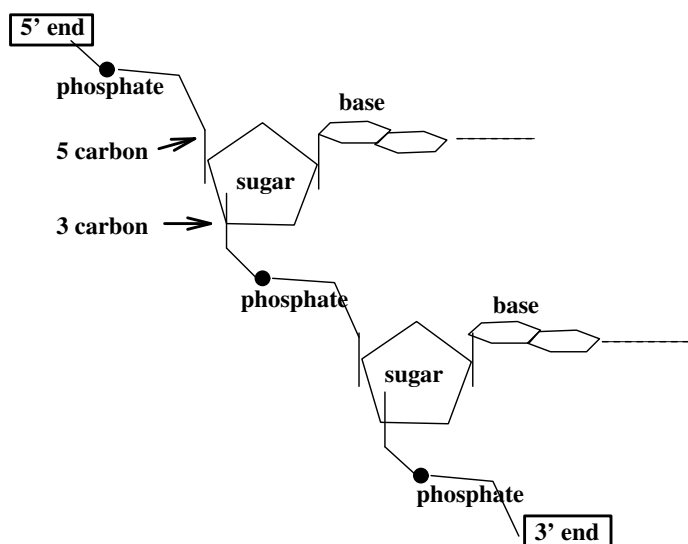


5 DNA

5.1 Background from biochemistry

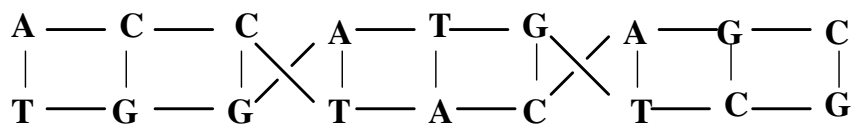
One of the most interesting applications of knot theory is to biochemistry, in particular to the study of the chemical which forms the primary genetic material of most organisms, DNA or deoxyribonucleic acid. In higher organisms, such as ourselves, DNA is found in the nucleus of the cell, but it is also present in organisms such as bacteria and viruses. DNA is not a single chemical substance but a very large family of substances, all built according to a common chemical structure, which we shall briefly describe. The exact composition of the DNA of an organism is special to that organism.

In its best known form, the DNA molecule consists of two linear strands which are wound round one another in the form of a double helix with a right hand thread. The backbone of each strand is made up of alternating sugar and phosphate units linked by the strong chemical bonds called covalent bonds, in which electrons are shared between neighbouring atoms.



Attached to each sugar and extending into the interior of the helix is one of four bases: adenine (A), thymine (T), cytosine (C) or guanine (G). Each base on one strand is attached by a weak chemical bond called a hydrogen bond to a corresponding base on the other strand. Base A can form bonds only with base T, and similarly base C with base G. Thus the sequence of bases on one strand fixes the sequence of bases on the other strand uniquely. This sequence of paired bases, which may range in length from several thousand to many millions of pairs

depending on the particular organism from which the DNA is taken, constitutes the so-called *genetic code*.



DNA is responsible for a number of fundamental biological processes. These include

- *replication*, or making a copy of the whole molecule,
- *transcription*, or copying a segment of the molecule, and
- *recombination*, or modifying the given molecule in one of a number of ways.

The exact mechanisms by which these changes in DNA take place are controlled by members of another chemical family called *enzymes*. The basic function of an enzyme is to act as a catalyst in a biochemical reaction. This means that the direction in which a reaction would naturally tend to occur, because of the particular environment in which the DNA finds itself, is accelerated by the presence of the enzyme. The enzyme itself is not altered as the reaction proceeds. A typical enzyme might consist of a chain of about 1000 amino acids, the basic building blocks of proteins.

