VASSILIEV KNOT INVARIANTS AND THE STRUCTURE OF RNA FOLDING.

Louis H. Kauffman * and Yuri B. Magarshak†

September 10, 1993

1 Introduction.

It is the purpose of this paper to introduce certain combinatorial structures into the study of RNA folding. These structures are useful for the classification of foldings and for the topological classification of the embeddings of these foldings into three-dimensional space. Both the abstract classification and the topological classification are highly relevant to problems in molecular biology - where these folded structures are instantiated as molecules in a three - dimensional ambient physical space.

The paper is organized as follows. In section 2 we discuss the basic idea of a folding (folded molecule) and graphical models for such foldings. We introduce the use of the Brauer monoid for the classification of non- embedded foldings. This introduces a multiplicative structure into the set of foldings and we discuss the structure of the resulting algebra. Section 3 discusses the relationship of foldings and topological invariants of embedded rigid - vertex graphs. Vertices arise in foldings as loci of a linear sequence of base pairs. We translate these folding vertices into standard 4-valent vertices and thereby obtain a translation of rigid vertex invariants to the category of folded molecular structures. This section discusses both Vassiliev invariants and a more general scheme that produces invariants of embedded foldings from any topological invariant of knots and links. Section 4 gives specific information about the Vassiliev invariants. In particular, we show how to construct a Vassiliev invariant of type 3, and we illustrate how Lie algebras give rise to Vassiliev invariants. The appendix discusses this last point in more detail. This key relationship between Lie algebras and Vassiliev invariants provides an interconnection among topological invariants, Lie algebras, Feynman diagrams and significant indices for protein

*Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, Illinois, 60680. † Biomathematical Sciences Department, Mount Sinai School of Medicine of City University of NY, NY 10029.

foldings. These connections are just beginning. We conjecture that these relationships occur at biological as well as topological levels of natural structure.

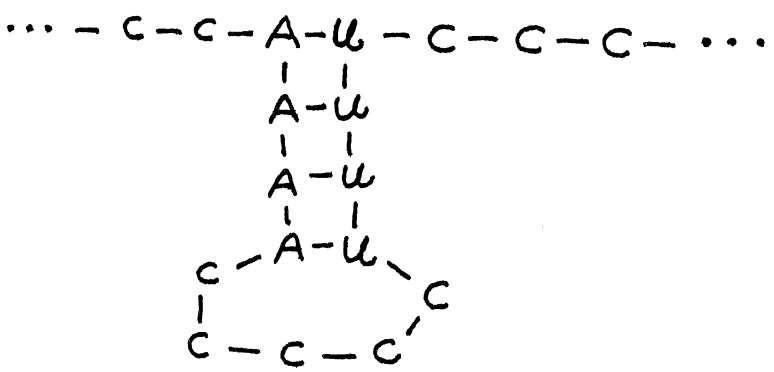
Acknowledgments. Research for this paper on the part of the first author was partially supported by the Program for Mathematics and Molecular Biology, University of California at Berkeley, CA, and by NSF Grant # DMS 9205277. On the part of the second author, research was partially supported by NSF Grant No. CHE 9123802.

2 Foldings and the Brauer Monoid.

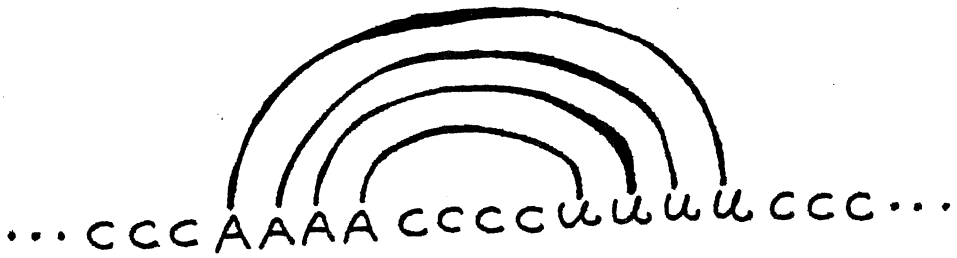
The purpose of this section is to introduce our abstraction of an RNA folding, and to give a method of enumerating such foldings in terms of the Brauer monoid [BR], an algebraic structure that generalizes the symmetric group on n letters. The Brauer monoid is of independent interest via its relationship with the theory of group representations [BR] and with the theory of invariants of knots and links ([BW], [K3], [KV]).

In order to begin, we need an appropriate mathematical abstraction for RNA. To this end, let us discuss some of the properties of the RNA molecule. The molecule is a long chain consisting in a sequence of the bases A (adenine), C (cytosine), U (uracil) and G (guanine). The pairs [A and U] and [C and G] are capable of bounding with each other. It is characteristic of RNA that the molecule can bond with itself. We say that two bases are paired if they are so bonded.

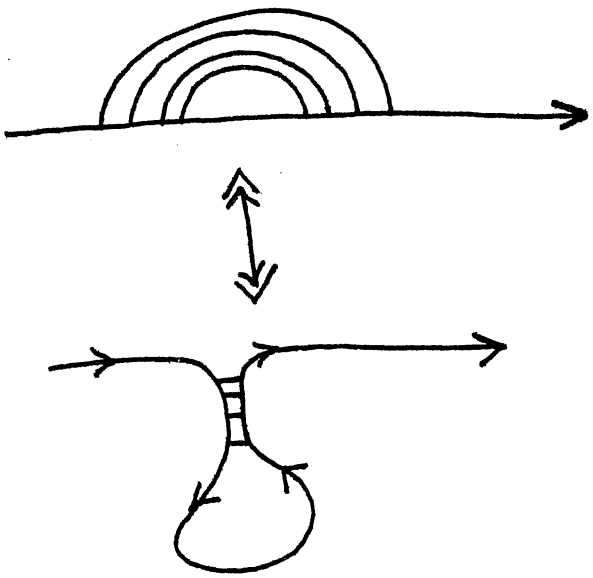
Thus, an abstract RNA molecule is just a linear sequence of the letters A, C, G and U. A *folding* of the molecule is a possible pairing structure with respect to the given sequence of bases. For example, we could have the chain $\cdots CCCAAAACCCCCCUUUUCCC\ldots\cdots$ and the corresponding folding



This folding can be indicated on the chain itself by a diagram (see also [M], [KMM]) with arcs connecting the paired bases:

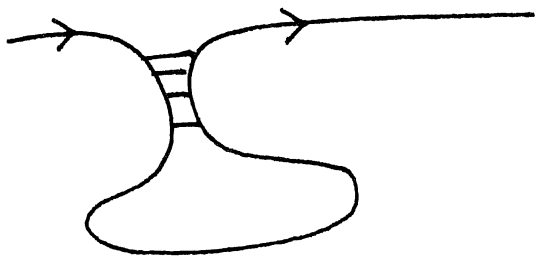


We may abstract this to a diagram that simply indicates the form of the pairing:



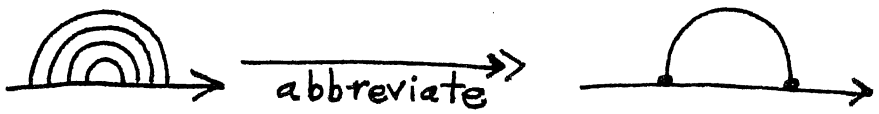
It is often the case that a sequence of repeated bases pairs with another sequence of repeated bases. This results in a *basic pairing node* of the form

Note the directions on the arcs of this node that correspond to the sequence of bases. The arcs are oppositely oriented just as in our example:

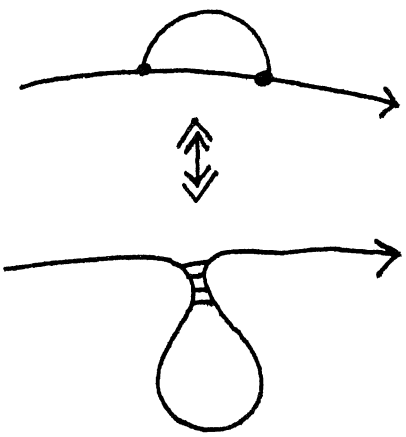


Any multiplicity of connecting arcs is possible, but we shall adopt the convention of four such arcs for pictorial purposes.

On the other hand, in an unfolded diagram it is useful to abbreviate a multiplicity of connecting arcs to a single connecting arc as in



Abbreviated arcs will be indicated by solid nodes  as shown above. The solid nodes will be called the *feet* of the connecting arcs. Thus, we have the correspondence



With these conventions, we can indicate the form of a great multiplicity of foldings (see Fig.1).

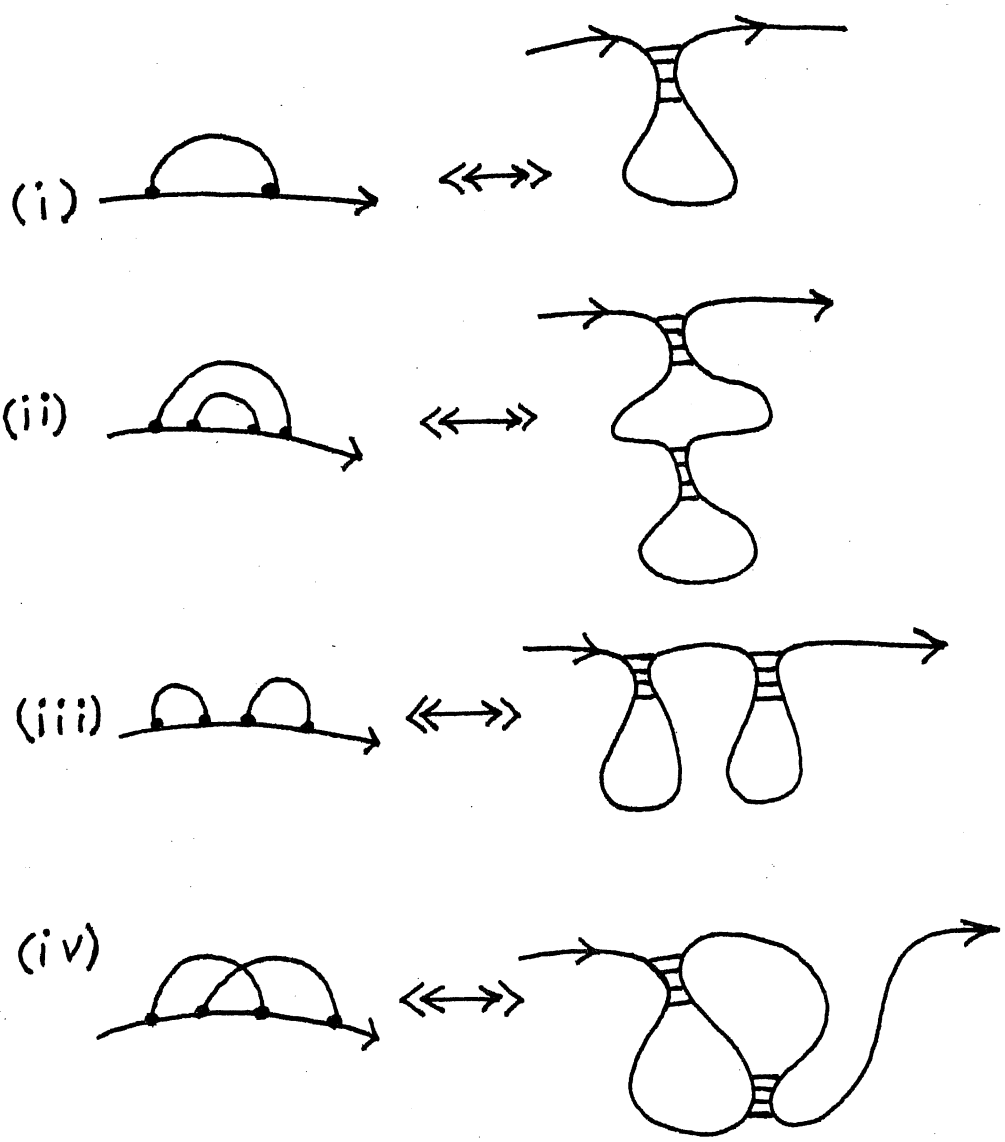
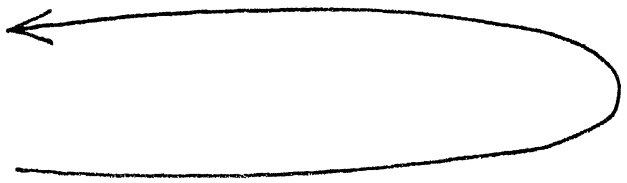


Figure 1

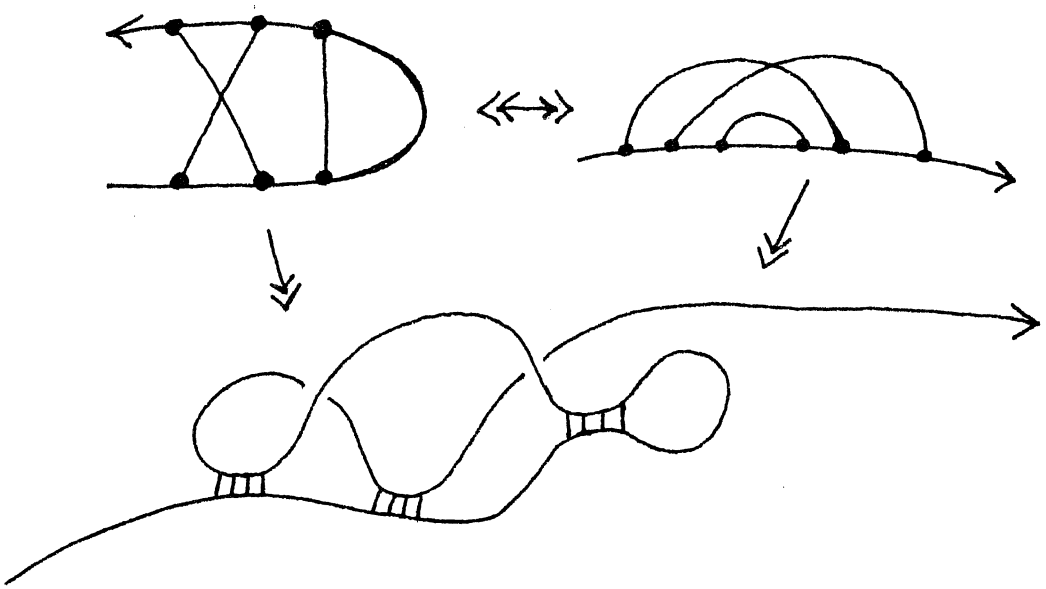
It is these forms that we are interested in classifying, first abstractly and then with respect to their possible embeddings in three dimensional space. The remainder of this section is devoted to the abstract classification.

First, note that the collection of arc diagrams bifurcates into those diagrams with non-intersecting arcs (examples (i), (ii) and (iii) in Fig.1) and those with arcs that necessarily intersect (example iv on Fig.1). We shall refer to foldings that correspond to arc diagrams free of intersections as secondary structures. The rest are *tertiary structures*. The simplest tertiary structure is the "pseudo-knot" illustrated as example (iv) in Fig.1.

We now discuss the following simple strategy for enumerating all arc diagrams (for secondary and tertiary structures). First bend the backbone (i.e. the line representing the linear sequence) into a "finger"as shown below.



A given arc diagram involves the pairing of 2N points for some positive integer N. Array the first N of these points on the bottom arc of the finger, and the second N on the top arc. Draw the connecting arcs in the bounded space of the finger. The example below shows this correspondence in a special case.



