Rose-Hulman Undergraduate Mathematics Journal

Volume 7
Issue 1

Article 13

Modeling DNA Using Knot Theory:￼ An Introduction

Jenny Tomkins
University of Texas at Tyler, jennytompkins@yahoo.com

Modeling DNA with Knot Theory: An Introduction

Jenny Tompkins

Advisors: Drs. Casey Mann and Jennifer Mcloud-Mann
University of Texas at Tyler
Undergraduate Research Project
Summer 2005

1 Introduction

In recent years, exciting new applications of mathematics to the field of molecular biology have been developed. In particular, knot theory gives a very nice way to model DNA recombination. The relationship between mathematics and DNA began in the 1950's with the discovery of the helical Crick-Watson structure of duplex DNA. The discovery of this model opened the door for mathematical analysis of DNA. One such mathematical model is the Tangle Model for Site-Specific Recombination, which was first introduced by De Witt Sumners [10]. This model uses knot theory to study enzyme mechanisms.

Knot theory is a subset of a larger branch of mathematics called topology. *Topology* is an area of mathematics which involves studying the properties of geometric figures which are unaltered by elastic deformations such as stretching or twisting. To a topologist, a sphere is the same as a cube, and a doughnut is the same as a coffee cup. *Knot theory* is an area of topology that deals with knots and links. A *knot* is a closed curve in space with no self-intersections (i.e. a knot is a simple closed curve). In layman's terms, a knot is a piece of string, tangled or not, whose ends are connected.

Knots first received scientific attention in the 1880's when Lord Kelvin hypothesized that all matter was made of a substance called ether and that atoms are knots in the ether. This began the first attempts at classification of knots and general understanding of knots in the mathematical sense. Once modern atomic theories were formulated, physicists and chemists lost interest in knot theory, but mathematicians were intrigued by the study of knots and continued in the field despite its lack of applications at the time. It wasn't until the 1980's that applications of knot theory in molecular biology were discovered.

The purpose of this article is to explain the details of this application of knot theory to DNA recombination. Of course, the reader is aware that DNA are long, thin molecules found inside the nucleus of a cell; these molecules are nature's way of encoding biological traits and are the mechanism for reproduction. To get a sense of the scale of things, imagine the cell nucleus as the size of a basketball. Inside a nucleus of that size, you would find that the DNA would resemble thin fishing line with 200 km packed inside. Because the DNA is so tightly packed into such a confined space, it is not surprising that it is a tangled and knotted mess. DNA must be topologically manipulated in order for vital processes such as replication, transcription, and recombination to take place. Nature's answer to the tangling problem is enzymes.

Enzymes act by manipulating DNA in several different ways. They may cause coiling up of DNA (*supercoiling* – Figure 1). They may switch a crossing of nearby strands of DNA (transient enzyme-bridged break – Figure 2), or they may break apart a pair of strands and recombine them to different ends (recombination – Figure 3). The last of these is a process called site-specific recombination which will be discussed further in Section 6.

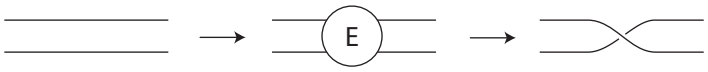

Figure 1: An enzyme (E) induced supercoil

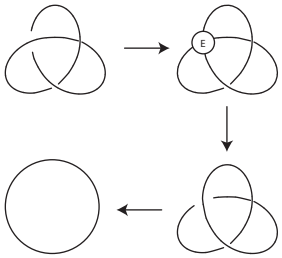

Figure 2: An enzyme (E) induced crossing change in a DNA trefoil knot

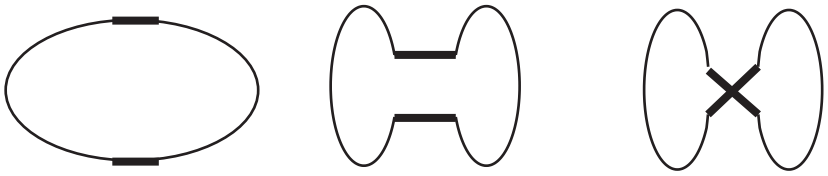

Figure 3: Recombination

This paper is organized as follows:

- Section 2 will introduce tangles, which will be our model for enzymes.
- Section 3 will discuss several operations which can be performed on tangles. These operations will be used to model the enzymatic actions on DNA.
- The DNA molecules themselves can be modeled using 4-plats or rational tangles, which will be discussed in Section 4.
- The theorems used to solve tangle equations will be explained in Section 5.
- Section 6 will introduce site-specific recombination.
- Section 7 will describe the tangle model for site-specific recombination.

- The last section we will look at an example where the tangle model has been successfully applied to analyze Gin site specific recombination.

2 Tangles

A *2-string tangle* is a pair (B, t), where B is a 3-ball and t is a pair of unoriented arcs (strings) properly embedded in B so that the end points of the arcs go to a specific set of 4 points on the equator of the ball (usually labeled NW, NE, SW, SE). A *tangle diagram* is the projection of the tangle on the plane of the equator as in Figure 4. We will label the endpoints in the diagram NW, NE, SW, SE. *Rational tangles* are defined as the family of tangles that can be transformed into the trivial tangle (see Figure 4b) by a sequence of twisting of the endpoints. Because rational tangles look like what is seen when studying DNA micrographs, they will be our focus in the paper. One should know that there are tangles that cannot be obtained in this fashion; they are the *prime tangles* and *locally knotted tangles*.