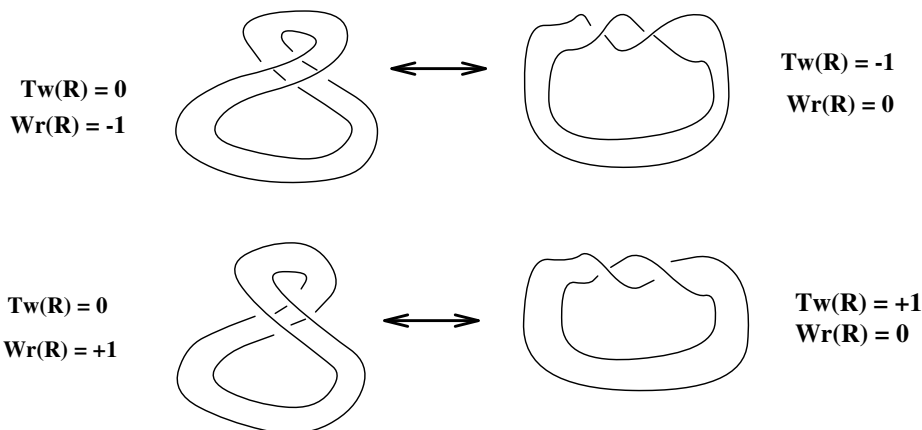
Although the basic information content of the DNA is completely determined by the sequence of bases, and does not depend on the spatial configuration of the molecule, its biological functions are affected by it. Two molecules that have the same chemical composition, but differ in structure, are called *isomers*. Isomers can differ in a geometrical property, such as left and right handed sugar molecules. In the case of DNA, it is possible for isomers to differ in a topological property, such as knotting or linking. Such isomers are called *topoisomers*.

The physical arrangement of the DNA molecule can be described in terms of geometry and topology. For this purpose, we can think of the molecule as a long rope made up of two strands which are wound round and round each other, like a piece of two-strand electric flex. Geometrical properties generally have to do with the rate of winding of one strand of the double helix around the other. Biologists talk of a molecule as *relaxed* if the axis of the helix lies flat, and as *supercoiled* if the axis itself coils through space. You can see this effect by adding extra twists to a length of electric flex. Supercoiling can be of two kinds, positive and negative. Positive supercoiling occurs when more twists are added to the DNA helix, and negative supercoiling occurs when there are fewer twists in the DNA than there are in its relaxed state. When the DNA double helix is in solution, as it is found in its natural state in the cell, one strand winds round the other at a rate of one complete revolution every 10.5 base pairs.

In order to study individual DNA rings in this way, it is first necessary to separate them. This is done by a process called *gel electrophoresis*, in which molecules of DNA are placed in a gel and an electric current is passed through to attract them towards an electrode. The more tightly supercoiled molecules are able to move faster through the gel, and can be separated out.

We next explain the three quantities involved in the equation $\text{Lk}(R) = \text{Tw}(R) + \text{Wr}(R)$, as they apply to closed circular duplex (double-stranded) DNA. In this form, The molecule of DNA may be modelled mathematically as a twisted ribbon in \mathbf{R}^3 . The two strands containing the bases can be regarded as the two curves which form the edges of the ribbon. Then the linking number $\text{Lk}(R)$ can be defined as the linking number of these two edge curves. It can equally well be taken as the linking number of one of the edge curves with the axis of the ribbon (why?). The linking number cannot be changed by any continuous deformation of the pair of curves, so long as no break is made in the curves. Also, we have seen that it can be calculated from any diagram of the ribbon, provided that we avoid projecting in certain exceptional directions.

Although $\text{Lk}(R)$ is a topological invariant, the other two terms in the equation, the twist $\text{Tw}(R)$ and the writhe $\text{Wr}(R)$, belong to differential geometry. These numbers are not necessarily integers, and each can be changed by continuous deformation of the ribbon R . To understand these numbers, let's go back to some of the diagrams of Section 2.5.