It is now a small step from the finger diagram to the *tangle diagram* consisting in two rows of N-points with arcs connecting the total 2N points in pairs. The arcs are restricted to the space between the two rows of points as shown below (Figure 2).

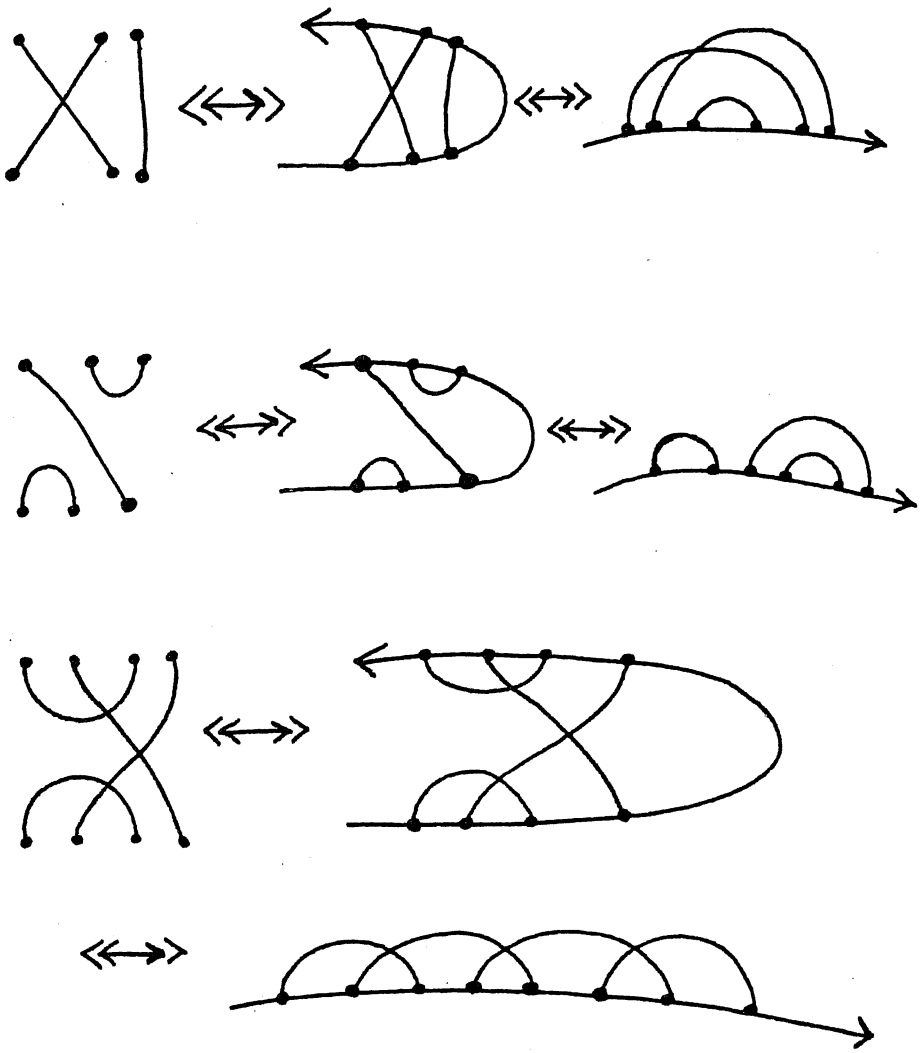
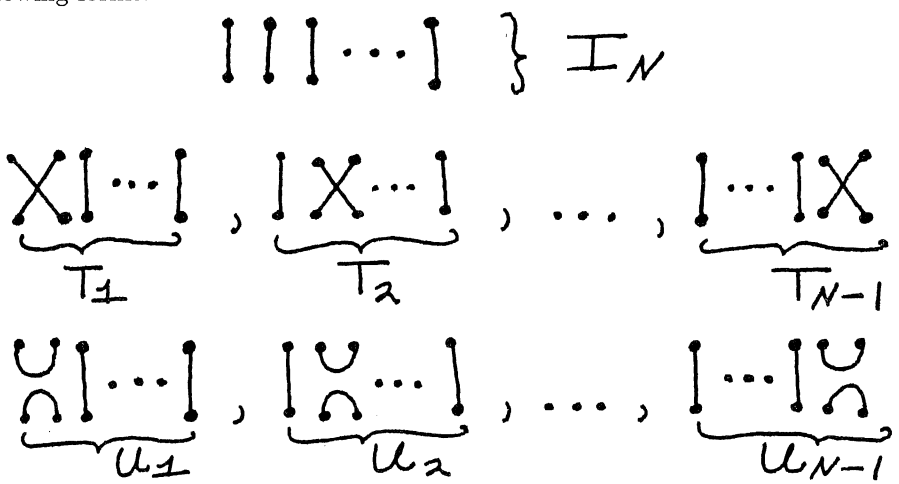


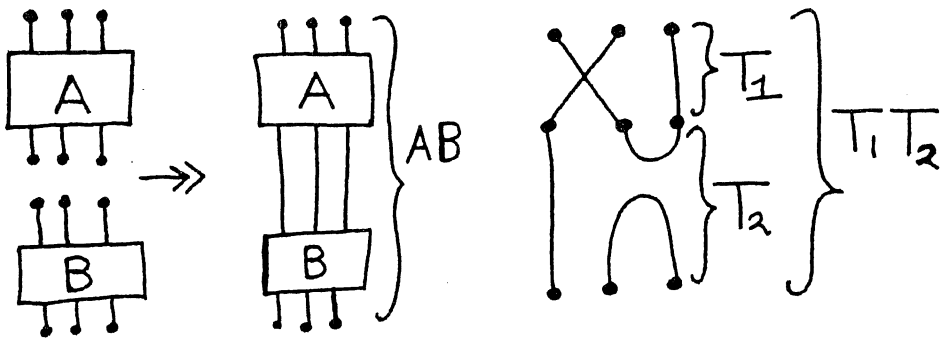
Figure 2

Thus we have shown that **foldings on 2N points are in one-to-one correspondence with tangle diagrams with two rows of N points.**

Let T_N denote the set of tangle diagrams with two rows of N points. Note that two such diagrams are equivalent if and only if they denote the same pattern of point connections. The pattern of intersections of arcs in the tangle is, however, relevant to the structure of the entire set of tangle diagrams. Every tangle diagram can be decomposed as a product of elementary diagrams of the following forms:



The product of tangle diagrams is obtained by attaching the bottom row of one diagram to the top row of the next diagram as indicated below.



Some products produce loops that are unattached to either row of points. For example letting δ denote the loop, we have $U_i^2 = \delta U_i$. See Fig.3. Thus, setting $\delta = 1$, we obtain an algebraic structure, the Brauer Monoid [B] (see also [BW],[K3]), on the set of foldings on 2N-points. The set of possible RNA-foldings has a rich algebraic structure. This is useful for classification and enumeration of foldings, and we suggest that it will eventually have even deeper implications

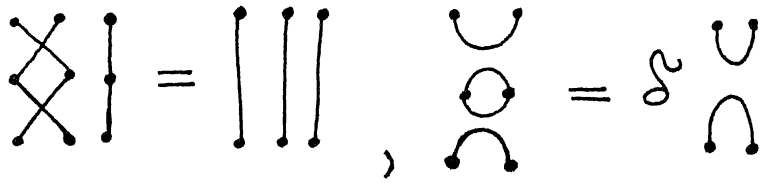
for molecular biology. It is worth stating some of the algebraic structure of the Brauer monoid explicitly for arbitrary loop value δ. Here is a summary:

i. δ commutes with every element of T_N.
ii. I_N is an identity element.
iii. $T_i^2 = I_n$, $T_iT_{i+1}T_i = T_{i+1}T_iT_{i+1}$
iv. $U_i^2 = \delta U_i$
v. $T_iU_i = U_iT_i = U_i$
vi. $T_iU_{i+1} = T_{i+1}U_iU_{i+1}$, $U_iT_{i+1} = U_iU_{i+1}T_i$.

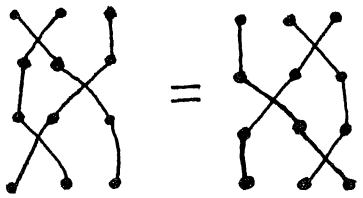
These relations abstractly specify the complete structure of (T_N, δ), the Brauer monoid of N-tangle diagrams with loop value δ. In Fig.3 we have indicated the diagrammatic picture that accompanies each of these relations. In Fig.4 we have listed all elements of T_3 as tangle diagrams, arc diagrams, and some corresponding foldings.

Fig.4 illustrates the sort of taxonomy provided by the Brauer Monoid. There are 15 structures in all, 15 being the number of ordered foldings on six sites. We have labelled these structures (1), (2),...,(15). Note that (1) $\rightarrow$ (5) are secondary structures. These form the subalgebra of the Brauer Monoid that is generated by $\{I_n, U_1, U_2, ..., U_{N-1}\}$. Call this subalgebra $T\mathcal{L}_N$. it is the multiplicative structure of the *Temperley-Lieb algebra* [BA] (see also [K3],[K1],[KV].) This algebra occurs in statistical mechanics, and it is the basis of the construction of the Jones polynonial in the theory of knots. The structures (6) $\rightarrow$ (10) are tertiary structures occurring as products of $T_1, T_2, ..., T_{N-1}$. The algebra generated by $\{I_N, T_1, T_2, ..., T_{N-1}\}$ has $N!$ elements and is isomorphic with the symmetric group on N letters. In other words, these foldings are in one-to-one correspondence with all permutations of N objects. Finally, the foldings (11) $\rightarrow$ (15) are mixed structures- products of U_i's and T_j's. The first pseudo-knot occurs at (7). We have illustrated particular embeddings in three-space associated with the structures (10) and (15). It will be interesting to compare this approach to the combinatorics of foldings with other methods. In particular, it appears to us that there is a fruitful interaction of the Brauer monoid technique with the methods of Magarshak and Benham in [MB]. This will be the subject of another paper. It is also of interest to compare our approach to that of Penner and Waterman [PW]. Penner uses the secondary structures, with some extra conditions, to create a cell structure for a moduli space for hyperbolic structures on Riemann surfaces. This gives a point of view on the topology of the space of all (unembedded) secondary structures.

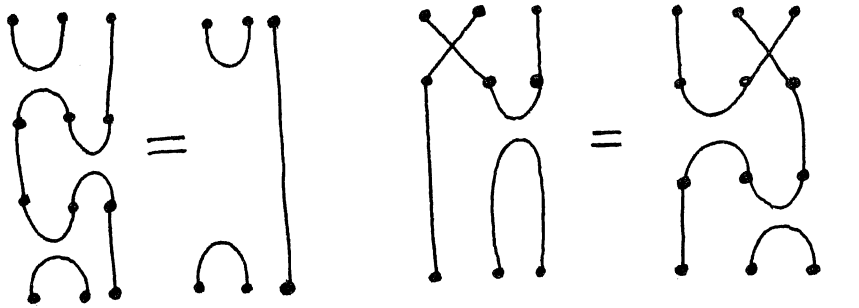
We now turn to the matter of embeddings.



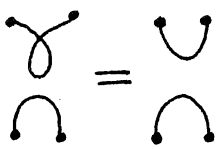
$$T_i^2 = I_N \qquad U_i^2 = \delta U_i$$



$$T_i T_{i+1} T_i = T_{i+1} T_i T_{i+1}$$



$$U_i U_{i+1} U_i = U_i \qquad T_i U_{i+1} = T_{i+1} U_i U_{i+1}$$



$$T_i U_i = U_i$$

Figure 3

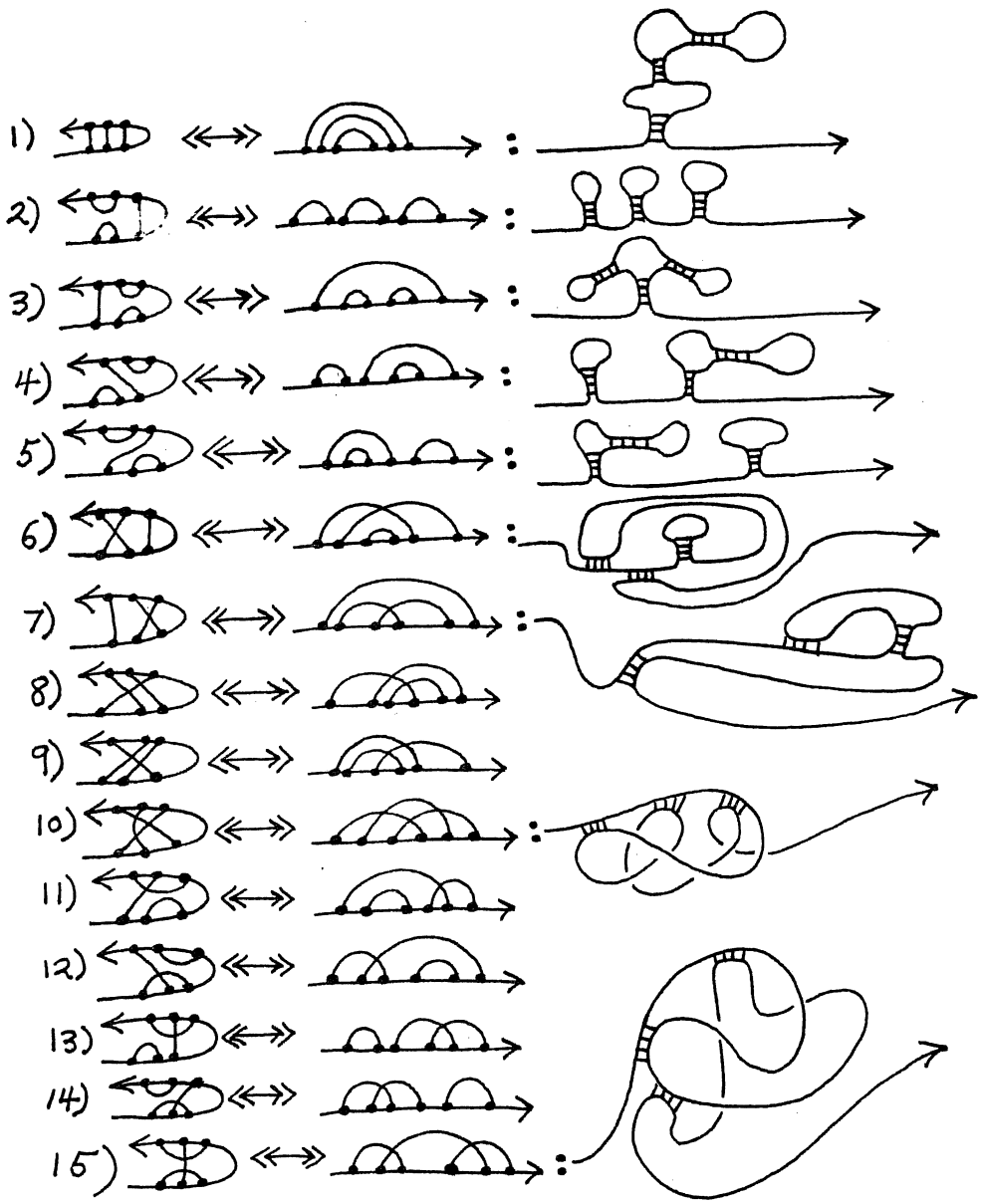
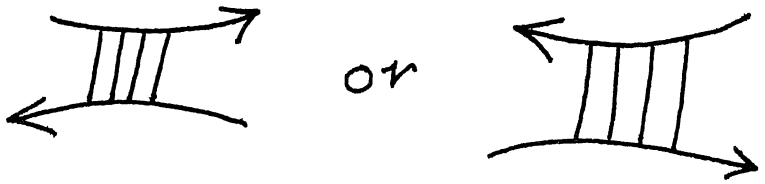