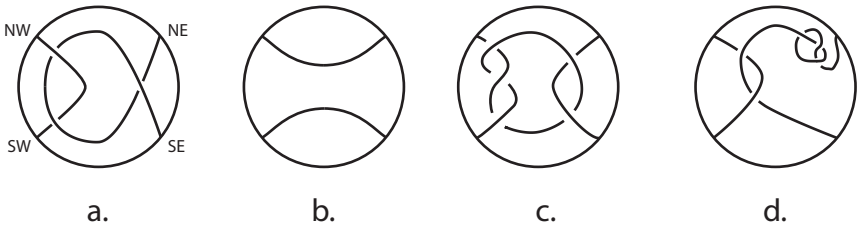

Figure 4: Examples of tangles: a. Rational, b. Trivial, c. Prime, d. Locally Knotted

Every rational tangle can be represented by a vector $(a_1, a_2, ..., a_n)$ where $a_i \in \mathbb{Z}$ for all i. This vector can be used to draw the tangle diagram in the following way: Start with a circle with points labeled NW, NE, SW, SE and and connecting the arcs (as in figure 4(b)). If n is even, start at the bottom (SW and SE) and do a_1 half-twists (using the convention of

right-hand twists for positive a_1 and left-hand twists for negative a_1). Next, do a_2 half-twists of the NE-SE side of the diagram. Then go back to the bottom, etc. If n is odd, start on the right and repeat the procedure as before. For example, the rational tangle $(2, 1, 2)$ is constructed in figure 5.

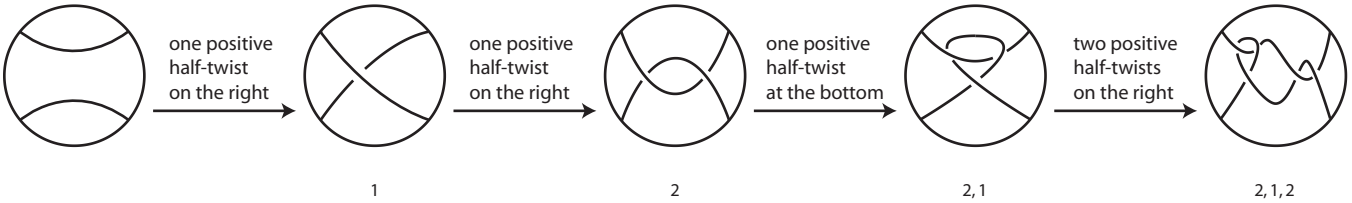

Figure 5: Draw the tangle (2, 1, 2)

Any vector with integer entries can be used to construct a continued fraction which is equal to a rational number $\frac{\beta}{\alpha}$. If Tangle T is represented by $(a_1, a_2, ..., a_n)$, one can construct the continued fraction $a_n + \dfrac{1}{a_{n-1} + \frac{1}{a_{n-2}+...+\frac{1}{a_1}}} = \dfrac{\beta}{\alpha}$. The rational number $\frac{\beta}{\alpha}$ is called the fraction of the tangle T.

Theorem 1 ([7]). *Two rational tangles are isotopic iff they have the same fraction.*

Two tangles are isotopic or equivalent if there is a mapping (called an ambient isotopy) which deforms one tangle to the other without moving the endpoints, breaking a string, or passing one string through another. This theorem says that two tangles have the same fraction representation if and only if they are equivalent. This theorem tells us that this fraction essentially describes the tangle.

Above we see that a fraction can be created from an integer-entry vector. Conversely, a fraction $\frac{\beta}{\alpha}$ of a tangle can be expanded into a continued fraction $\dfrac{\beta}{\alpha} = a_n + \dfrac{1}{a_{n-1} + \frac{1}{a_{n-2}+...+\frac{1}{a_1}}}$. From the continued fraction, you can create an integer-entry vector representation of the tangle $(a_1, a_2, ..., a_n)$. Since the continued fraction expansion of a rational number is not

unique, more than one vector may represent the same tangle. For example, the vectors $(3, -2, 2)$ and $(2, 2, 1)$ represent the same tangles. This can be seen by computing the rational number that corresponds to $(3, -2, 2)$, which is $2 + \dfrac{1}{-2 + \frac{1}{3}} = \dfrac{7}{5}$. Then by expanding $\frac{7}{5}$ in to a continued fraction $1 + \dfrac{1}{2 + \frac{1}{2}}$ the vector $(2, 2, 1)$ is obtained. By theorem 1 both vectors represent the same tangle. However, every rational tangle (with the exception of $\{(0), (\pm 1), (\infty)\}$) has a unique canonical vector representation called the Conway symbol. A vector $(a_1, a_2, ..., a_n)$ is said to be in canonical form if $|a_1| > 1$, $a_i \neq 0$ for $1 \leq i \leq n-1$, and all nonzero entries have the same sign. The Conway symbol of the example above is $(2, 2, 1)$. The Conway symbol of the four tangles excluded from the canonical form are $\{(0), (\pm 1), (\infty)\}$.

The next theorem, which is a direct result of Conway's theorem will gives us a means of classifying rational tangles by way of their fractions.