Broadly speaking, the twist $\text{Tw}(R)$ measures how much the ribbon R twists about its axis A , while the writhe $\text{Wr}(R)$ measures how much the axis A itself writhes around in space. When the axis of the ribbon lies flat in a plane, as in the right hand diagrams above, $\text{Wr}(R) = 0$. In the left hand diagrams, the ribbon doesn't twist about its axis at all, giving $\text{Tw}(R) = 0$. In these cases, the value of the other quantity is of course the same as the linking number $\text{Lk}(R)$.

In Chapter 6, we shall define the writhe $\text{wr}(D)$ of a knot or link *diagram* D to be the sum of the signs of all the crossings in D . This is not a topological invariant, because it is obviously changed by Reidemeister I moves. (In fact,

we'll see later that it *is* invariant under RII and RIII moves.) From a geometric viewpoint, the writhe of the diagram given by projecting the axis curve A of a DNA ribbon on to a plane will be different for different choices of projection plane.

So what do we mean by $\text{Wr}(R)$? To define this quantity in general, we *average* the writhe of the diagrams obtained by projecting R in all possible directions. Thus $\text{Wr}(R)$ is defined to be the surface integral

$$\text{Wr}(R) = \frac{1}{4\pi} \int \text{wr}(D) ds,$$

taken over the unit sphere in \mathbf{R}^3 , where $\text{wr}(D)$ is the writhe of the diagram D representing the axis A in a projection, and ds measures area on the unit sphere, so that $\int ds = 4\pi$. Of course, it isn't usually easy to evaluate $\text{Wr}(R)$ directly from this definition, but at least we can see that this definition gives 0 when the axis A lies in a plane.

The twist $\text{Tw}(R)$ of the ribbon R is also defined by an integral, so that it is additive on segments of R . We can measure the twist by placing a small arrow on the ribbon perpendicular to its axis and pointing to one of its edges. As the arrow moves along the ribbon, it rotates around the axis, and each complete rotation contributes $+1$ or -1 to $\text{Tw}(R)$, depending on the direction of rotation. More formally, we denote this arrow by a unit vector \mathbf{v} and form the line integral

$$\text{Tw}(R) = \frac{1}{2\pi} \int (\mathbf{T} \wedge \mathbf{v}) \cdot d\mathbf{v}$$

taken along the axis curve A , where \mathbf{T} is the unit tangent vector to A . For example, when the axis A is the z -axis in \mathbf{R}^3 and the edge of R is the right handed helix

$$x = r \cos t, y = r \sin t, z = t,$$

then $\mathbf{v} = (\cos t, \sin t, 0)$, $\mathbf{T} = (0, 0, 1)$, and $\text{Tw}(R)$ increases by 1 when t increases by 2π .

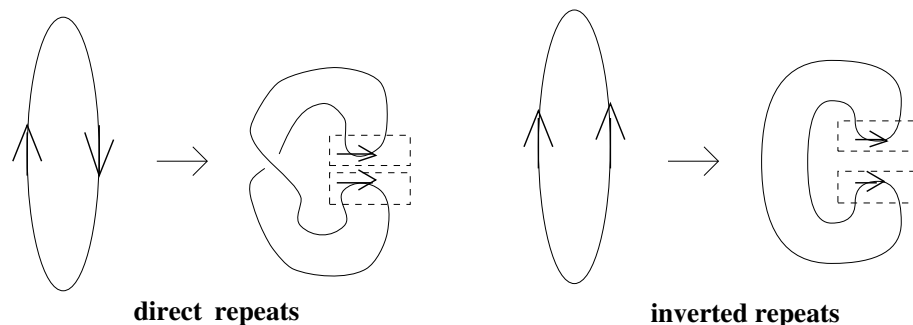
We have mentioned that in its relaxed state duplex DNA is modelled by a right handed double helix with one complete rotation about the axis for every 10.5 base pairs. Thus a closed circular loop of duplex DNA in the relaxed state with 5250 base pairs would have $\text{Tw} = 500$. (These values are approximately correct for the simian virus SV40.) If the axis lies in a plane, then this would also be the value of $\text{Lk}(R)$, since then $\text{Wr}(R) = 0$. Usually, however, in its native free state, SV40 DNA has a smaller linking number than this, typically about 475. This is because of negative supercoiling, meaning that the axis has a negative writhe. The observed value of the writhe, measured by counting crossings in projections and averaging as explained above, is typically about -18. This gives a twisting number of $475 + 18 = 493$ rather than 500. It seems that the deficiency in linking number from the relaxed state is compensated partly by a change in Tw