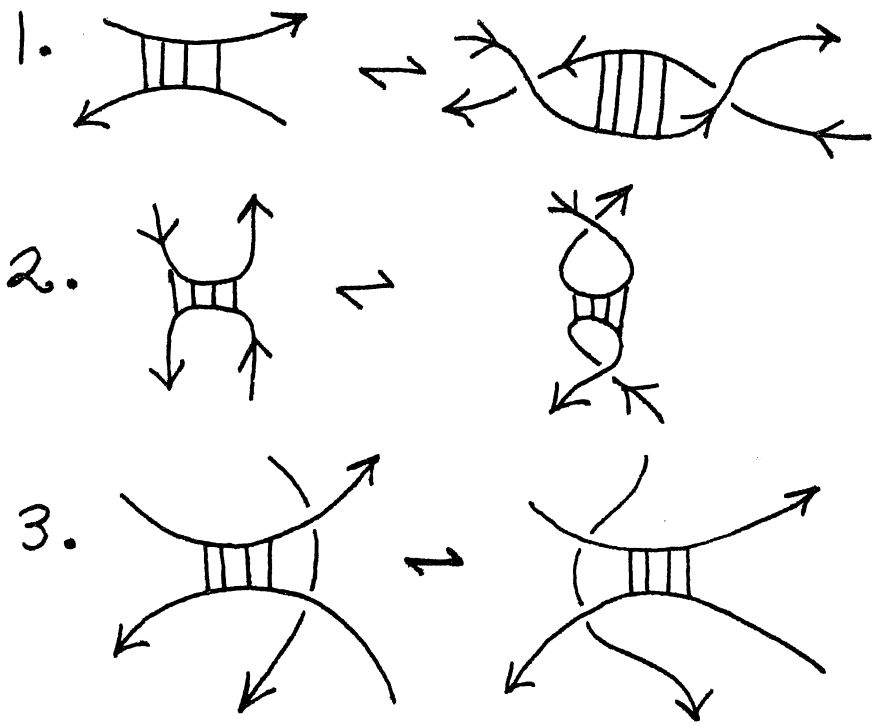
Figure 4

3 Embedded Foldings, Graph Invariants and the Vassiliev Invariants.

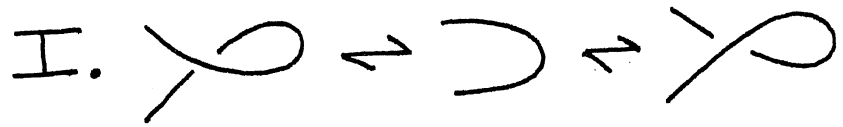
In order to study the topology of RNA foldings in three dimensional space it is necessary to specify an appropriate mathematical model for this topology. We take the lead for this model from the form of our basic bond vertex:

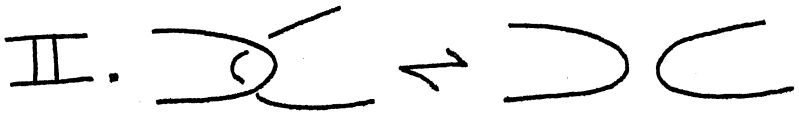


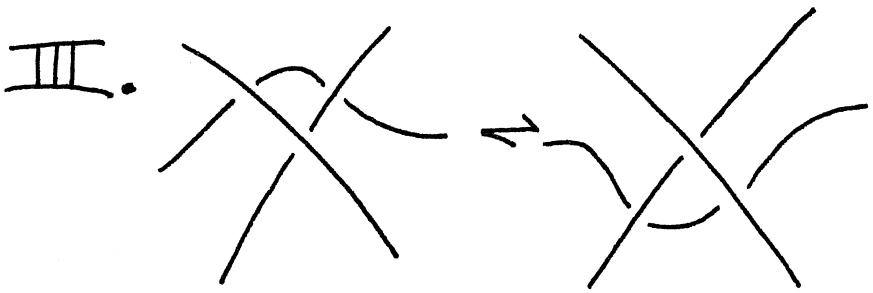
We take this to be a *rigid* vertex by which is meant that the configuration of bonding arcs is rigid (not subject to any twisting) while the oriented arcs that enter or leave the vertex are topologically flexible. This means that the following moves are available with respect to the vertex:



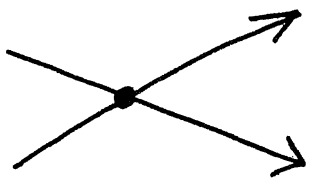
These moves (and obvious symmetries obtained by mirror imaging) plus the usual Reidemeister moves [K1] away from the bonds constitute our topological model for rigid vertex isotopy [K4]. For the record, the basic Reidemeister moves are shown below:







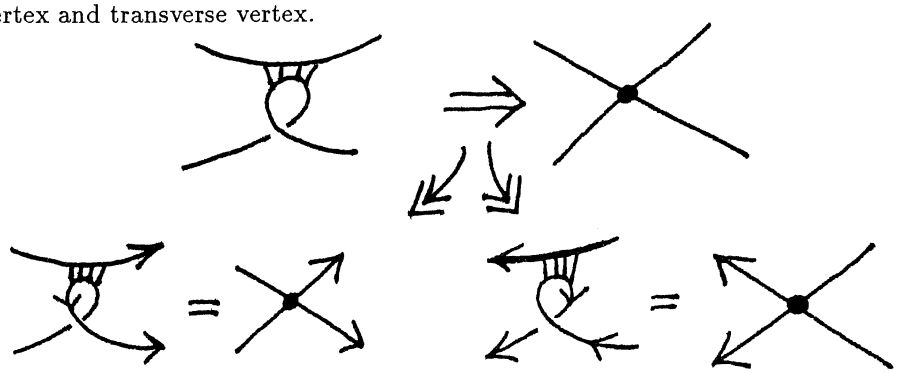
The invariants that we are about to discuss are indeed invariants of rigid-vertex graph embeddings in three dimensional space. However, they are formulated in the mathematical literature with respect to a rigid vertex with a different structure. This structure is as shown below.



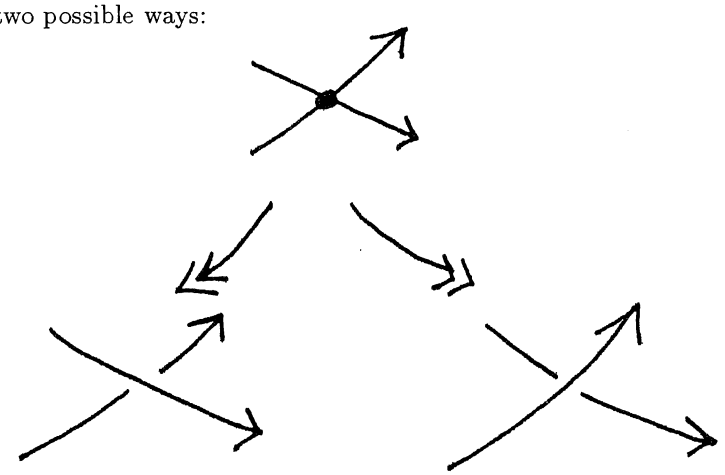
In this vertex the strands that bond go cross-wise to one another, forming a vertex with two in-going and two out-going lines. Mathematical formulations with respect to this *transverse vertex* are particularly convenient and symmetrical. Consequently, we shall define a conventional relationship between the transverse vertex and the bond vertex so that they can (up to a translation) be

used interchangeably.

Definition. By convention, we take the following relationship between bond vertex and transverse vertex.

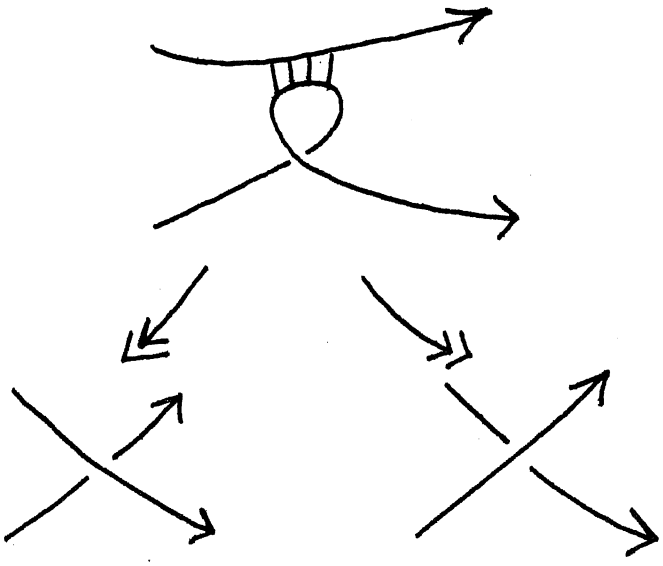


From the point of view of RNA-folding, this serves as the definition of the transverse vertex. An invariant of rigid isotopy for transverse vertices will automatically be an invariant of rigid isotopy for bond vertices (and conversely). Thus we shall discuss invariants in the context of the transverse vertices. The translation between bond vertices and transverse vertices is simple, but it does involve a definite shift of context. For example (and this is quite important), we can "resolve" a transverse vertex into a crossing of two lines that do not touch in two possible ways:



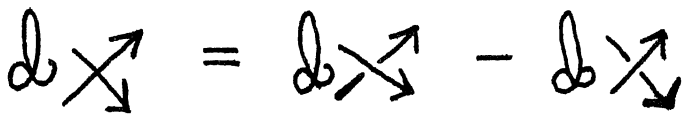
These resolutions are most natural in the transverse context, where *we visualize the vertex as a stage in the act of passing one line through the other.*

Translating this scenario to the bond vertex, we find



Thus the resolution of the bond vertex involves a recombination that may have no biological significance in the context of RNA. It is interesting to speculate about the possible meaning of mathematical operations in terms of biology. In this case, we justify including such recombinations because they allow us to calculate topological invariants.

Now, let us begin the topology. First of all, suppose that $\mathcal{I}_K$ is an invariant that assigns a number to an oriented knot or link K so that if K and K' are related by a sequence of Reidemeister moves, then $\mathcal{I}_K = \mathcal{I}'_K$. Given such an $\mathcal{I}_K$, we can define an extension of $\mathcal{I}$ to include rigid vertex graphs. We do this by the equation

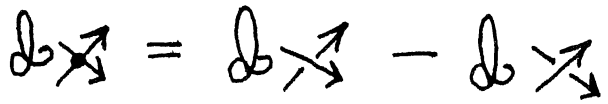


(this is part of a more general scheme, see [KV]). More specifically, let G be a graph embedding with rigid vertices $V_1, V_2, ..., V_n$. Let $\vec{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ be a vector with $\epsilon_i = \pm 1$ for each i. Let $[2^n]$ denote the set of these vectors. Let $G(\vec{\epsilon})$ denote the link or knot that is obtained from G by replacing V_i by a crossing of type ϵ_i. Here  has type $+1$ and  has type -1. Let $|\vec{\epsilon}|$ denote the

number of $\epsilon_i = -1$ in $\vec{\epsilon}$. Now define $\mathcal{I}_G$ by the formula

$$\mathcal{I}_G = \sum_{\epsilon \in [2^n]} (-1)^{|\vec{\epsilon}|} \mathcal{I}_{G(\vec{\epsilon})} \tag{1}$$

This formula gives a well-defined value to $\mathcal{I}_G$ in terms of the values of $\mathcal{I}$ on knots and links and it is obvious that $\mathcal{I}$ satisfies the formula

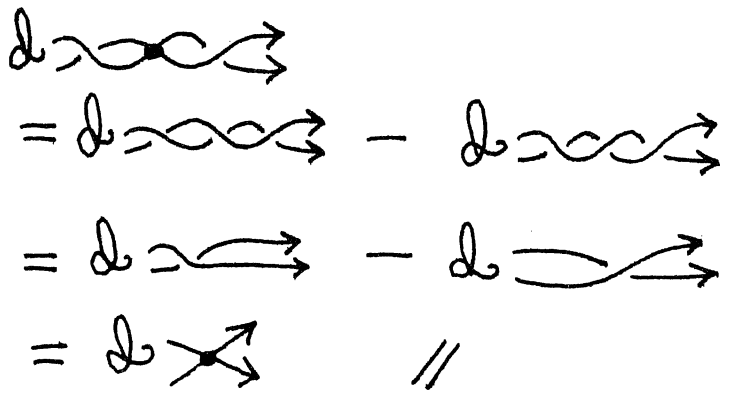


where the small diagrams are regarded as parts of an otherwise unchanged larger diagram.

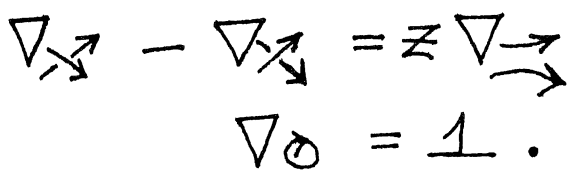
We now have the following basic lemma [KV].

Lemma. If $\mathcal{I}$ is an ambient isotopy invariant of knots an links, then its extension to rigid vertex graphs, as defined above, is an invariant of rigid vertex isotopy.

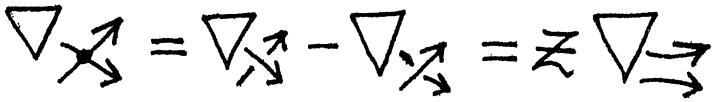
Proof. We just check one case of rigid vertex isotopy, leaving the rest as an exercise for the reader.



Example. Let $\nabla_K(z)$ denote the Conway (Alexander) polynomial of K, for K a knot or link. Then ∇_K is determined by the axioms [K1]



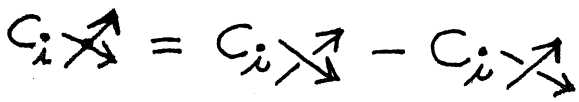
Thus, in the graph extension, we have



From this we see that if G has n vertices, then ∇_G is divisible by z^n. If we think of ∇_K as a set of numerical invariants (the polynomial coefficients), then we have

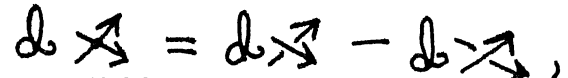
$$\nabla_K = \sum_{i=0}^{\infty} c_i(K) z^i$$

(c_i eventually zero for any given knot or link K). The statement that G with n vertices $\Rightarrow z^n | \nabla_G$ then becomes: $c_i(G) = 0$ if G has $> i$ vertices. We say that the invariants c_i are of finite type. Note that they also satisfy the identity



This leads to the following

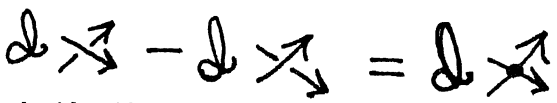
Definition. An invariant $\mathcal{I}$, of rigid-vertex graphs, is said to be a Vassiliev invariant of finite type i, if it satisfies the following rules [V], [BL]:

1. 
2. $\mathcal{I}_G = 0$ if G has more than i vertices.

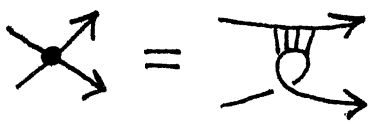
Thus we have shown that the rigid- vertex graph extensions of the coefficients of the Conway polynomial, $c_i(G)$, are Vassiliev invariants of type i.