Theorem 2 (Rational Tangle Classification Theorem [5]). *There exists a 1-1 correspondence between classes of rational tangles and the extended rational numbers* $\frac{\beta}{\alpha} \in \mathbb{Q} \cup \{\frac{1}{0} = \infty\}$ *where* $\alpha \in \mathbb{N} \cup \{0\}, \beta \in \mathbb{Z}$ *and* $gcd(\alpha, \beta) = 1$.

In addition to the vector notation and tangle fractions, tangles can also be represented as a matrix. This matrix representation will be used in Lemma 5. Given any even-length vector representative for the tangle $\frac{\beta}{\alpha}$, we can compute a 2×2 matrix representative by the following equation:

$$\begin{bmatrix} u & v' \\ v & u' \end{bmatrix} = \begin{bmatrix} 1 & a_{2k} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ a_{2k-1} & 1 \end{bmatrix} ... \begin{bmatrix} 1 & 0 \\ a_1 & 1 \end{bmatrix}$$

Here is an example. Let $\frac{\beta}{\alpha} = \frac{23}{17} = 1 + \frac{1}{2 + \frac{1}{1 + \frac{1}{5}}} = (5, 1, 2, 1)$ the matrix representative is

$$\begin{bmatrix} u & v' \\ v & u' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 5 & 1 \end{bmatrix} = \begin{bmatrix} 23 & 4 \\ 17 & 3 \end{bmatrix}.$$

3 Tangle Operations

There are several operations that can be performed on tangles. Given tangles A and B, the sum $A + B$ is formed by connecting the NE and SE endpoints of one, to the NW and SW endpoints of the other, respectively.

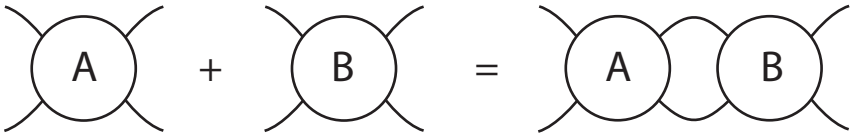

Figure 6: Tangle Addition

Given a tangle T, the numerator closure, $N(T)$, is formed by connecting the NW and NE endpoints and the SW and SE endpoints.

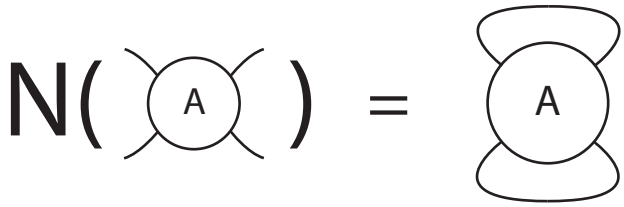

Figure 7: Numerator Closure

The denominator closure of T, $D(T)$, is formed by connecting the NW and SW endpoints and connecting the NE and SE endpoints. For example, the numerator closure of the tangle (2) is the hopf link $< 2 >$.

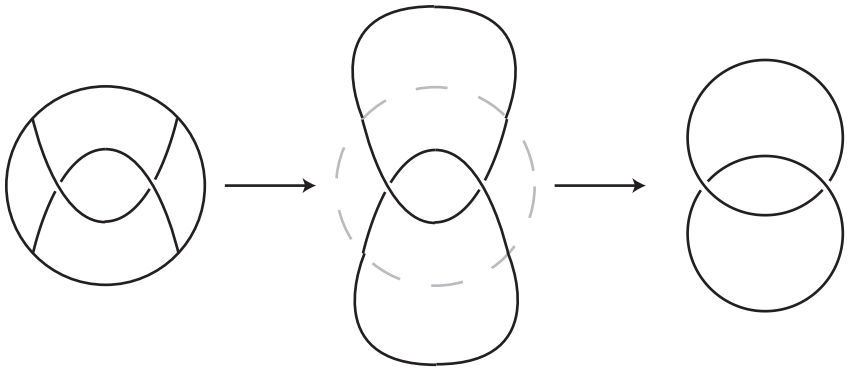

Figure 8: Example of Numerator Closure

The operations, tangle addition and numerator closure can be combined to form tangle equations. An example of a tangle equation is $N((2,0)+(1)) =< 3 >$ where $< 3 >$ is the trefoil knot (see Figure 9). This equation is of the form $N(A+B) = K$, where K is a knot. Later in this paper we will consider tangle equations of this general form where K is a knot or link.

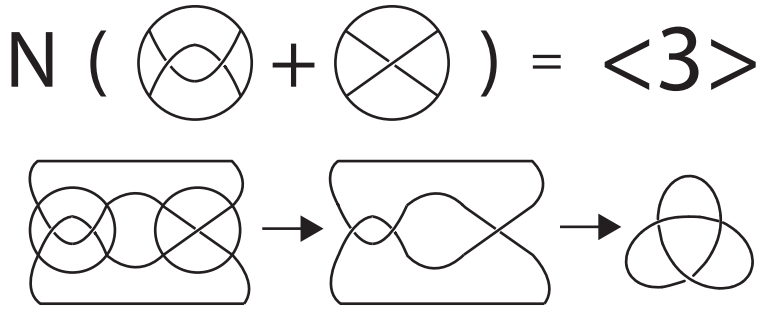

Figure 9: Tangle Equation

Each tangle has a *parity* of $(0), (1)$, or $(0,0)$. If the string which starts at the NW position of a tangle T ends at the NE position, we say T has parity (0) and we denote this by $T \approx (0)$. If the string which starts at the NW position ends at the SE position, then T $\approx (1)$. And if the string which starts at the NW position ends at the SW position, T has parity $(0,0)$ or (∞).

