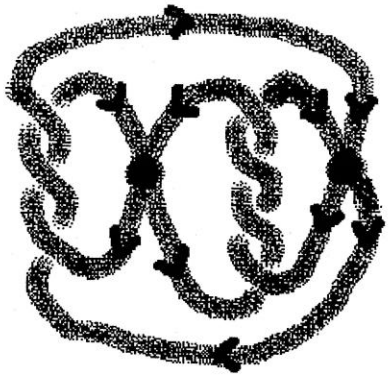
In the case of standard orientation we can take a given rigid vertex embedding and consider all the knots and links that are obtained by replacing a vertex either with a single crossing, or with a smoothing. The ambient isotopy class of this set of links is an invariant of the original graph.

If we have available an ambient isotopy (or regular isotopy) invariant $R(K)$ of oriented links K, then this construction gives rise to an extension of the invariant R to the category of rigid vertex graphs. For the insertion set described above (insert a crossing or a smoothing) this extension is expressed by the formula $R[G|V] = aR[G|V_+] + bR[G|V_-] + cR[G|V_0]$ where $G|V$ denotes a graph embedding G with a selected vertex V, $G|V_+$ and $G|V_-$ denote the embeddings obtained from $G|V$ by replacing V by a positive or a negative crossing and $G|V_0$ denotes the result of smoothing G.

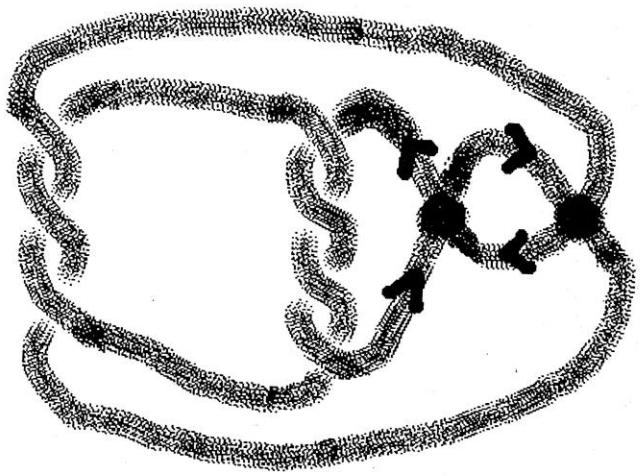
This formula gives an implicit expansion of $R[G]$ as a polynomial in a, b, c with coefficients in the ring of values of the invariant R restricted to knots and links. The expansion is well-defined at the outset, so that we only need verify that the extension of R to rigid vertex graphs is indeed invariant under the generalized Reidemeister moves. This follows directly from Proposition 1.

Example. The case $a = +1$, $b = -1$ and $c = 0$ is the well-known case of the Vassiliev invariants [2], [7]. A Vassiliev invariant $V[G]$ is said to be of *finite type* i if $V[G] = 0$ for all graphs with more than i vertices. For the structure of invariants of knots and links, these are a very important class of invariants. As a class of graph invariants the Vassiliev invariants have some obvious limitations, for it is possible to give two distinct graphs with exactly the same Vassiliev invariants for all choices of Vassiliev invariant. Such an example is shown below.

Example.



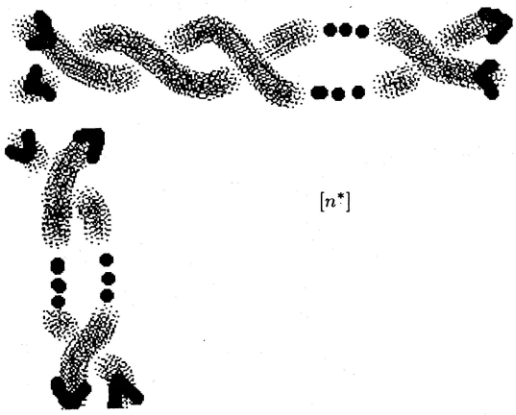
Vassiliev invariants cannot distinguish this graph embedding from the one shown below because the only substitution allowed is a single crossing inserted at the 4-valent vertex. Each such insertion on the graph above is ambient isotopic to the corresponding insertion on the graph below.