and partly by a change in W_r . Biologists believe that a deficit in linking number provides the DNA molecule with stored energy that is used when the two strands are separated. Such separation is needed in many biological processes.

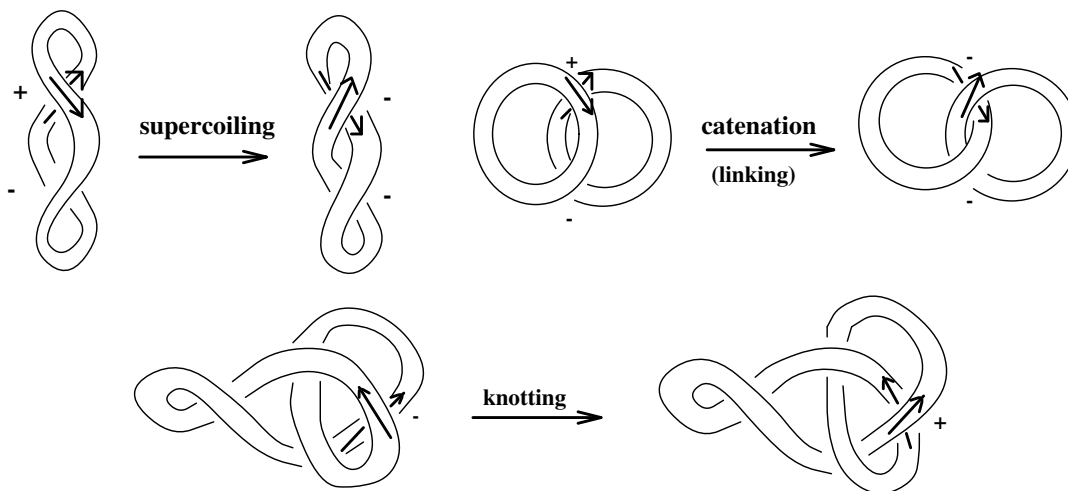
5.3 Studying enzyme actions using knot theory

Certain enzymes, called *topoisomerases*, act on the DNA molecule in such a way as to alter the topological arrangement of the molecule in space. They do this by cutting and rejoining the strands in various ways, the specific action being characteristic of the particular enzyme used. Knot theory has been applied in an attempt to work out the action of the enzymes by analysing the products of these reactions and observing the type of knots produced.

We discuss a particular type of action by an enzyme called *site-specific recombination*. In this process, an enzyme attaches itself to two specific sites on two strands of DNA, called the *combination sites*. The two sites have the same sequence of base pairs, which may be on the same or on different strands of DNA. When they are on the same circular DNA molecule, the combination sites may lie in the same direction round the molecule or in opposite directions. The first case is called *direct repeats* and the second is called *inverted repeats*.



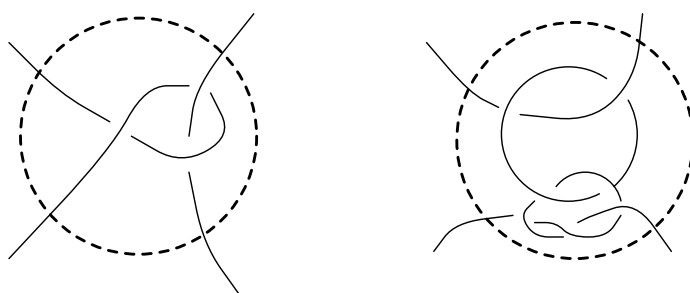
The enzyme lines up the two combination sites, then cuts the two strands open and recombines the ends in some manner. Some possible actions of an enzyme on cyclic duplex DNA are represented schematically below. Here the DNA is modelled mathematically as a ribbon whose two edges correspond to the two base sequences. The linking number between these two strands can be changed by the reactions.