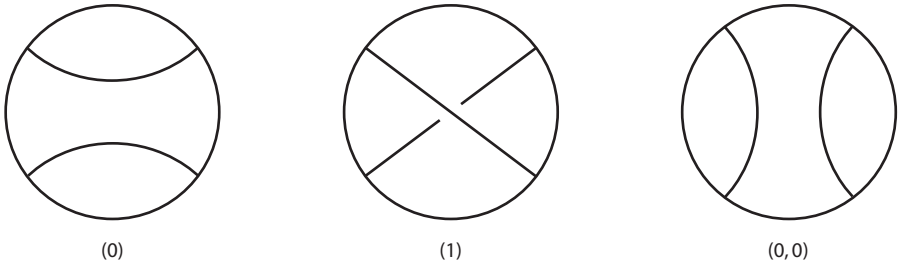

Figure 10: Parity

Note that if A and B are rational tangles, $A+B$ is not necessarily rational. For example, if A and B are both of parity $(0,0)$, then $A+B$ would not be a rational tangle because it would contain a circle in addition to the two arcs. For example look at the tangle $(2,0)$, which has parity $(0,0)$; the sum $(2,0)+(2,0)$ is not a rational tangle (Figure 11).

Figure 11: (2,0)+(2,0)

It may seem like all hope is lost, since tangle addition is not well behaved, but although the sum of two rational tangles is not necessarily rational, the numerator closure of the sum of two rational tangles is always a well understood kind of knot called a 4-plat. A 4-plat is a type of knot which will be discussed in the following section and will be important for modeling DNA molecules.

4 4-Plats

A 4-plat is a knot (link) made by braiding four strings and connecting the ends as shown below in Figure 13 [1]. 4-plats, also known as two-bridge or rational knots, admit a diagram

in which one of the strings is free from crossings. All prime knots with less than 8 crossings and all prime, two-component links with less than 7 crossings are 4-plats. A prime knot is a knot other than the unknot which cannot be expressed as a composition of two other knots, neither of which is unknotted [1].

4-plats can be represented by an integer-entry vector, much like rational tangles. The 4-plat vector representative is an odd-length vector $< c_1, ..., c_{2k+1} >$ where $c_i \geq 1$ for all i and where each integer represents a half-twist between strings. Also, like rational tangles, the vector can be used to draw the 4-plat diagram. Start with four strings (Figure 15a), do c_1 half-twists between the middle two strings, bringing the bottom string on top (Figure 15b). Next, do c_2 half-twists between the top and second string, this time bring the top string down. Go back to the middle two strings and repeat this process until you have completed the twists for all integers in the vector. Last, connect the ends as shown in Figure 15e.

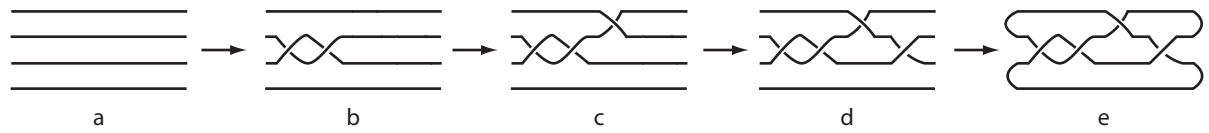

Figure 12: Drawing 4-Plats

This vector representation is called the *Conway symbol* for the 4-plat and corresponds to a minimal, alternating diagram of the 4-plat. Two 4-plats are the same if and only if they have the same Conway symbol or if one is the same as the reverse of the other; $< c_1, ..., c_{2k+1} >$ is the same four plat as $< c_{2k+1}, ...c_1 >$. With the exception of the unknot $< 1 >$ and the unlink of two components $< 0 >$, the Conway symbol can be used to compute a classifying rational number $\frac{\beta}{\alpha}$ with $0 < \beta < \alpha$ where $\frac{\beta}{\alpha} = \frac{1}{c_1 + \frac{1}{c_2 + ...}}$. The 4-plat $\frac{\beta}{\alpha}$ is denoted $b(\alpha, \beta)$.

Figure 13: 4-Plats

Theorem 3 (4-Plat Classification Theorem [2]). *Two 4-plats $b(\alpha, \beta)$ and $b(\alpha', \beta')$ are equivalent (as unoriented knots or links) iff $\alpha = \alpha'$ and $\beta^{\pm 1} \equiv \beta'(\mod \alpha)$.*

For example, look at the two 4-plats $b(17, 5)$ and $b(17, 7)$. The Conway symbol corresponding to $b(17, 5)$ is $(3, 2, 2)$, and the Conway symbol corresponding to $b(17, 7)$ is $(2, 2, 3)$. Therefore, the two are equivalent 4-plats. Notice $17 = 17$ and $5^{-1} \equiv 7(\mod 17)$.

Rational tangles and 4-plats are closely related by means of the rational number representation. If given a rational number $\frac{\beta}{\alpha}$ with $0 < \frac{\beta}{\alpha} < 1$, the denominator closure of the rational tangle $\frac{\beta}{\alpha}$ gives the 4-plat $b(\alpha, \beta)$ and if $\frac{\beta}{\alpha} \geq 1$ the numerator closure of the tangle $\frac{\beta}{\alpha}$ gives the 4-plat $b(\beta, -\alpha)$. For any integer x, $D((d_1, ..., d_{2k+1}, x)) =< d_1, ..., d_{2k+1} >$ and $N((d_1, ..., d_{2k+1}, x, 0)) =< -d_1, ..., -d_{2k+1} >$. Also, as mentioned before, the numerator closure of the sum of two rational tangles is a 4-plat. The following theorem discusses equivalence of rational knots obtained by taking the numerator closure of rational tangles. We know that the numerator closure of a rational tangle is a 4-plat, so this theorem is much like the 4-Plat Classification Theorem.