The virtue of the Vassiliev invariants is that, being of finite type, they are determined by their behavior on a finite collection of graphs. These graphs can be interpreted as RNA foldings! Thus the Vassiliev invariants can give us information about the structure of embeddings of RNA foldings *and* they are also a way to look at the abstract structure of these foldings.

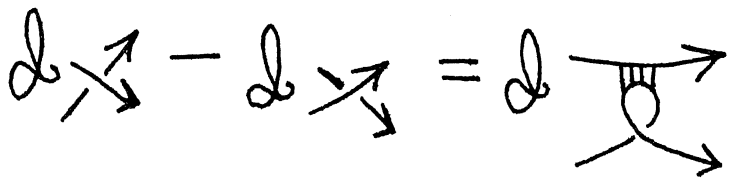
We shell give specific examples shortly. But first, it is necessary to look more closely at the translation from foldings to graphical nodes: in a Vassiliev invariant we have



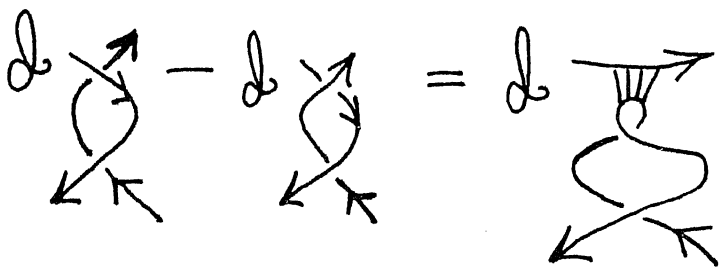
and we have made the identification



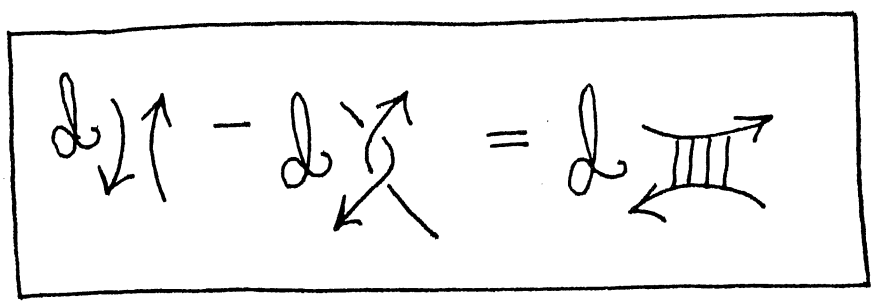
Therefore



whence,



Hence

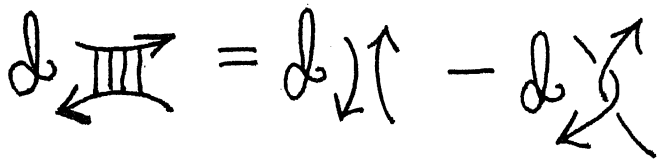


This is the basic equation for computing the Vassiliev invariant in the language of foldings.

It is this equation that may be of direct use to the microbiologist interested in the topology of foldings. Note that we can prove the basic invariance lemma

directly:

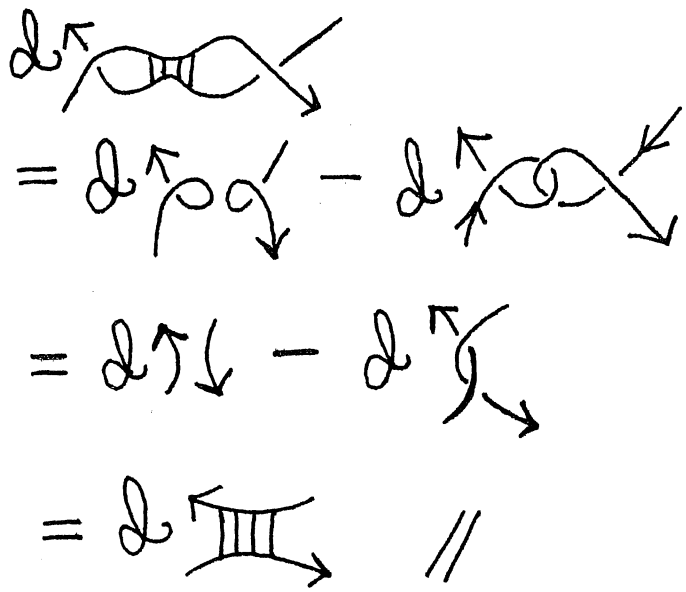
Lemma. Let $\mathcal{I}_K$ be an ambient isotopy invariant of oriented knots and links K. Define a function on foldings via



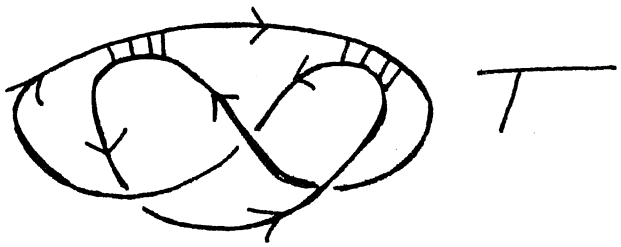
(This is formalized just as in our discussion of $\mathcal{I}$ ).

Then, for a folding F, $\mathcal{I}_F$ is an invariant of rigid-vertex isotopy of the folding.

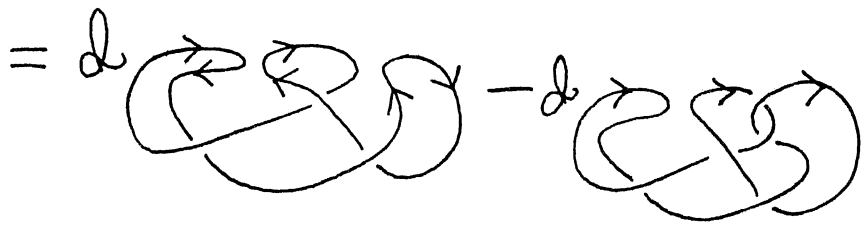
Proof. Again we just check one of the cases of rigid vertex twist:

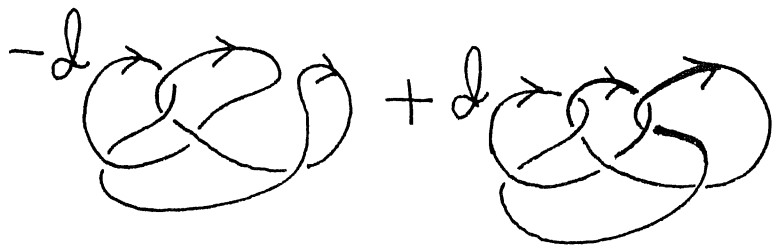


As an example, consider the embedding of the tertiary structure T shown below:

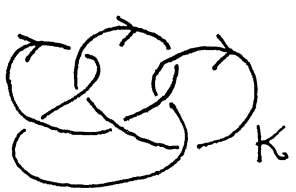


$\alpha_T =$ 

$=$ 

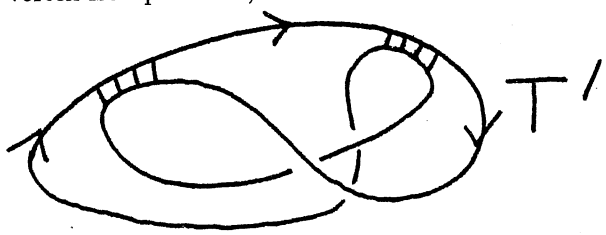


$= \alpha\bigcirc -$  $-$ 

$+$ 

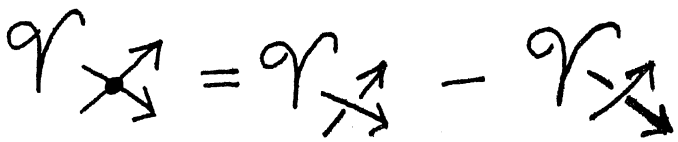
$[E \sim E^*$ (mirror image), but $K \not\sim K^*]$

Then by using the facts that the two knots shown in the above expansion are actually knotted and inequivalent to their mirror images, we conclude that T is not rigid-vertex isotopic to T',



nor is T rigid vertex isotopic to its mirror image. In this case, we have not assumed that our invariants were of finite type. Nevertheless, the formulation

corresponds directly, via our conventions, to the basic identity for Vassiliev invariants

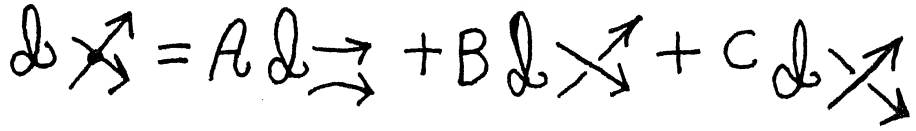


and hence, computations of $\mathcal{I}_K$ can be interfaced with computations of Vassiliev invariants.

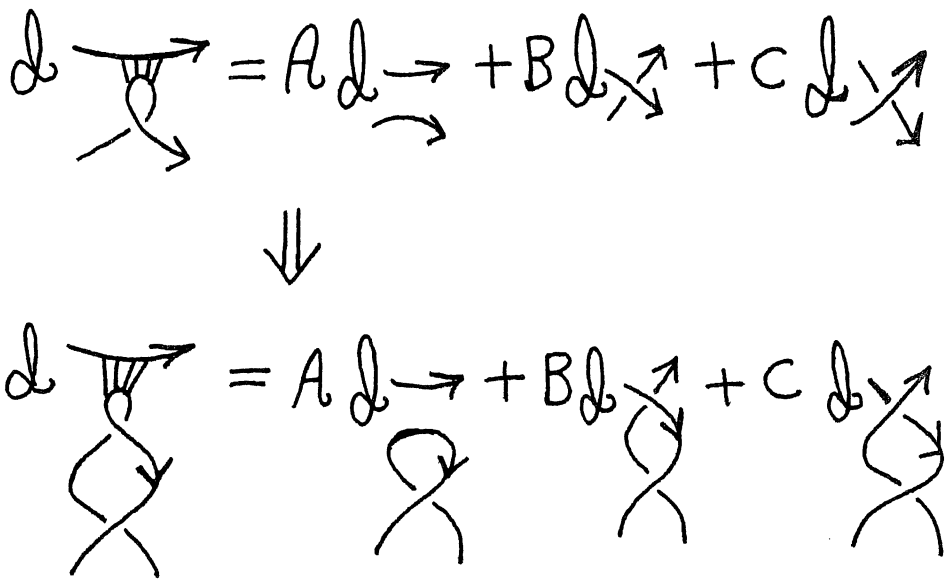
In the next section we shall supply more information about Vassiliev invariants. For the reminder of this section, we show how to reformulate our more general rigid vertex graph invariants [K4] for the case of protein folding.

4 Generalizing The Invariant $\mathcal{I}$.

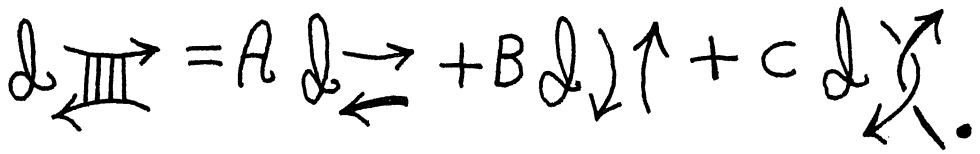
Let



define a graph invariant corresponding to a given oriented link invariant (as explained in [K4]). Then we can reformulate $\mathcal{I}$ in folding terms via:



Hence (assuming $\mathcal{I}$ on links is an ambient isotopy invariant)



This completes our description of the *generalized invariant of foldings*. It is quite useful for studying the topology of foldings via direct unfolding, recombination and linked recombination. If one knows either an ambient isotopy invariant $\mathcal{I}$ or the specific isotopy classes of the knots and links that occur through this process of resolution, then a great deal of information about the topology of the folded molecule is obtained as a consequence.

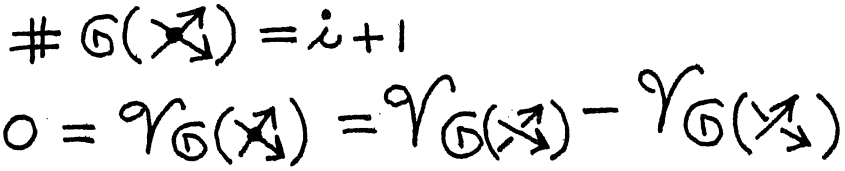
5 More about Vassiliev invariants.

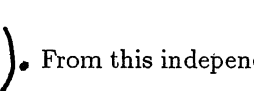
So far, except in the case of the Conway polynomial, we have not discussed the crucial matter of finite type in relation to graph invariants satisfying

Therefore, suppose that $\mathcal{V}$ is of type i. Let # G denote the number of 4-valent nodes in a graph G. Then we have the important

Fact. If $\mathcal{V}$ is of type i, then for # G=i, $\mathcal{V}$ is independent of the embedding type of G in R^3.

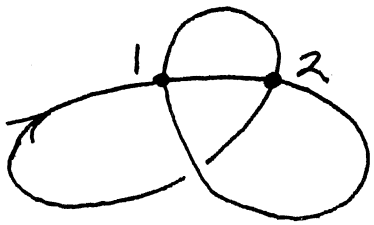
Proof. Suppose G is embedded with a crossing of the form  in its list of crossings. Then let $G($  $)$ denote this embedding and $G($  $)$ the embedding obtained by switching this given crossing. Then we have:



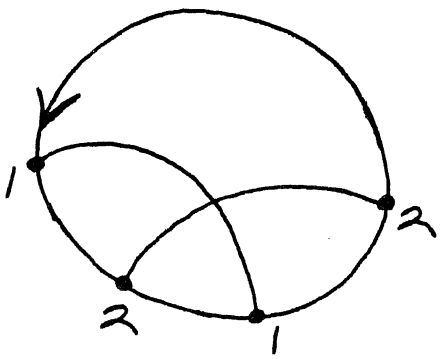
Thus  . From this independence of crossings, it follows

that $\mathcal{V}_G$ depends only upon the abstract graph G. This completes the proof. //

In this section we shall work entirely in the language of 4-valent nodes. Thus a diagram

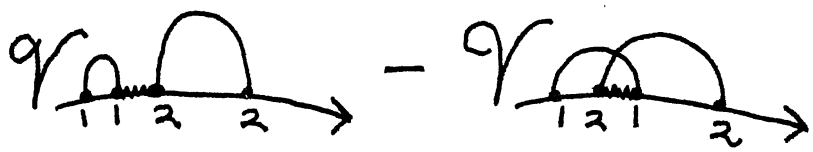


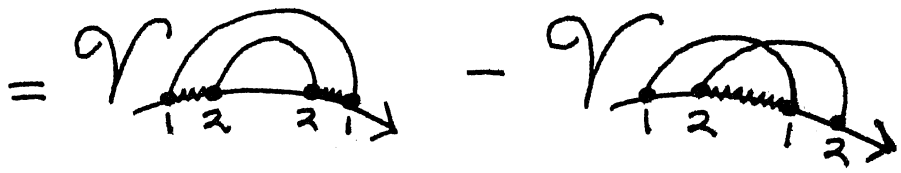
is an embedded graph with nodes labelled 1 and 2. The abstract structure of G is represented by the pairing diagram



This same diagram represents a folding structure. We see that *Vassiliev invariants of type i assign (topological) indices to abstract folding structures with i pairings.* These indices do not depend upon the embedding type of foldings with i pairings *and* they can be used to obtain information about embeddings of foldings with fewer than i pairings (as we did in the last section).