Note that the two graphs in this example can be distinguished by smoothing all the vertices (a replacement not available for the Vassiliev invariant). One graph yields a link with two knotted components while the other yields a link with one knotted component and one unknotted component. Thus, by Proposition 1, the two graphs are not rigid vertex isotopic.

The RNA Vertex

For an alternate vertex, we can insert $[n^*]$ for n an integer, where $[n^*]$ denotes a pair of oppositely oriented strands with linking number n as shown below. Note that there are two ways to insert $[n^*]$ into the given alternate vertex. For the alternate vertex the simplest examples of this insertion are $[1^*]$, $[-1^*]$ and the smoothing $[0^*]$.



$[n^*]$

$(2n$ half twists).

For the RNA application it is convenient to diagram an alternate vertex with extra structure as shown below.

One can think of this vertex as two oppositely oriented strands that are bound together at the site indicated by the arcs drawn between them. Call such a vertex an ***RNA vertex.***

In an RNA vertex we can make the distinction between the two types of insertion of the tangle $[n^*]$. We call the insertion that preserves the lines in the RNA vertex a ***horizontal insertion,*** and the insertion that recombines these lines a ***vertical insertion.***

With this distinction, we are in position to delineate invariants of graphs with RNA vertices by allowing the horizontal and vertical insertions of $[n^*]$ for each integer n. Let $[Hn^*]$ denote a horizontal insertion and $[Vn^*]$ denote a vertical insertion. If it is desirable to obtain an invariant with new variables for each insertion (extending a given ambient isotopy invariant R) this can be done by restricting the insertion to a range of integers between $-N$ and $+N$. We then define the extension of R via the expansion

$$R[G|v] = \sum_{n=-N}^{N} \{a_n R[G|Hn^*] + b_n R[G|Vn^*]\}$$

where $G|v$ denotes a graph with a selected RNA vertex v and $G|Hn^*$, $G|Vn^*$ denote the results of inserting n^* either horizontally or vertically at this vertex.

Repeated application of this formula yields a state summation for $R[G]$, where the states S are all possible vertical and horizontal insertions of n^* (with $-N < n < N$) into the vertices of G. Call the $\{a_i\}$ and the $\{b_i\}$ the (horizontal and vertical) *vertex weights* for the state S. That is, each vertex v of G is associated with an insertion in the state S of $[Hn^*]$ or of $[Vn^*]$. The vertex weight for the insertion $[Hn^*]$ is a_n and the vertex weight for the insertion $[Vn^*]$ is b_n. Let $[G|S]$ denote the product of the vertex weights associated with the state S, and let $L[S]$ denote the link obtained from G by the insertions for the state S. Then the state summation for $R[G]$ is given by the formula

$$R[G] = \sum_S [G|S] R[L(S)].$$

It follows directly from Proposition 1 that $R[G]$ is a rigid vertex isotopy invariant of G whenever R is an ambient isotopy invariant of oriented knots and links. This gives a wide range of invariants to consider for the applications to the topology of RNA folding.

Finally, it is worth mentioning that the invariants obtained by using only horizontal insertions match the possibility that the RNA bond has some flexibility around a horizontal axis. We shall discuss this point further in the next section.

The General Rigid Vertex

The same approach that we have outlined for the case of 4-valent rigid vertices works for all cases of rigid vertices of *even* valence. With a vertex of even valence it is possible to insert a tangle at that vertex (a line will be left over in the case of odd valence). The generalized Reidemeister moves for rigid vertices of arbitrary valence are just what one expects: One must add a slide move that allows a strand that underpasses (overpasses) a subset of adjacent edges to a vertex to be moved so that it underpasses (overpasses) the complement of these edges at the vertex; one must allow twisting of the vertex that does not change the cyclic order of the strands emanating from it. The same arguments that we have already discussed

yield the direct generalization of Proposition 1 to graphs with even valence vertices. Once again, this means that specific insertions can be used to test the topology of a given embedded graph, and systematic insertions can be used to create extensions of invariants of knots and links to invariants of graphs with even valent vertices.