Theorem 4 ([7]). *Suppose the rational tangles with reduced fractions $\frac{p}{q}$ and $\frac{p'}{q'}$ are given. If $N(\frac{p}{q})$ and $N(\frac{p'}{q'})$ denote the corresponding rational knots obtained by taking numerator closures of those tangles, then $N(\frac{p}{q})$ and $N(\frac{p'}{q'})$ are topologically equivalent iff $p = p'$ and $q^{\pm 1} \equiv q'(\mod p)$.*

Obviously, if $p = p'$ and $q = q'$, then the $N(\frac{p}{q}) = N(\frac{p'}{q'})$. This theorem also says that if $p = p'$ and if $q^{-1} \equiv q'(\mod p)$, then the numerator closure of the two tangles will yield the same knot.

5 Solving Tangle Equations

A tangle equation, introduced in Section 3, is an equation of the form $N(A+B) = K$, where A and B are tangles and K is a knot or link. Solving equations of this type will be useful in the the tangle model described in Section 7, and therefore important in gaining a better understanding of certain enzyme mechanisms.

When working with tangle equations, there are several possible situations to consider. It may be that the two tangles A and B can be used to solve for the unknown knot K. But it could also be that one or both of the tangles are unknown and the knot K is known. The following theorems are used in solving equations of these types.

If the two tangles in the equation are known, Lemma 5 can be used to solve for the 4-plat that results from taking the numerator closure of the sum of the two tangles.

Lemma 5 ([5]). *Given two rational tangles $A_1 = \frac{\beta_1}{\alpha_1}$ and $A_2 = \frac{\beta_2}{\alpha_2}$, then $N(A_1 + A_2)$ is a 4-plat which is equal to $b(\alpha, \beta)$, where $\alpha = |\alpha_1\beta_2 + \alpha_2\beta_1|$ and β is determined as follows:*

1. *if $\alpha = 0$, then $\beta = 1$;*

2. *if $\alpha = 1$, then $\beta = 1$;*

3. *if $\alpha > 1$, then β is uniquely determined by the following: $0 < \beta < \alpha$ and $\beta \equiv \sigma(\alpha_1\alpha_2' + \beta_1\beta_2'(\mod \alpha)$, where $\sigma = sign(\alpha_1\beta_2 + \alpha_2\beta_1)$ and α_2' and β_2' are the entries*

in the second column of any matrix representative for the tangle $\frac{\beta_2}{\alpha_2}$.

Example 6. Let $A_1 = 2$ and $A_2 = \frac{23}{17}$. Then $\alpha = |1 * 23 + 17 * 2| = 57$ and getting α_2' and β_2' from the example of the matrix representative for $\frac{23}{17}$, $\beta = (1 * 4 + 2 * 3)(\mod 57) = 10$. Therefore, $N(A_1 + A_2) = b(57, 10)$.

It may be that one of the tangles in the equation is unknown and the other tangle and the knot K are known. Although we know that given tangles A and B and the tangle equation $N(A + B) = K$, if A and B are rational, then K is a 4-plat. On the other hand, if we have the tangle equation $N(X + A) = K$ where A is rational and K is a 4-plat, then there is not enough information to say that X is rational. In fact, there could be infinitely many prime tangle solutions. However, using advanced techniques one can sometimes prove in specific cases that X must be rational [5] [8] [10] [11]. Since the tangle model concerns only rational tangles, these are the only solutions the theory is equipped to handle. The following theorem gives all of the rational tangle solutions.

Theorem 7 ([5]). *Let* $A = \frac{\beta}{\alpha} = (a_1, a_2, ..., a_{2n})$ *be a rational tangle and* $K =< c_1, c_2, ..., c_{2k+1} >$ *be a 4-plat. The rational tangle solutions to the equation* $N(X + A) = K \neq< 0 >$ *are the following:* $X = (c_1, ..., c_{2k+1}, r, -a_1, ..., -a_{2n})$ *or* $X = (c_{2k+1}, ..., c_1, r, -a_1, ..., -a_{2n})$, *with* r *any integer. If* $K =< 0 >$, *then* $X = (-a_1, -a_2, ..., -a_{2n})$ *is the unique solution.*

The previous theorem shows us that an equation of the type $N(X + A) = K \neq< 0 >$ where X is the unknown, has infinitely many rational tangle solutions. On the other hand, if given two equations of this type with one unknown, the following corollary says that there are at most two distinct rational solutions.

Corollary 8 ([5]). *Let A_1, A_2 be distinct rational tangles, and K_1, K_2 be 4-plats. There are at most two distinct rational tangle solutions to the equations $N(X + A_1) = K_1$ and $N(X + A_2) = K_2$.*

Proof: Let $X = \frac{u}{v}$, $A_1 = \frac{\beta_1}{\alpha_1}$, $A_2 = \frac{\beta_2}{\alpha_2}$, $K_1 = b(\alpha, \beta)$, and $K_2 = b(\alpha', \beta')$. Then by Lemma 5, we have $\alpha = |v\beta_1 + \alpha_1 u|$ and $\alpha' = |v\beta_2 + \alpha_2 u|$. In the (u, v)-plane, these equations describe two pairs of parallel straight lines. These lines intersect in at most 4 points. Since $\frac{u}{v} = \frac{-u}{-v}$, these four points of intersection describe at most two distinct rational tangle solutions for the equations in the hypothesis $\square$

The following is an example of this corollary.