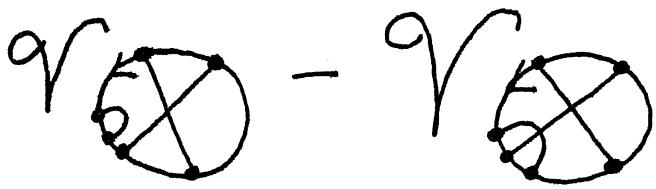
We shall call the assignment $\mathcal{V}_G$ of a Vassiliev invariant of type i to graphs with i nodes a *top row* of the Vassiliev invariant V. It turns out that the topology $[S]$ dictates a necessary and sufficient condition for these indices in the form of a set of relations. In terms of the top row these relations can be written symbolically as shown below:

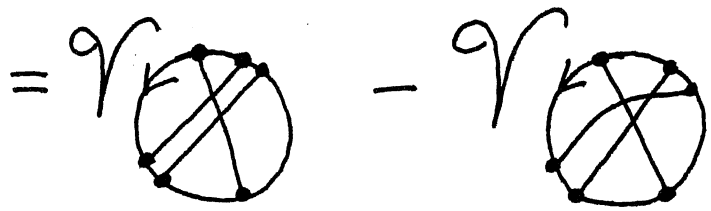




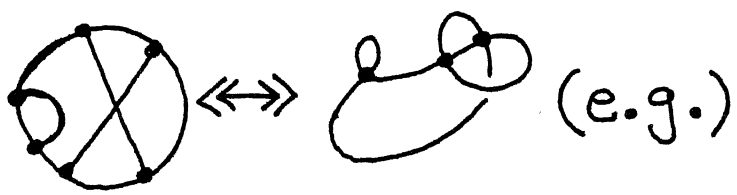
Here the wiggly line indicates that the indices 1 and 2 on either end of it are actually neighbors with no other intervening connections.

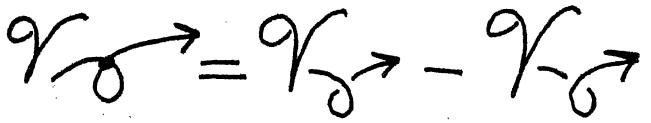
For example, at i=3 we have:



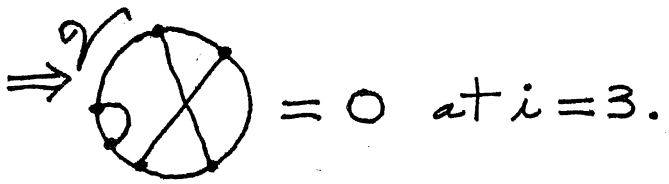


However, note that

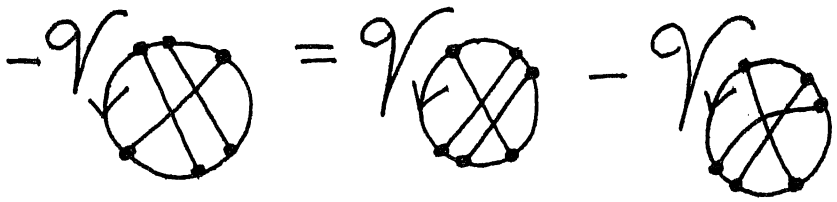




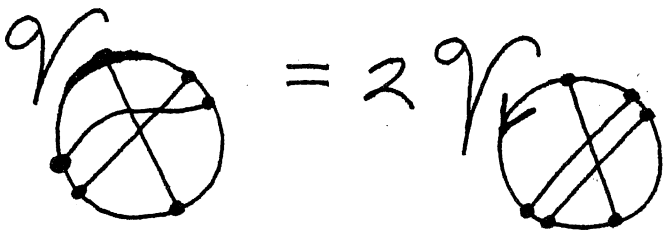
$= 0.$



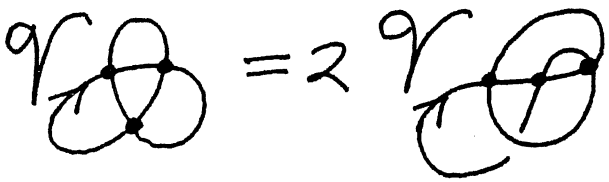
Thus



Hence



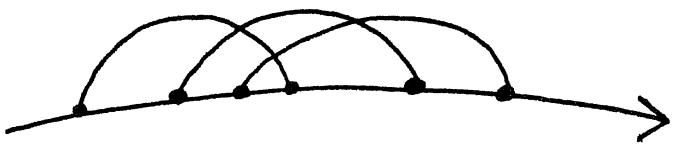
Whence



if  has *type 3*.

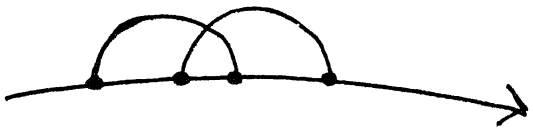
In terms of constructing Vassiliev invariants of type 3, this means that one only has to consider the 3-noded graph

or its corresponding folding

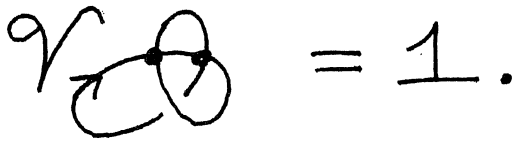


This makes Vassiliev invariants of type 3 rather easy to compute. We assign

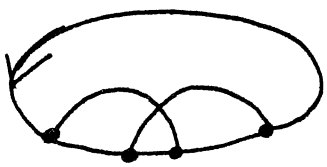
Now go back to 2-noded graphs such as



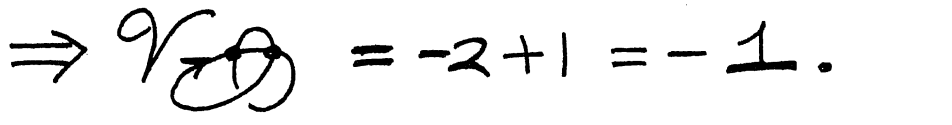
In fact, we see that abstractly this is the only 2-noded graph of relevance. Assign it the abstract value 1, and *define*



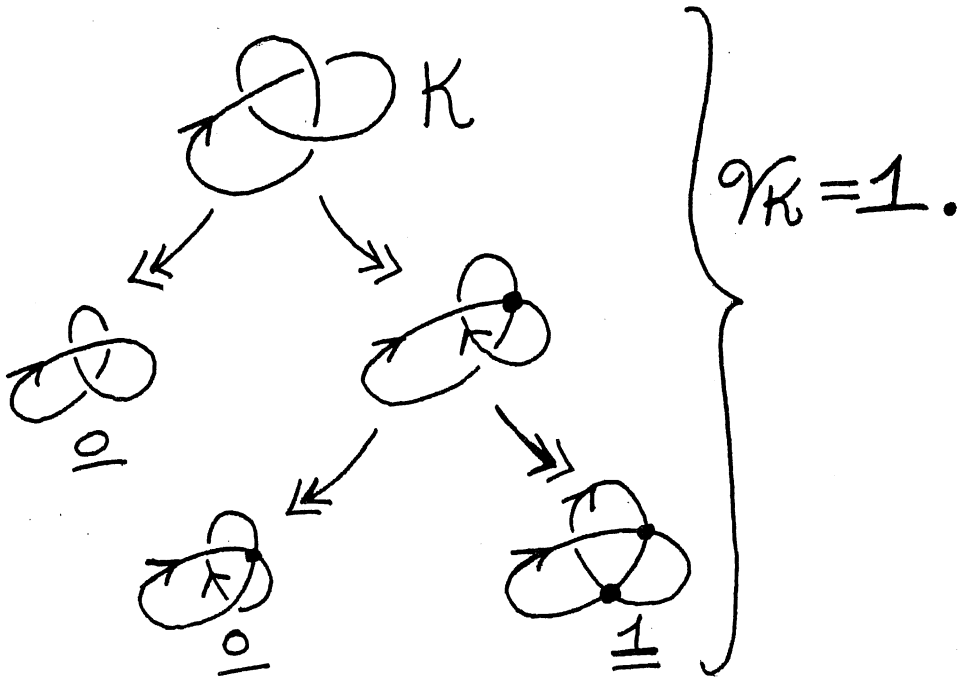
Then, any other embedding of



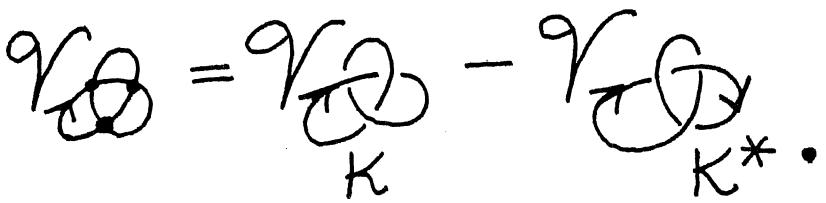
is determined by the values at level 3 and the switching relations. For example,



Similarly



Since it is easy to check that



This shows that this Vassiliev invariant of type 3 detects the topological difference between the trefoil knot K and its mirror image K^*.

By the same token, we have shown that the graph embeddings

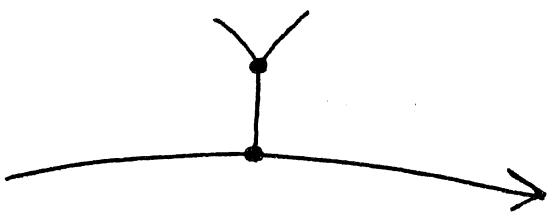
are not rigid vertex isotopic, and hence neither are the folded embeddings shown below isotopic.

There are non-trivial Vassiliev invariants of all orders. These can be obtained from the well-known skein polynomials via truncation of power-series substitutions. For example, we have the theorem of Birman and Lin.

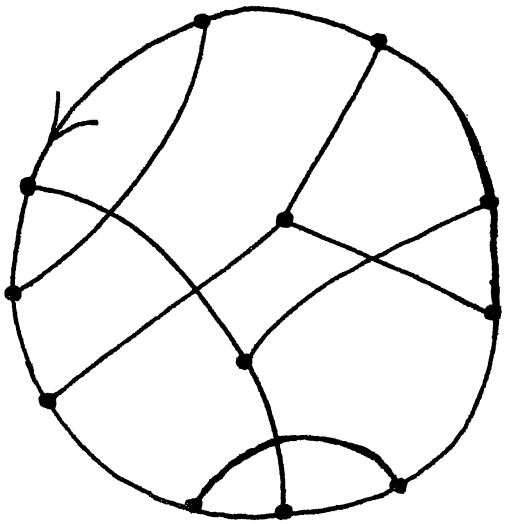
Theorem.[BL] Let $V_K(t)$ denote the original Jones polynomial [J] as a Laurent polynomial in K. Let $V_K(e^x) = \sum_{n=0}^{\infty} v_n(K)x^n$ be the power series resulting from substituting e^x for t. Then the coefficients $v_n(K)$ are Vassiliev invariants of type n.

Proof. See [BL] or [K5].

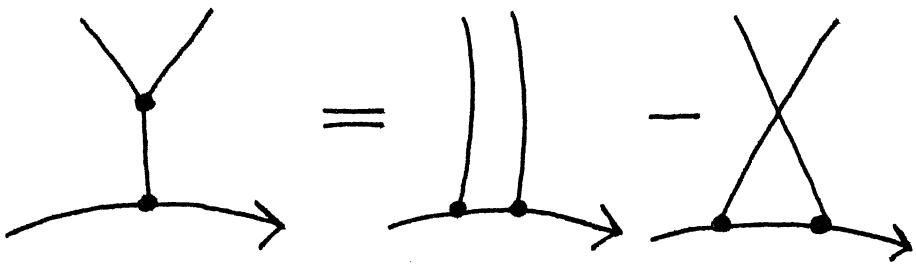
Another very striking result is the construction of Top Rows by Bar-Natan [BAR] via Lie algebra and "Feynman diagrams". In Bar-Natan's construction, the "chord diagrams" (our folding diagrams) are extended to allow a 3-valent interior vertex such as



or



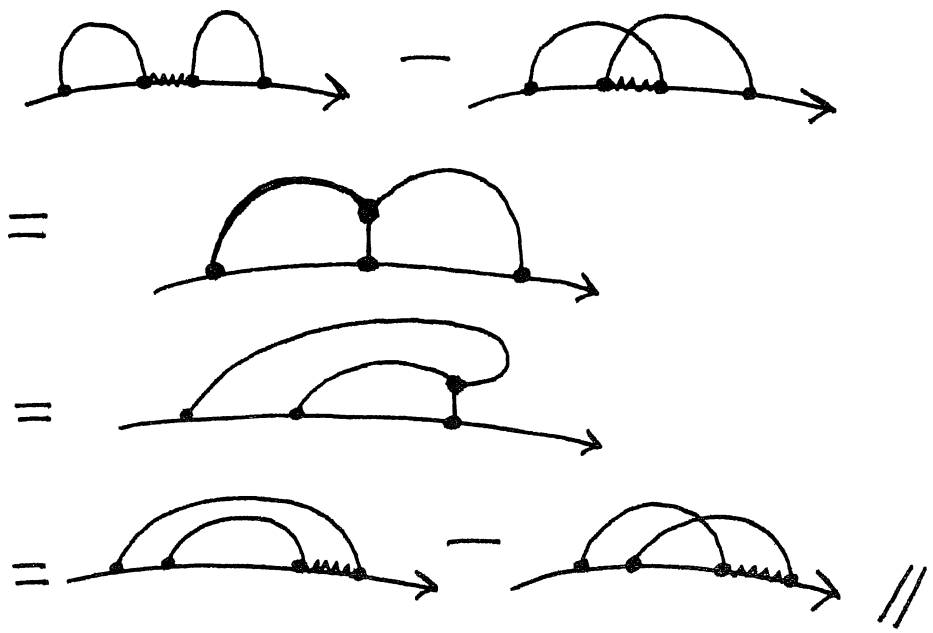
Bar-Natan takes as axiomatic the relation



Call this relation the *STU* relation.

Proposition. The STU relation implies the topological 4-term relation on top row diagrams.

Proof.