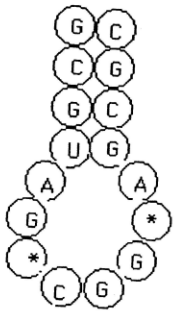
2. A Model for RNA Topology

An RNA molecule is a long string of bases with the characteristic feature that the bases on the string can pair with themselves. Thus the molecule folds on itself to form complex configurations. These configurations then perform many duties in the operations of the cell, including the manufacture of specific proteins. In this section we give a simple graph-theoretic model for studying the topology of the RNA as it is embedded in three dimensional space. This same model can be used to study the intrinsic topology of the folded RNA structures independent of their 3-dimensional embedding.

We begin with a description of the structure of an RNA folding. First there is a long molecular string of bases. The bases are denoted A (adenine), U (uracil), G (guanine) and C (cytosine). A and U can pair with each other; G and C can pair with each other. Thus a string such as

$$\cdots GCGUAG^*CGG^*AGCGC\cdots$$

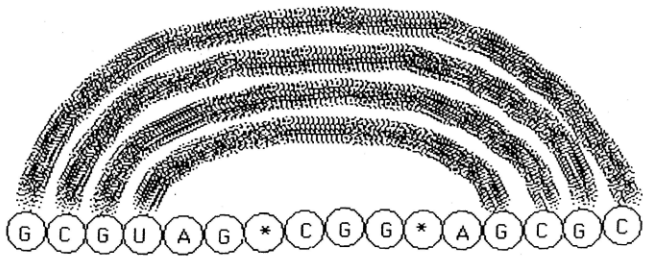
may fold up into the structure indicated below.



(Here we have indicated by * the possible presence of other molecules in the chain.)

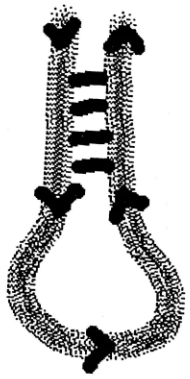
Long regions of double bonding can twist into double helical form reminiscent of the configurations of DNA.

The key structural feature of RNA is that it folds on itself. A given chain can have numerous possible foldings and it is a challenging combinatorial problem to catalog these foldings and to understand which of them are relevant to actual biological processes. Here we wish to consider the structure of a given folding. To this purpose we abstract the long chain to a long segment of directed line, and we indicate paired sites by arcs that join this line to itself as shown below.



The arcs can be drawn short to indicate a bonded configuration, or they can be drawn as semicircles to indicate how the folding is to proceed from a straight chain of molecules.

Note that in a folding involving a sequence of consecutive bonding sites the chain is matched to itself with the orientation reversed as shown below.

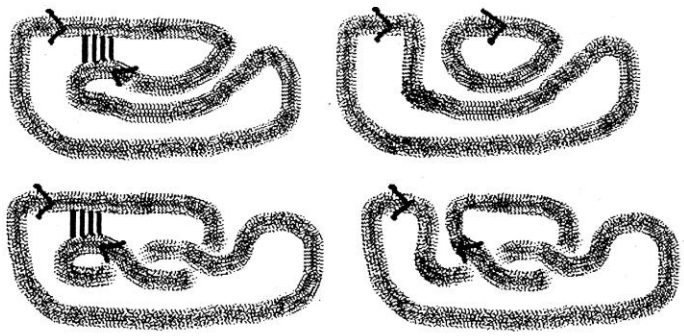


For a short chain of consecutive bonding sites such a configuration can be regarded as a rigid vertex. *It is exactly the RNA vertex of Section 1.* This completes our description of the way in which the methods of Section 1 can be applied to a model for the topology of RNA. It is easy to produce examples for topological analysis, and we refer the reader to [9] for examples of this sort.