Example 9 ([5]). Let $A_1 = \frac{1}{3}$, $A_2 = \frac{5}{17}$, $K_1 = b(5, 3)$, and $K_2 = b(29, 17)$. Then,

$$|v + 3u| = 5$$

$$|5v + 17u| = 29$$

Solving this system equations:

$$v + 3u = 5$$

$$5v + 17u = 29$$

we get $X = -\frac{2}{1}$. Next, a second solution can be obtained by solving:

$$v + 3u = 5$$

$$5v + 17u = -29$$

so we get $X = -\frac{27}{86}$. The previous example shows a case where there are two distinct rational solutions to the equations.

In order to describe the enzyme mechanism in the topological approach to enzymology, we must be able to give a unique solution for R to a system of tangle equations as described above. In [6] the following result does exactly that.

Theorem 10 ([6]). *A system of four simultaneous tangle equations $N(O + iR) = K_i$ for $0 \leq i \leq 3$ where K_i are 4-plats and $\{K_1, K_2, K_3\}$ represent at least 2 different link or knot types has at most one simultaneous solution $\{O, R\}$ for some integral tangle R and a tangle O which is either a rational tangle or the sum of two rational tangles. Moreover, if there exists a solution, then at least one of the 4-plats K_i must be chiral (or not equivalent to its mirror image).*

Note that an integral tangle is a tangle of the form (k) where k is an integer.

6 Site-Specific Recombination

Deoxyribonucleic acid (DNA) are long, thin molecules that are tightly packed into the cell nucleus. *Duplex* (double-stranded) DNA consists of two backbone strands. Each strand consists of a sugar phosphate backbone with a nitrogen base attached to each sugar. The four possible bases are: A-adenine, G-guanine, C-cytosine, and T-thymine. The strands form hydrogen bonds between each other, where A only bonds with T, and C only bonds with G. This is what forms the double helix shape of DNA. Therefore, by reading the letters of one strand, you know that the other strand is a dual copy with A replaced with T and C replaced with G and vice versa. The sequence of letters obtained by reading down one strand is called the DNA's *genetic sequence*. DNA is twisted in a right-hand helical fashion, with an average pitch of approximately 10.5 base pairs for each full twist. Each half twist is called a *supercoil*. As discussed in the introduction, DNA must be topologically manipulated by enzymes in order for vital life processes to occur. One of these enzymatic actions is called Site-Specific Recombination.

Site-Specific Recombination is a process by which a block of DNA is moved to another position on the molecule or a block of viral DNA is integrated into a host genome. Recombination is used for gene rearrangement, gene regulation, copy number control, and gene therapy. This process is mediated by an enzyme called a *recombinase*. A small segment of the genetic sequence of the DNA that is recognized by the recombinase is called a *recombination site*. A pair of sites on the same molecule or different molecules, once recognized, are aligned and then bound by the enzyme. This is the stage of the reaction called *synapsis*. The DNA molecule(s) and the enzyme itself are called the *synaptic complex*. Before recombination the DNA molecule is called the *substrate* and after recombination it is called the *product*. Once bound to the DNA, the enzyme breaks the DNA at the two sites and then recombines the ends by exchanging them. Each of the recombination sites is oriented by the order in which the bases appear as one reads around the DNA strand in some predetermined order. If the orientations of the sites agree, the site configuration is called *direct repeats*. If they disagree, this is called *inverted repeats*

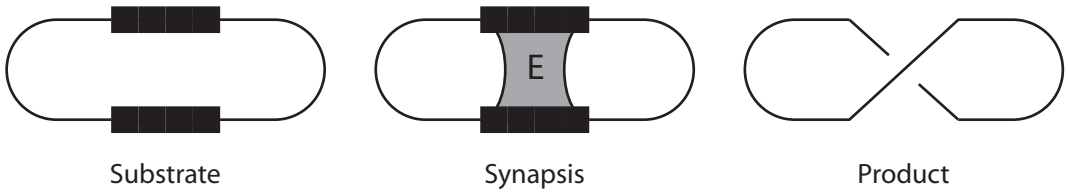

Figure 14: The dark bands at left denote the recombination site. The middle figure depicts the recombinase bound to the sites. At right is the product after the enzyme has reacted and detached from the site

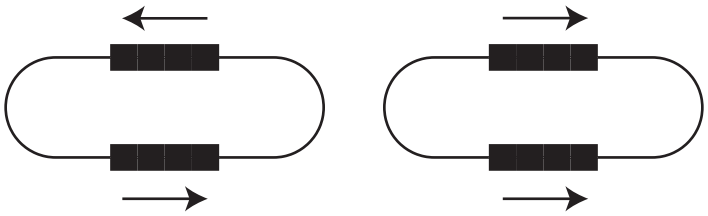

Figure 15: Direct repeats (left) and inverted repeats (right)