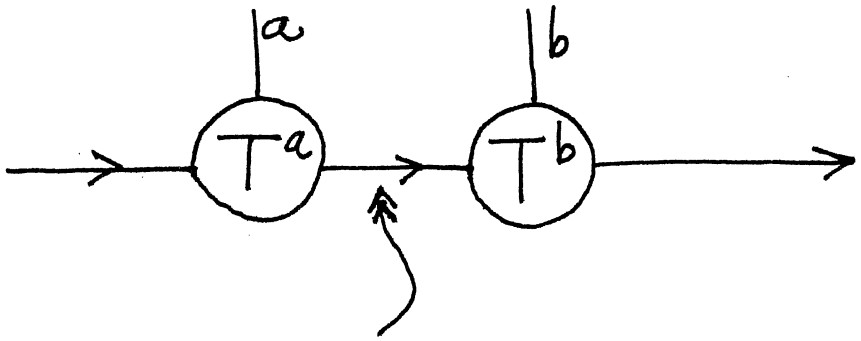
The fantastic thing about this observation is that (in the context provided by the Vassiliev invariants) it provides the core explanation why Lie algebras implicate topological invariants of knots, links and graphs! For the STU relation is actually an abstract way to state that in representing a Lie Algebra, *the representation of a commutator is the commutator of the representations.* A Lie algebra has a basis $\{T^a | a = 1, 2, ..., n\}$ and a basic commutator formula

$$[T^a, T^b] = f_c^{ab} T^c$$

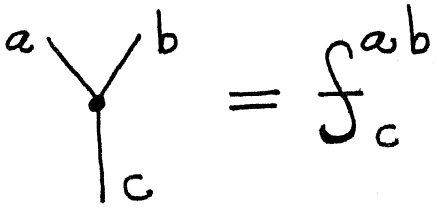
(sum on c), or

$$T^a T^b - T^b T^a = f_c^{ab} T^c$$

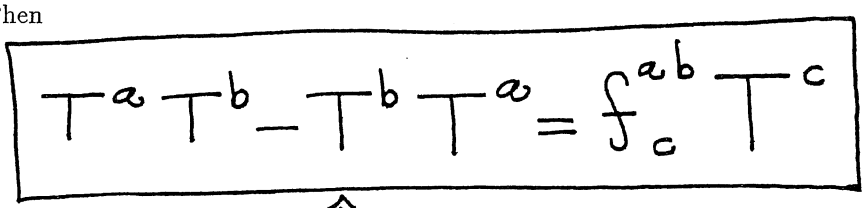
Now diagram this relation via

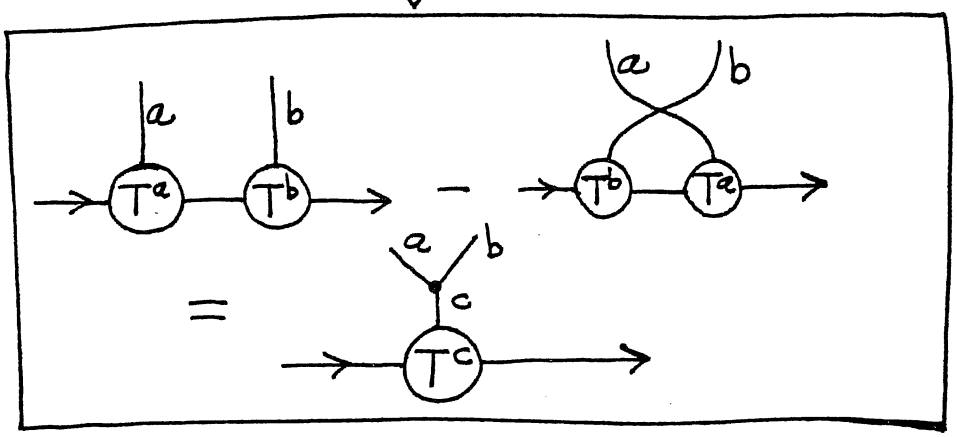


product of T^a with T^b.



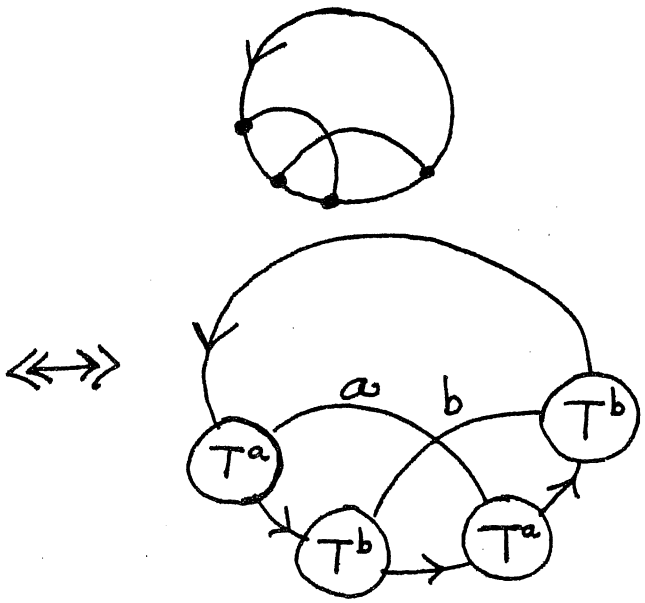
Then





Thus, we see that by appropriately labelling the trivalent diagrams with Lie algebra generators (or their representing matrices) we shall obtain weight systems that give rise to Vassiliev invariants (We have deliberately left out certain technicalities about the Killing form in the Lie algebra. See [BAR].)

For pairings themselves, the weights are obtained by arranging Lie algebra generators at the paired points and summing and tracing the corresponding matrices. Thus



$$\longleftrightarrow tr\left[\sum_{a,b} T^a T^b T^a T^b\right]$$

(tr denotes matrix trace.)

(again we have deliberately left out the Killing form). The approach sketched here is valid when f_c^{ab} is totally anti-symmetric in the three indices- such bases are available for the classical semi-simple Lie algebras.

It is fascinating to speculate on deeper relationships between molecular biol-

ogy and this pattern of assigning Lie algebra elements to the pattern of paired bases in a protein folding.

6 Discussion.

1. **4D-Interval and nucleotide algebra.** Self-splicing and RNA enzymatic activity suggest that not only secondary structure of RNA is biologically important, and under control, but secondary structure transitions are important as well ([PW], [MK]). In this paper secondary structures of RNA have been analyzed without particular analysis of nucleotide sequence. To study secondary structure transitions, it is necessary to find mathematical formulation of complementarity of nucleotides, as well as basepairing.

   As has been demonstrated recently (see [M] and [M2]), from the formal definition of nucleotides as two hierarchical negations, follows the representation of nucleotides by four 4×4 matrices of the form:

$$[a,b|c,d] = \begin{pmatrix} a & d & b & c \\ c & a & d & b \\ b & c & a & d \\ d & b & c & a \end{pmatrix} \tag{2}$$

   Namely, one pair of complementary nucleotides is represented by unit matrices $[1,0|0,0]$, $[0,1|0,0]$. Another pair is represented by matrices $[0,0|1,0]$ and $[0,0|0,1]$. One can check that **multiplication of four matrices**

$$\check{1} = [1,0|0,0], \quad -\check{1} = [0,1|0,0], \quad \check{i} = [0,0|1,0], \quad -\check{i} = [0,0|0,1] \tag{3}$$

   coincides with that of unit complex numbers 1, -1, i and $-i$ **correspondingly** (see [M2]).

   From the postulate that *complementary nucleotide matrices sum to zero* it follows, that

$$0_W = [1,1|0,0] \leftrightarrow 0 \qquad and \qquad 0_C = [0,0|1,1] \leftrightarrow 0 \tag{4}$$

   This condition is satisfied if one determines *four-dimensional complementary interval*, generated by linear combinations

   $[a,b|c,d] = a[1,0|0,0]+b[0,1|0,0]+c[0,0|1,0]+d[0,0|0,1]$ with non-negative coefficients a,b,c,d of unit matrices.

   Let the 4D-complementary interval be defined as follows:

$$\mathcal{I}(a,b|c,d) = \begin{cases} (a-b,0|c-d,0) = (a-b)\check{1} + (c-d)\check{i} & if \quad a \geq b \text{ and } c \geq d \\ (a-b,0|0,d-c) = (a-b)\check{1} + (d-c)(-\check{i}) & if \quad a \geq b \text{ and } d \geq c \\ (0,b-a|c-d,0) = (b-a)(-\check{1}) + (c-d)\check{i} & if \quad b > a \text{ and } c \geq d \\ (0,b-a|0,d-c) = (b-a)(-\check{1}) + (d-c)(-\check{i}) & if \quad b > a \text{ and } d > c \end{cases} \tag{5}$$

So defined, the 4D-interval is a 4×4 matrix. Multiplication and addition of intervals 5 is defined as usual matrix operation. It is rather straightforward to prove that the set of intervals 5 is a field. Moreover, 4D-intervals 5 is a commutative division algebra of the order 2 over the algebra of 2D-intervals

$$\mathcal{I}[a,b] = \begin{cases} (a-b)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & if\, a \geq b \\ (b-a)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \text{otherwise} \end{cases} \tag{6}$$

The set of all 2D-intervals, with multiplication and addition defined as corresponding matrix operations, is isomorphic to the algebra of real numbers [M2].

The set of all 4D-intervals, with multiplication and addition defined as corresponding matrix operations, is isomorphic to the algebra of complex numbers.

One can interpret the 4D-complementary interval geometrically (see details is [M2]). To any matrix $[a,b|c,d]$ there corresponds another matrix $P = I[a,b|c,d]$. The matrix P can be considered as a projection of matrix $[a,b|c,d]$. In this interpretation, the 4D-interval I is a projection operator. The set of matrices $P = I[a,b|c,d]$ with $a,b,c,d \geq 0$ constitutes a 4D-octant. Interval (projective) matrix P belongs to the first quadrant of one of four planes $[X,0|Y,0]$, $[X,0|0,Y^c]$, $[0,X^c|Y,0]$, $[0,X^c|0,Y^c]$. These **four mutually orthogonal quadrants with the projective operation 5 make a complex plane over the set of matices** $[a,b|c,d]$ **(with** $a,b,c,d \geq 0$**)** ! This surprising result proves the validity of complex number representation of nucleotides, proposed recently by Magarshak and Benham [MB].

2. **Complex numbers representation of nucleotides.** Now we sketch the approach of Magarshak and Benham [MB] for the representation of secondary structures. In this approach complementary base pairs G and C receive the labels 1 and -1, while A and T receive the labels i and $-i$ ($i^2 = -1$). In a diagram for the folding the bases are labelled 1,2...,n and an $n \times n$ matrix M is formed so that

$$M_{ij} = \begin{cases} -1 & \text{i is paired with j } (i \neq j) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The sequence of bases is mapped to a vector $\vec{v}$ with entries 1, -1, i and $-i$ as explained above. It is then easy to see that $M\vec{v} = \vec{v}$. Thus, *for a given nucleotide vector* $\vec{v}$ *the set of possible foldings is in 1-1 correspondence with the matrices* M *such that* $M\vec{v} = \vec{v}$.

Matrix M is a solution with eigenvalue 1 of the equation $M\vec{v} = \vec{v}$, where (what is unusual), vector $\vec{v}$ is given, and matrix M is unknown [MB].

The matrix M can be interpreted either as a structure matrix, or as a transition operation, which transforms free (i.e.totally dispaired) state of RNA molecule into the final secondary structure. More generally, one can define transition matrix $T = M_2 M_1 = M_2 M_1^{-1}$ which transforms initial secondary structure M_1 into the final secondary structure M_2. A well formulated account of this method is given in [MB].

Vassiliev knot invariants, analyzed in this paper, can be interpreted as these of solutions of the equation $M\vec{v} = \vec{v}$. In a sequel paper the topological invariants of transition matrices and secondary structure dynamics will be studied.

3. Non-Watson-Crick basepairs and Hermitian forms. The theory presented in the present paper is developed specifically for the analysis of the RNA and DNA structures with Watson-Crick basepairing. But it can be generalized. If non-Watson-Crick nucleotides U and G (number j and k) are paired, the structure matrix S becomes Hermitian. Let the transition from the secondary structure S_1 to the secondary structure $S_2 = T_{21} S_1$ be performed. The transition matrix of the inverse process T_{21} is complex conjugate to the matrix T_{12}^*, so any transition matrix $T_{21} = S_2 S_1^{-1} = S_2 S_1$ is unitary.

Let analyze Hermitian form

$$\mathcal{H}(S) = \mathrm{g}^\dagger S \mathrm{g} = \sum_{i=1}^{n} \sum_{k=1}^{n} S_{ik} g_i^* g_k \tag{8}$$

If complementary nucleotides g_i and $g_k = g_i^c$ are *real*, i.e. equal to ± 1, then element (i,k) of the Hermitian form 8 $\mathcal{H}(S)_{ik} = \mathcal{H}(S)_{ki} = S_{ik} g_i^* g_k = S_{ki} g_k^* g_i = (-1)(-1)1 = 1$. If complementary nucleotides g_i and $g_k = g_i^c$ are *imaginary*, i.e. equal to $\pm\imath$, then any nonzero element (i,k) of the Hermitian form 8 $\mathcal{H}(S)_{ik} = \mathcal{H}(S)_{ki} = 1$. We see that in any case the Hermitian $\mathcal{H}(S)$ is positive. Moreover, for arbitrary nucleotide sequence g and any secondary structure S, which this sequence can form, $\mathcal{H}(S) = const = n$, i.e. is just equal to the number of nucleotides in the chain. Sure, this trivial model can be modified. We begin with definition of the matrix $\mathcal{E}$ such that:

i. element (k,k) is equal to the energy E_k, associated with non-paired nucleotide number k,

ii. element (j,k) is equal to $-(E_j + \Delta E_{jk}/2)$, if and only if nucleotides j and k are paired, and

iii. all other elements of matrix $\mathcal{E}$ are equal to zero.

The Hermitian form

$$\mathcal{H}(\mathcal{E}) = \mathrm{g}^\dagger \mathcal{E} \mathrm{g} = \sum_{i=1}^{n} \sum_{k=1}^{n} \mathcal{E}_{ik} g_i^* g_k \tag{9}$$

is increasing by the ΔE_{jk} if the bond (jk) is making. This property can be used for calculation of a *partition function* of RNA secondary structure formation.

4. Partition function of complementary structures (Co-partition).

   Finally, we wish to point out here that the Magarshak-Benham approach to foldings gives rise to a natural partition function of the form

$$Z_{\vec{v}} = \sum_{M:M\vec{v}} = \vec{v}e^{-\frac{1}{kT}E(M)} \tag{10}$$

   where $E(M)$ is appropriately chosen energy functional for the foldings, and T is temperature, k Bostzmann's constant. For example, one may take $E(M) = 3[G-C] + 2[A-U]$ where $[G-C]$ denotes the number of $G-C$ pairs while $[A-U]$ denotes the number of $[A-U]$ pairs. (The 3 and the 2 are the number of hydrogen bonds needed in each case)[1].

   Analysis of the algebra, generated by nonconventional basepairs, goes out of the frame of this contribution and will be discussed elsewhere. Partition function analysis also will be the subject of a sequel to this paper.

5. In this paper the 3-dimensional disposition of atoms was not introduced directly. But the model can be modified. Namely, one can introduce a complementary field, which acts on complementary nucleotides only [M2]. This is a short-range force field, responsible for basepairing. In complex-numbers designations, complementary force is proportional to $\delta_{g_i,-g_k}$, where g_i and g_k are complex values of nucleotides number i and number k. In other words, compelmentary force attracts nucleotides, if and only if they satisfy the trivial equation:

$$g_i + g_k = 0 \tag{11}$$

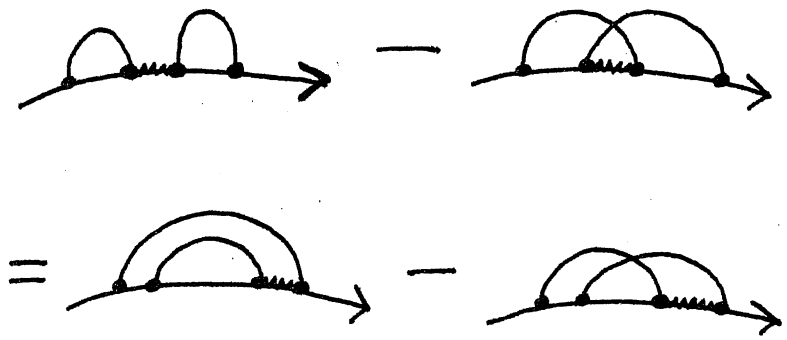
   and orients them in accord with their disposition in DNA duplex. One must include the complementary force into the model in addition to the force fields, typical for double-stranded DNA (see, for instance, [SO]). So there is a hope to compute 3-dimensional structures of one-stranded RNA, as well as RNA secondary structure formation.