The real question remains as to the applicability of these methods to molecular biology. For the mathematics itself, the concept of an RNA vertex and the problems of examining the classification of graphs containing RNA vertices is intrinsically interesting. Nevertheless it is worthwhile to propel these methods into practical use. A key problem about the structure of RNA is to understand why a given chain, with many possible foldings, pairs with itself to produce the particular folded RNA that performs its duties in the cell. It is not yet clear what role the topology of the molecules plays in these processes. (Compare with the discussion in [1].) However, there are some structures of embedding that are so simple that they must occur in reality by sheer probability. For example, consider the folding shown below. In this case the strand has undergone a simple twist before attaching itself to itself.

The resulting configuration is topologically distinct from the simpler direct folding shown below.

In this paper we have made precise just what is this topological difference.



In our terms, we make each of the folded molecules into a closed loop, and then apply Proposition 1 of Section 1. In this case, we can distinguish the two embeddings by performing a single insertion of vertical type in each graph. The resulting links are linked with linking number one in the twisted case and unlinked in the untwisted case. These operations are illustrated above. This gives an easy proof of the topological difference between the foldings. More complex cases can be handled in a similar way.

The use of the invariants defined in this paper will depend upon the availability of appropriate data about RNA folding. The models and methods that we have articulated are conceptually of great simplicity. We look forward to seeing how this clarity will affect the understanding of biological process.

Acknowledgement

Research for this paper on the part of the first author was partially supported by the Program for Mathematics and Molecular Biology, University of California at Berkeley and by NSF Grant No. DMS9205277. On the part of the second author, research was partially supported by NSF Grant No. CHE 9123802.

References

[1] C.J. Benham and M.S. Jafri, *Disulfide bonding patterns and protein topologies,* Protein Science, No. 2 (1993), Cambridge Univ. Press, pp. 41-54.

[2] Dror Bar-Natan, *On the Vassiliev Knot Invariants,* (preprint 1992).

[3] L.H. Kauffman, *Knots and Physics,* World Sci. Pub. (1991), Second Edition 1993.

[4] L.H. Kauffman, *State models for link polynomials,* L'Enseignment Math. No. 36 (1991), pp. 1-37.

[5] L.H. Kauffman, *An invariant of regular isotopy,* Trans. Amer. Math. Soc., Vol. 318, No. 2 (1991), pp. 697-710.

[6] L.H. Kauffman, *Invariants of graphs in 3-space,* Trans. Amer. Math. Soc., Vol. 311, No. 2 (1989), pp. 697-710.

[7] L.H. Kauffman, *Vassiliev invariants and the loop states in quantum gravity,* In Knots and Quantum Gravity, ed. J. Baez., Oxford Univ. Press (1994).

[8] L.H. Kauffman and P. Vogel, *Link polynomials and a graphical calculus,* Journal of Knot Theory and Its Ramifications, Vol. 1, No. 1 (March 1992), pp. 59-104.

[9] L.H. Kauffman and Y. Magarshak, *Vassiliev knot invariants and the structure of RNA folding,* (to appear in Knot Theory and Its Applications, ed. by L.H. Kauffman, World Sci. Pub. (1994)).

[10] Y. Magarshak, *Quaternion representation of RNA sequences of tertiary structures,* In Computer Genetics, ed. P. Pevsner and M. Gelfand, Elsvier Science Pub. (1993).

[11] Y. Magarshak and C.J. Benham, *A algebraic representation of RNA secondary structures,* Journal of Biomedical Structure and Dynamics, ISSN 0739-1102, Vol. 10, No. 3 (1992), pp. 465-488.

[12] R.C. Penner and M.S. Waterman, *Spaces of RNA secondary structures,* (to appear in Advances in Mathematics).

[13] D.W. Sumners, *Knot theory and DNA,* Proceedings of Symposia in Applied Mathematics, Vol. 45 (1992), Amer. Math, Soc., pp. 39-72.