The enzyme may act by performing more than one recombination event while bound to the DNA. This is called *processive recombination*. If more than one recombination event occurs during separate binding encounters, it is called *distributive recombination*. An enzyme may act both processively and distributively in regards to recombination. Determining the topology of the protein-DNA complex in solution is difficult to do directly, so indirect methods are used. One such method is called the *topological approach to enzymology*. Using cloning techniques, circular substrate molecules can be genetically engineered. Experiments can then be performed on these molecules. Using circular substrates has the advantage of making the topological changes easier to detect. Once the reaction has taken place, the changes are observed experimentally. *Gel electrophoresis* is used to separate the reaction product into different knot and link types. First, the DNA products are put at the top of an agarose gel. Then a positive charge is put at the bottom of the gel which attracts the negatively charged DNA. The smaller (or more knotted up) the DNA product, the faster it will travel. After gel electrophoresis the DNA knots and links are directly observed using electron microscopy.

7 The Tangle Model

The purpose of the tangle model, which was introduced by DeWitt Sumners in 1980, is to deduce mathematically what happens during recombination. That is, given the geometry and topology of the substrate and product DNA, can we figure out exactly what the enzyme is doing? In the electron micrographs, DNA strands can be observed winding about each other. Since rational tangles and 4-plats are formed by twisting strings, we find them to be the perfect candidate for modeling DNA. Recall that the definition of a tangle is a pair (B, t), where B is a 3-ball and t is a pair of unoriented arcs (strings) properly embedded in B. A tangle can be used to model the enzyme-DNA complex with the enzyme being the 3-ball and the two strings being the two recombination sites. Most observed products of recombination experiments are 4-plats. Since the numerator closure of a sum of rational tangles is a 4-plat, it is conceivable that we could model the enzyme-DNA complex and the changes taking place with tangle equations. But, before we can use the tangle model to deduce the enzyme mechanism, there are several assumptions that must be made. The first assumption is that the enzyme-DNA complex can be represented as a sum of tangles. E is the enzyme, O_b is the part of the DNA which is bound to the enzyme, but unchanged during the reaction, and P is the site which is changed during the reaction. Therefore, we can write the enzyme-DNA complex as $E = O_b + P$. We also need to consider the free DNA that is not bound to the enzyme. We'll call the tangle formed by this part of the DNA O_f. We now have one tangle equation, $N(O_f + O_b + P) = K_0$, the substrate molecule. The second assumption is that recombination acts by tangle surgery where the site tangle P is replaced by the recombinant tangle R after one round of recombination. By this assumption, we can model one round of

recombination by replacing the P in our equation for the substrate molecule with R. Here is this model for one recombination round:

$$\begin{array}{ll} N(O_f + O_b + P) = K_0 & \text{(substrate)} \\ N(O_f + O_b + R) = K_1 & \text{(product)} \end{array}$$

Next, we must assume that the mechanism of recombination is constant, independent of substrate geometry and topology. This means that if all of the substrate molecules are all of the same knot type, then the tangles O_f, O_b, P and R won't change from one event to another. If the substrate molecules are of different knot types, the only tangle that would change is O_f. The only exception to this is that we do need to consider site orientation. The last assumption will be that processive recombination acts like tangle addition. This means after n rounds of recombination, P becomes $nR = R + R + ... + R$. Under these assumptions, the model for processive recombination is given by the system of tangle equations is:

$$\begin{array}{cc} N(O + P) = K_0 & \text{(substrate)} \\ N(O + R) = K_1 & \text{(product of the first round)} \\ \vdots & \vdots \\ N(O + nR) = K_n & \text{(product of nth round)} \end{array}$$

where $O = O_f + O_b$ and O, P, and R are unknown.

8 Example

When using the tangle model to analyze a specific enzyme, one first must prove rationality of the tangles in question which requires deep results in topology such as the Cyclic Surgery theorem [4]. Once rationality is shown, the experimental results are used to set up the system of tangle equations which can be solved for the enzyme action.

In 2002, Mariel Vazquez and De Witt Sumners used the tangle model to analyze Gin site-specific recombination [11]. This section will discuss their findings. This is just one example where the tangle model has been used to analyze a specific enzyme mechanism. Gin is a site-specific recombinase which is encoded by bacteriophage Mu. A bacteriophage is a virus that infects bacteria. The phage genome has two recombination sites, called gix L and gix R, which the Gin recognizes. Once bound to the DNA, the Gin makes a break at each site, rotates the ends, and then reconnects the ends. Gin acts by processive recombination which can result in more than one recombination event during a single binding. The results of tangle analysis of Gin recombination on unknotted substrate molecules with inversely repeated gix sites is as follows:

$$\begin{array}{cc} K_0 =< 1 > & \text{(the unknot)} \\ K_1 =< 1 > & \text{(the unknot)} \\ K_2 =< 3 >= 3_1 & \text{(the trefoil knot)} \\ K_3 =< 2,1,1 >= 4_1 & \text{(the figure-8 knot)} \\ K_4 =< 2,2,1 > & \text{(the 5-twist knot)} \end{array}$$

In 2004, De Witt Sumners and Mariel Vazquez gave the following result which solves the simultaneous system of the first four equations above and accurately predicts the fifth equation [11].