6. Topological code. Topological properties of equations, which describe some key biological processes, often determine the solutions of these equations. For instance, it is known that solution of equations of stationary enzyme kinetics is determined by the graph of corresponding enzyme reaction (see, for instance [VG] and [VM]). The same statement is true for electron transfer rate in proteins and some other biologically important macromolecules

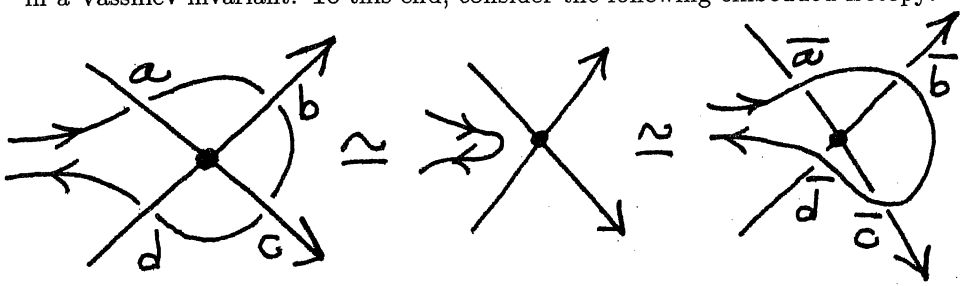
[1] We thank Nancy Wood for suggesting this particular energy function.

(see [OBG] and [MMJ]). In this paper, a relationship between RNA secondary structure and Vassiliev polynomials has been found. So the hypothesis, that **one of the languages, which nature uses in vivo, is topological language**, seems to be reasonable.

Appendix 1. It is the purpose of this appendix to give a quick derivation of the necessity of the 4-term relation (see [S])



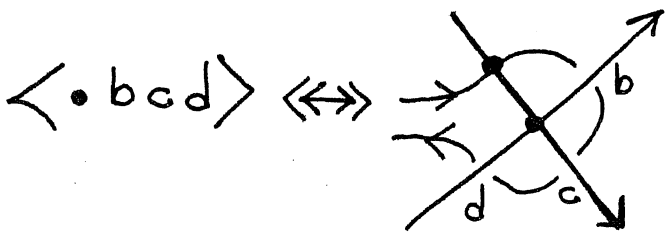
in a Vassiliev invariant. To this end, consider the following embedded isotopy:



Let us write the symbol $\langle abcd \rangle$ for the Vassiliev invariant of the left-hand configuration, and $\bar{x}$ for the switch of a given crossing labelled x. Then we can write

$$\langle abcd \rangle - \langle \bar{a}bcd \rangle = \langle \bullet bcd \rangle$$

where $\langle \bullet bcd \rangle$ denotes the replacement of the crossing a by a 4-valent node ($\bullet$).



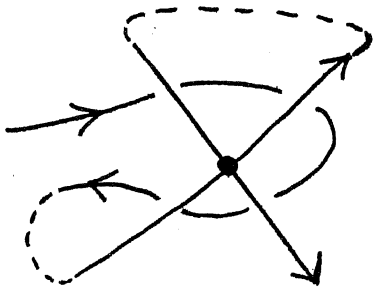
Then we have the equations

$$\begin{cases} \langle abcd \rangle - \langle \bar{a}bcd \rangle = \langle \bullet bcd \rangle \\ \langle \bar{a}bcd \rangle - \langle \bar{a}\bar{b}cd \rangle = -\langle \bar{a} \bullet bcd \rangle \\ \langle \bar{a}\bar{b}cd \rangle - \langle \bar{a}\bar{b}\bar{c}d \rangle = -\langle \bar{a}\bar{b} \bullet d \rangle \\ \langle \bar{a}\bar{b}\bar{c}d \rangle - \langle \bar{a}\bar{b}\bar{c}\bar{d} \rangle = \langle \bar{a}\bar{b}\bar{c} \bullet \rangle \end{cases} \tag{12}$$

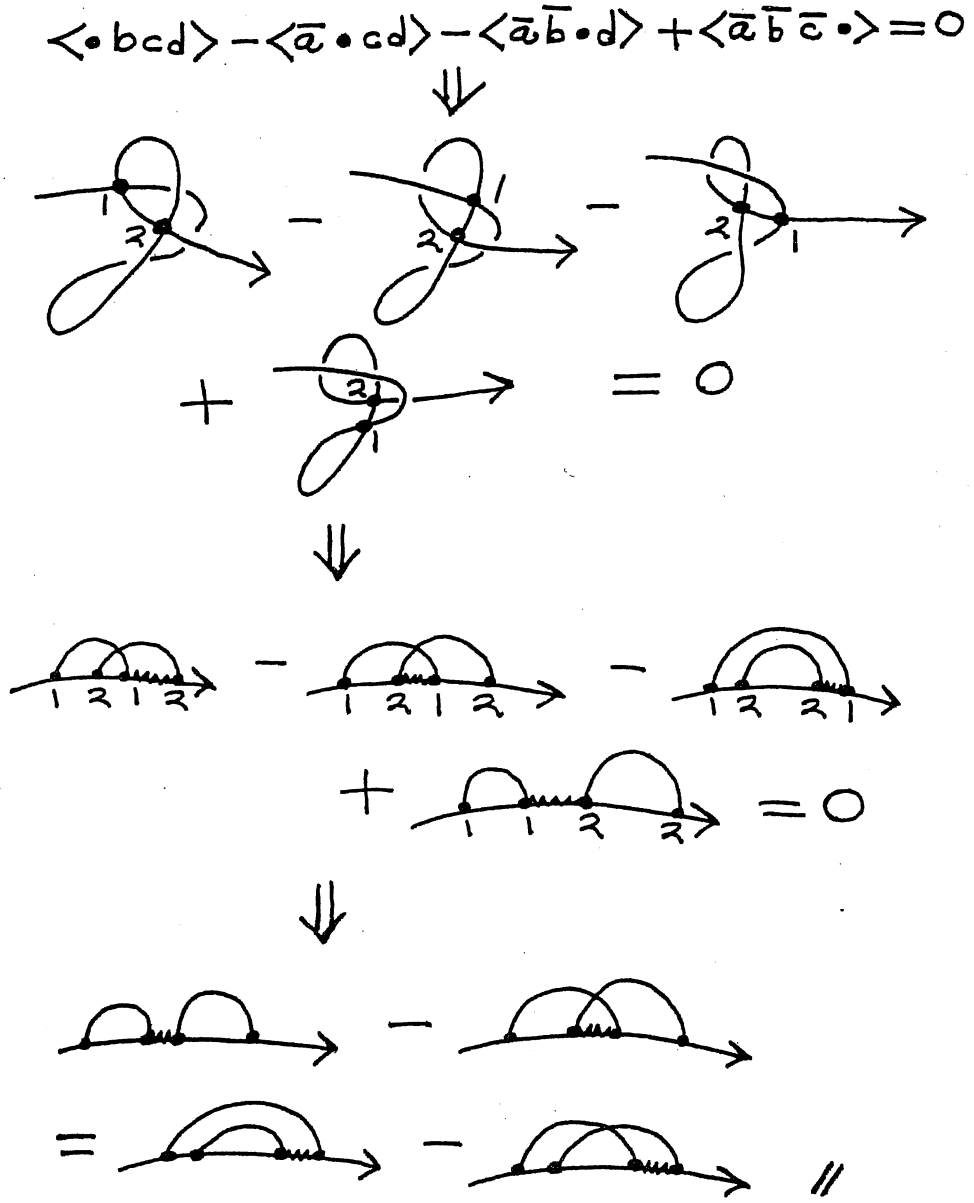
Since the isotopy shows that $\langle abcd \rangle = \langle \bar{a}\bar{b}\bar{c}d \rangle$, we conclude that

$$\langle \bullet bcd \rangle - \langle \bar{a} \bullet bcd \rangle - \langle \bar{a}\bar{b} \bullet d \rangle + \langle \bar{a}\bar{b}\bar{c} \bullet \rangle = 0$$

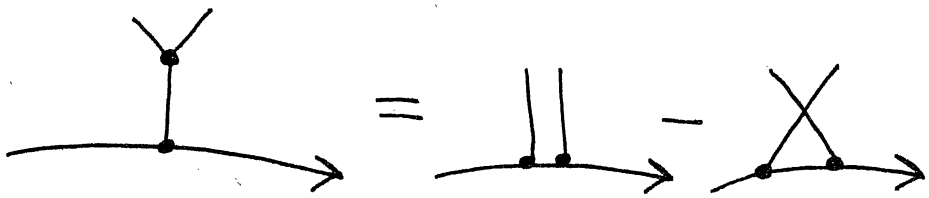
This relation translates into the 4-term relation on a top row. To see this, choose a given external connectivity for the diagram $\langle abcd \rangle$:



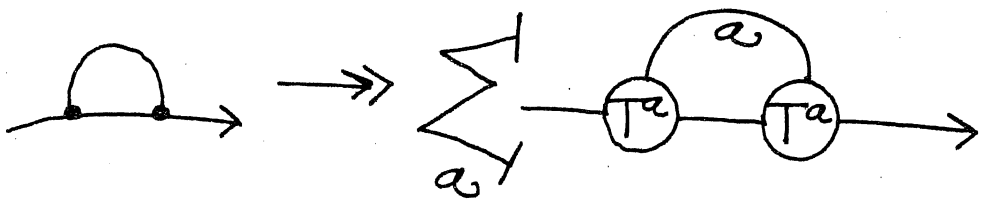
and translate into chord diagram language:



We must emphasize once again that it is quite extraordinary how this simple topological requirement is tied so directly to the Lie algebra patterns (see end of the last section of this paper) via the intermediate STU identity



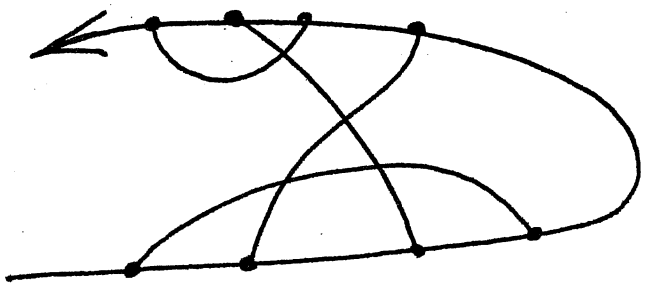
It should also be mentioned that



is the Casimir insertion into the Wilson line [K1] appropriately discretized for the content of combinatorial link invariants.

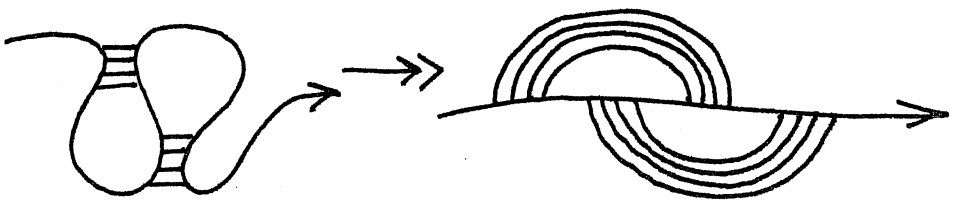
Appendix 2: Complexity and Proximity of Foldings.

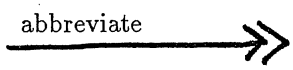
In this appendix we make some remarks about the possible use of a braided generalization of the Brauer monoid for measuring the proximity of a folded RNA molecule to its unfolded version. In other words, we are interested in a measure of the complexity of the folding. Recall that in section two we have represented abstract foldings via patterns such as



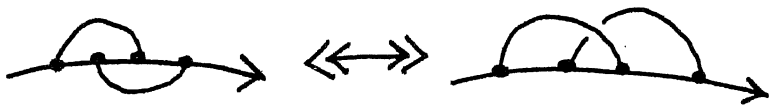
giving a correspondence with elements in the Brauer monoid.

An embedded folding can be (if the chain itself is unknotted) "pulled" into Brauer monoidal form, but the monoid strands may be woven about one another as in the simple pull of a pseudo-knot shown below.

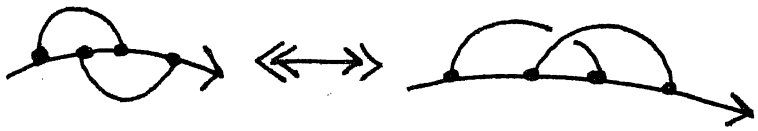




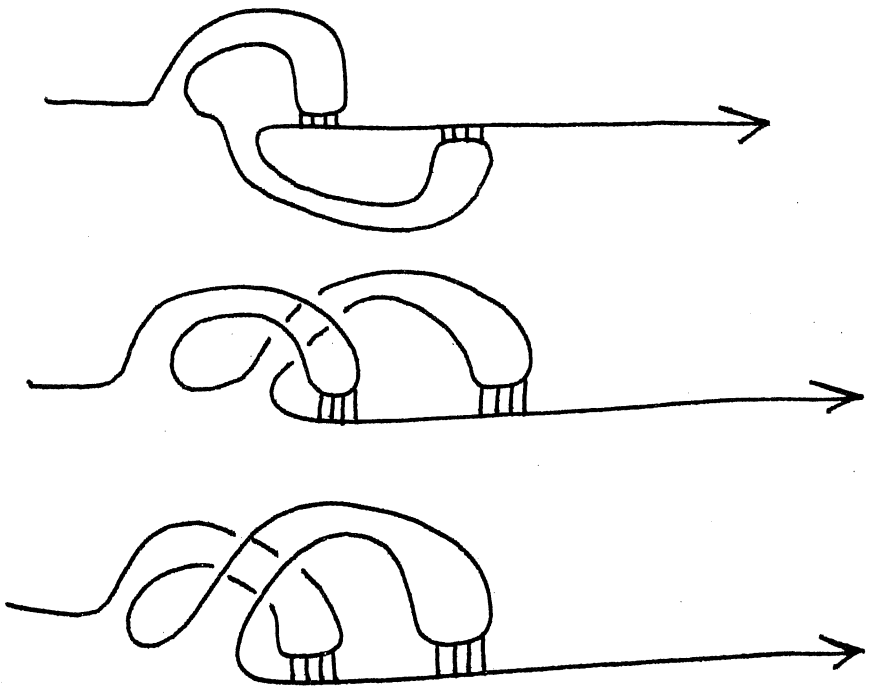
It is clear that it makes sense to measure the complexity of the pseudoknot by including the braiding structure of the attaching arcs. However, in some cases, more than one braiding structure will correspond to the same structure of attachment. For example, in the case above we have



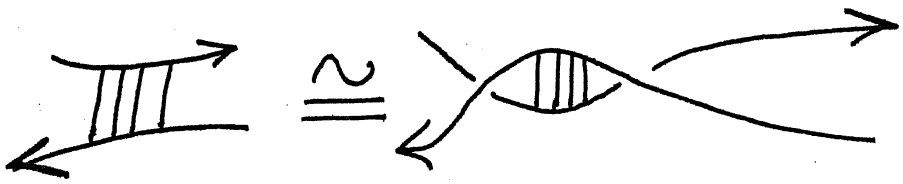
and also



(by swinging the given arc in one of the two possible directions around the oriented axis). This is a reflection of the fact that the following three pseudo-knots are almost isotopic:



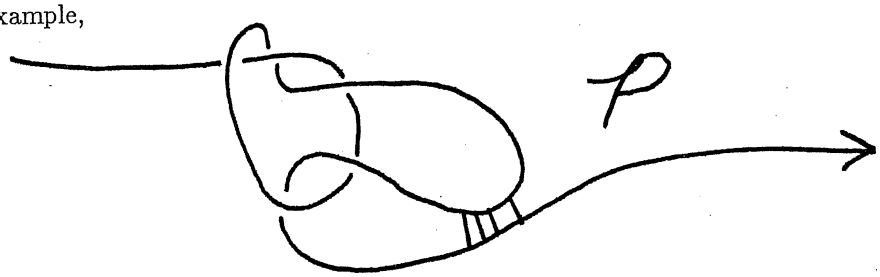
We say almost isotopic, because the three pictured pseudo-knots are isotopic up to twists of the sort shown below in a regular isotopy:



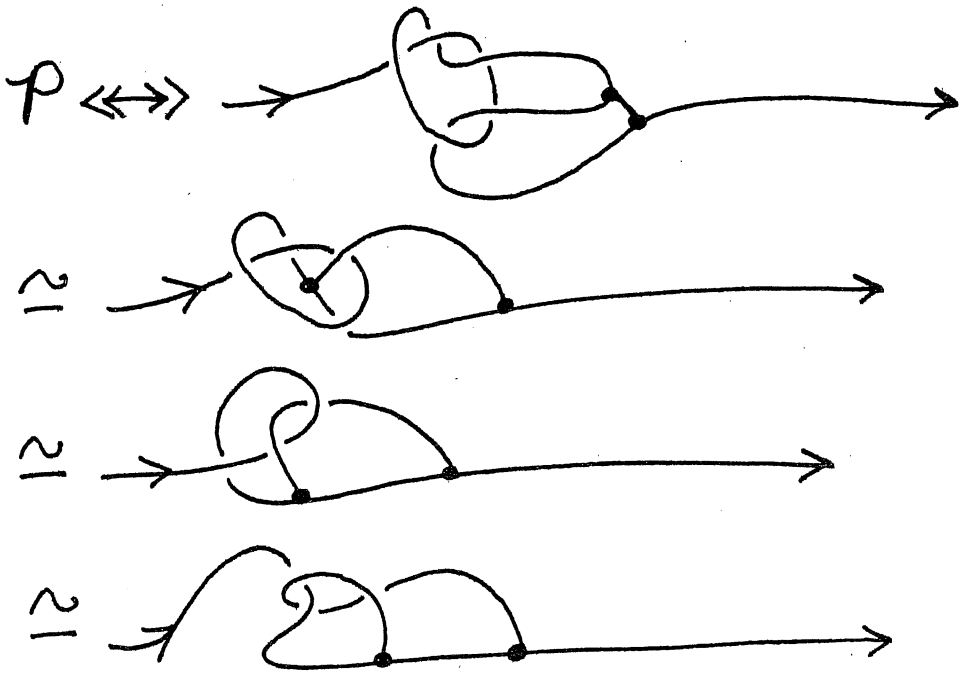
There are some interesting subtleties in even the simplest examples. For example, call a pseudoknot simple, if it gives rise to an unknotted embedding of its axis when all the pairing nodes are eliminated:

The standard pseudo-knot is simple:

More complex examples of simple pseudo-knots are easily manufactured. For example,



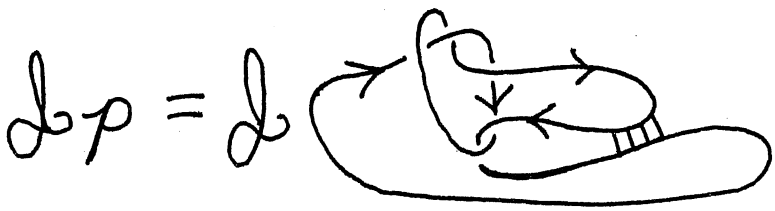
$\mathcal{P}$ is a simple pseudoknot with only one (multiple) pairing node. Translating into the language of attaching arcs, we have

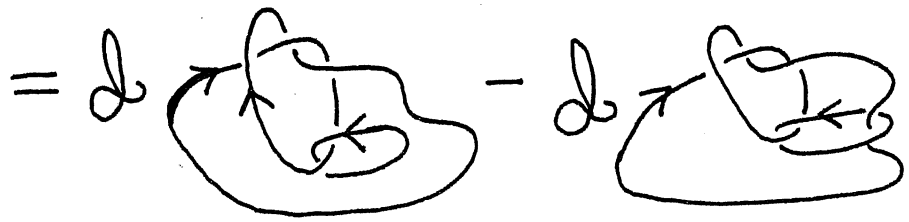


In this last representation we have an unknot · axis, but the attaching arc is entangled with the axis. This weaving reflects the complexity of the pseudoknot $\mathcal{P}$.

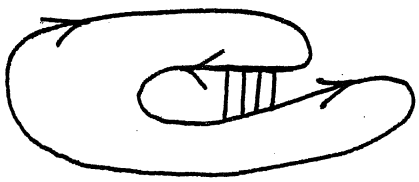
One way to analyze the complexity of $\mathcal{P}$ is to compute the invariant $\mathcal{I}_{\mathcal{P}}$ defined via

(see section 3 of this paper). This is the analogue of the Vassiliev invariant for folding. We obtain:



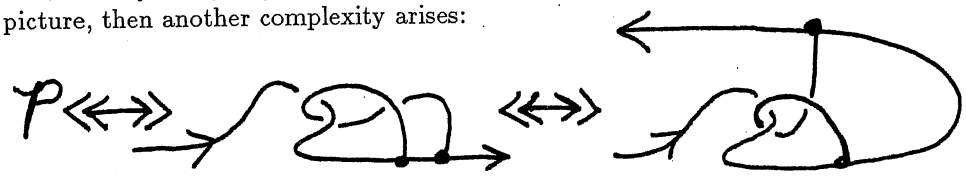


The two links obtained in this resolution are non-trivial and so we see that this pseudoknot is quite distinct from



and in fact it is quite complex. For example, it is topologically distinct from its mirror image.

If we try analyzing $\mathcal{P}$ via the braid monoid (generalized Brauer monoid) picture, then another complexity arises:



The entwinement of the axis with the attaching arc yields a new phenomenon for articulation. We can capture this phenomenon by adding to the vocabulary of the braid monoid top and bottom fixture elements of the form:

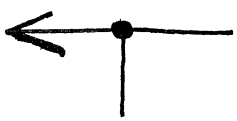
$= D_1 D_3 S_4 D_5 D_7 S_8 S_9$

or

$= d_1 d_3 s_4 d_5 d_7 s_8 s_9$

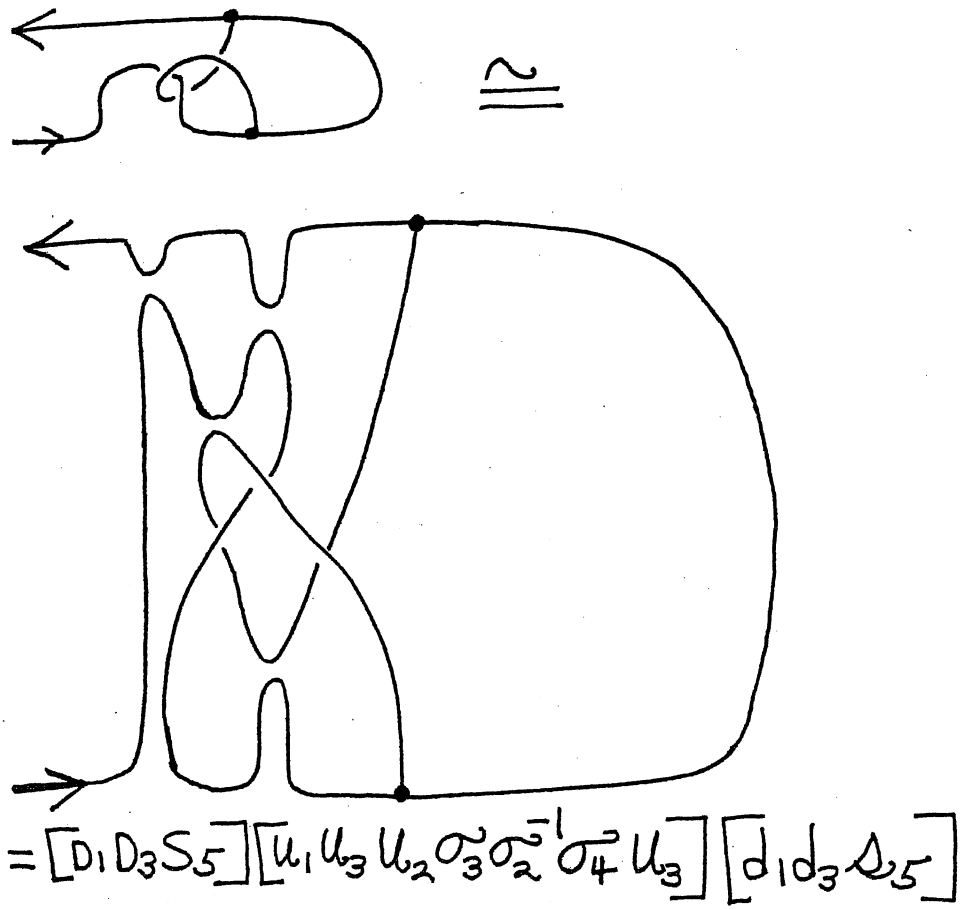
The basic parts of a top fixture consist in double strands such as

and an attachment as in

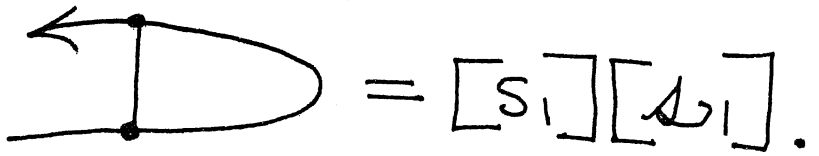


We use lower case letters to describe bottom fictures.

Then, any simple pseudo-knot can be expressed in the form of a product:
(Top Fixture)(Word in Braid Monoid)(Bottom Fixture). For example,

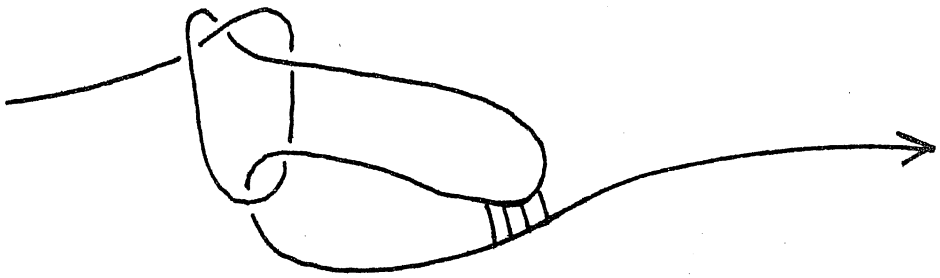


A simpler example is



The difference in complexity of $[S_1][s_1]$ and
$[D_1D_3D_5][U_1U_3U_2\sigma_3\sigma_2^{-1}\sigma_4U_3][d_1d_3s_5]$ can be regarded as a measure of the "topological proximity" of the foldings

and



We have concentrated in this appendix, on the possibility of measuring the proximity of different simple pseudoknots because this is a topic that can have actual application in microbiology. A long strand molecule has a high probability for unknotted self- entanglement. If this molecule can undergo self-binding (as in RNA) then some of these unknotted states of the strand can become simple pseudoknots of varying degrees of complexity.

REFERENCES.

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

[BAX] R.J.Baxter. *Exactly Solved Models in Statistical Mechanics.* Academic Press (1982).

[BL] J.S.Birman and X.Lin.*Knot polynomials and Vassiliev's invariants. Columbia Univ.preprint* (1991).

[OBG] J.N.Onuchic, D.N. Beratan and H.B.Gray. *Pathway Analysis of protein electron-transfer reactions.* Ann. Rev. Biophys. Biomol.Struct. 21, (1992) pp. 349-377.

[BR] R.Brauer. On algebras which are connected with the semisimple continuous groups. Annals of Math. Vol.58, N^0 4, October (1937).

[J] V.F.R.Jones. *Hecke algebra representations of braid groups and link polinomials.* Ann.of Math. 126 (1987) pp.335-388.

[K1] L.H.Kauffman.*Knots and Physics.* World Sci.Pub. (1991).

[K2] L.H.Kauffman. *State models for link polynomials.* L'Enseignment Math.36 (1991) pp.1-37.

[K3] L.H.Kauffman. *An invariant of regular isotopy.* Trans. Amer.Math. Soc., Vol. 318, N^o 2, (1991), pp.417-471.

[K4] L.H.Kauffman. *Invariants of graphs in 3-space.* Trans. Amer. Math.Soc. Vol. 311, 2 (Feb.1989) pp.697-710.

[K5] L.H.Kauffman. *Vassiliev Invariants and the Skein Polynonials*, (to appear).

[KV] L.H.Kauffman and P.Vogel. *Link polynomials and a graphical calculus.* Journal of Knot Theory and its Ramifications. Vol.1 N^o 1 (March 1992). pp.59-104.

[MK] Mironov,A.A and Kister,A.E. *RNA secondary Structure Formation during TranscriptionJ. Biomol.Str. Dyns.* 1986,v. 4, pp. 1-9

[KMM] A.Kister, Y.Magarshak, J.Malinsky, *The Theoretical Analysis of the process of RNA molecule self-assembly, in book Computer Genetics,* Elsevier Science Publishers (**1993**).

[M] Y.Magarshak, *Quaternion Representation of RNA Sequences and Tertiary Structures.* in book *Computer Genetics*, editors P.Pevsner and M.Gelfand, Elsevier Science Publishers, (**1993**).

[M2] Y.Magarshak. *Nucleotide algebra and nucleic relativity.* To appear.

[MB] Y.Magarshak and C.J.Behnam. An algebraic representation of RNA secondary structures. J of Biom.Str. & Dyn. ISSN 0739-1102, Vol.10, N^o 3 (1992) pp. 465-488.

[MMJ] Y.Magarshak, J.Malinsky & A. Joran. Diagrammatic techniques for solving Schwinger-Dyson equation: Electron transfer pathways in biological molecules, J.Chem Phys., 95,1 (1991), pp.418-432.

[PW] R.C.Penner and M.S.Waterman. Spaces of RNA Secondary Structures (to appear in Advances in Mathematics).

[PI] S.Piunikhin. Weights of Feynman diagrams, link polynomials and Vassiliev knot invariants (to appear in Journal of Knot Theory and its Ramifications).

[S] T.Stanford. Finite-Type invariants of knots, links and graphs (preprint 1992).

[SO] T.Schlick and V.K.Olson. *Supercoiled DNA energetics and Dynamics by Computer Simulation.* J.Mol.Biol. 223, (1992), pp.1089-1119.

[V] V.A.Vassiliev. *Cohomology of knot spaces in Theory of Singularities and Its Applications* (ed.by V.I.Arnold), Adv. in Soviet Math., Vol.1, AMS (1990).

[VG] M.V.Volkenstein and B.N. Goldstein.*A new method for solving the problems of the stationary kinetics of enzymological reactions.* Biochimica et Biophysica acta, 115, (1966) pp.471-477.

[VM] M.V.Volkenstein, Yu.B.Magarshak.*Application of the diagrammatic method of solving kinetic equations of cyclic enzymatic reactions.* Biophysics, 15, 5 (1970), , pp.805-813.