Theorem 11 (Inversely Repeated Sites Theorem). *The simultaneous solution* (O, R) *for the system of equations*

$$\begin{array}{cc} (i) & N(O+P) =< 1 >= \textit{the unknot} \\ (ii) & N(O+R) =< 1 >= \textit{the unknot} \\ (iii) & N(O+R+R) =< 3 >= \textit{the trefoil knot} \end{array}$$

for tangles O, P, and R is either $((-2,0),(1))$ *or* $((4,1),(-1))$*. Furthermore, if*

$$(iv) \quad N(O+R+R+R) =< 2,1,1 >= \textit{the figure-8 knot}$$

then there is a unique solution, namely $(O,R) = ((-2,0),(1))$*.*

The complete proof of this theorem can be found in [11]. Here we do not prove the rationality of O and R, rather we look at how the tangle equations were solved. First, recall that Lemma 5 said that given two rational tangles $A_1 = \frac{\beta_1}{\alpha_1}$ and $A_2 = \frac{\beta_2}{\alpha_2}$, then $N(A_1 + A_2)$ is a 4-plat which is equal to $b(\alpha, \beta)$ where $\alpha = |\alpha_1\beta_2 + \alpha_2\beta_1|$. Equations (ii) and (iii) from above give the following system of equations:

$$|u + rv| = 1$$
$$|u + 2rv| = 3$$

with u, r, v as unknown integrals. You can obtain ten solutions for the ordered pair $(\frac{u}{v}, r)$. Thus ten solutions for the rational tangle pair (O,R). The solutions are the pairs $((-2,0),(1)),((1),(-2)),((5),(-4)),((-2,-2,),(2)),((4,1),(-1))$ along with their mirror images. With aid of the following result, one can throw out several of these solutions.

Theorem 12 (Inversely Repeated Sites Claim ([11])). *The tangles involved in equation (i), (ii) and (iii) of the following theorem arising from Gin recombination on inversely repeated sites satisfy the following properties:* $O \approx (0,0)$*,* $R \approx (1)$ *and* $P \approx (0)$*.*

Since $O \approx (0,0)$ we can eliminate all solutions for which O is an integral tangle since it is easy to see that an integral tangle has parity (0) or (1). Additionally, if $R \approx (1)$, then we can get rid of any solution where $R = (2)$ since even integral tangles have parity (0) . We

can also discard the mirror images because the product knot of equation (iii) is chiral (i.e. not equivalent to its mirror image). Therefore we are left with only two solutions, of which only one satisfies the tangle equation (iv).

Through this tangle analysis Sumners and Vazquez have shown that when Gin reacts on a substrate with gix sites in the inverted orientation each round of recombination the enzyme mechanism adds one positive crossing to the substrate.

Instead of Gin acting on substrates with inverted directly repeated gix sites, it can also act on substrates with directly repeated sites. That is, the orientation of the sites is the same instead of opposite. For Gin, we get the following theorem from [11].

Theorem 13 (Directly Repeated Sites Theorem). *The simultaneous solution* (O, R) *for the system of equations*

$$(i) \qquad N(O+P) =< 1 >= \textit{the unknot}$$

$$(ii) \qquad N(O+R) =< 3 >= \textit{the trefoil knot}$$

$$(iii) \quad N(O+R+R) =< 1,2,2 >= \textit{the (-5) twist knot}$$

for tangles O, P, and R is either $((-2,0),(2))$ *or* $((2,1,1,2),(-2))$. *In addition, if*

$$(iv) \quad N(O+R+R+R) =< 1,4,2 >= \textit{(-7) twist knot}$$

then $(O,R) = ((-2,0),(2))$ *and*

$$(v) \quad N(O+nR) = \textit{the -(2n + 1) twist knot}$$

for all $n \geq 4$.

In this example, the tangle model has been used to mathematically show the enzyme mechanism of Gin. The result shows that with each round of recombination on inversely

repeated sites adds (1) to the tangle. In other words, $R = (+1)$. For directly repeated sites, $R = (+2)$.

References

[1] C. Adams, *The Knot Book*, New York: W.H. Freeman and Company, 1994.

[2] G. Burde and H. Zieschang, *Knots*, De

[3] J. Conway, On enumeration of knots and links and some of their related properties, *Computational Problems in Abstract Algebra; Proc. Conf. Oxford* (1970), 329-358.

[4] M. Culler, C. McA. Gordon, J. Luecke, and P.B. Shalen, Dehn surgery on knots, *Ann. of Math.* (2) **125** (1987), 237-300.

[5] C. Ernst and D.W. Sumners, A Calculus for Rational Tangles: applications to DNA recombinations, *Math. Proc. Camb. Phil. Soc.* 108 (1990), 489-515.

[6] C. Ernst and D.W. Sumners, Solving Tangle Equations Arising in a DNA Recombination Model, *Math. Proc. Camb. Phil. Soc.* 126 (1999), 23-36.

[7] L. H. Kauffman and S. Lambropoulou, From Tangle Fractions to DNA, to appear in *Contemp. Math.*.

[8] D. W. Sumners, Knot Theory and DNA, *Proceedings of Symposia in Applied Mathematics* 45 (1992), 39-72.

[9] D. W. Sumners, Lifting the Curtain: Using Topology to Probe the Hidden Action of Enzymes, *Notices of the AMS* 42 (1995), 528-537.

[10] D. W. Sumners, Untangling DNA, *The Mathematical Intelligencer* 12 (1990), 71-80.

[11] M. Vazquez and D. W. Sumners, Tangle Analysis of Gin Site Specific Recombination, *Math. Proc. Camb. Phil. Soc.* 136 (2004), 565-582.