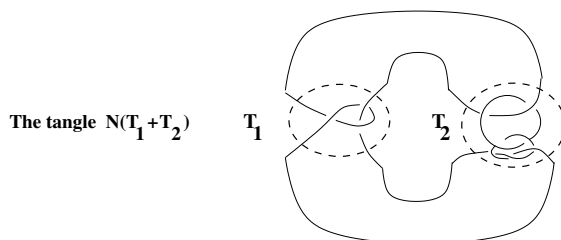
The evidence suggests that this is done in a way which depends only on the particular enzyme. However, the exact recombination process that has occurred cannot be observed directly. Thus we should like to deduce it by studying the topology of the products of the recombination and their relation to the topology of the *substrate*, the original DNA on which the enzyme action is taking place.

The action of a given enzyme can be analysed with the help of *tangle* theory. This is an idea introduced by John Conway to study knots, and in particular to provide a more systematic notation for knots than the traditional enumeration found in tables.

A *tangle* in a knot or link diagram is a circular region in the projection plane with the property that the link crosses the bounding circle at exactly four points. We always regard these four points as occurring in the NW, NE, SE and SW directions from the centre of the circle.



Tangles can be used as building blocks for link diagrams. You can read about this in a number of books, such as Chapter 2 of *The Knot Book* by Colin Adams. The only thing we need for the DNA application is one way to combine two tangles to get a link. This is done as shown below, and it is denoted by $N(T_1 + T_2)$. The idea is similar to the closure of a braid.



Now we think of the substrate, *i.e.* the circular DNA molecule on which the enzyme is to act, as being made up of two tangles, the *substrate tangle* S , which is unchanged by the enzyme, and the *site tangle* T where the enzyme acts. The enzyme replaces the site tangle with a new tangle, the *recombination tangle* R . Thus the substrate is $N(S + T)$ and the product of the reaction is $N(S + R)$.

We assume that we can determine the knot or link types of the substrate and of the product by experiment. This gives us two equations, one for $N(S + T)$

and one for $N(S + R)$. However, we cannot expect to be able to determine the three unknown tangles R , S and T from this information alone.

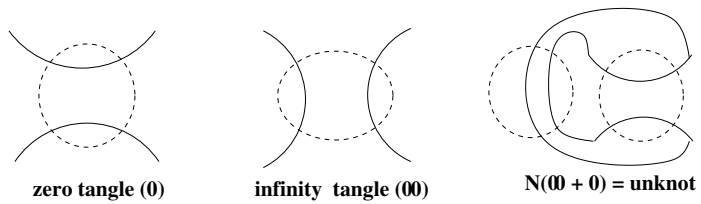
However, experimental observation suggests that the tangles T and R depend only on the *enzyme* and not on the topology of the molecule on which it acts. In certain circumstances, this additional information has been used to press the mathematical analysis a bit farther.

One example of a topoisomerase is *TN3 resolvase*, which acts on two matched sites in a cyclic duplex DNA molecule. Usually, the enzyme replaces the site tangle T with a single copy of the R tangle. However, sometimes it repeats the operation two, three or more times before releasing the molecule. In a series of experiments, the various products of these reactions were analysed, with results as follows.

$$\begin{aligned} N(S + T) &= 0 \text{ (the unknot)} \\ N(S + R) &= H^+ \text{ (the positive Hopf link)} \\ N(S + R + R) &= 4_1 \text{ (the figure eight knot)} \\ N(S + R + R + R) &= L \text{ (the Whitehead link, page 45)} \end{aligned}$$

From this set of equations, Ernst and Sumners deduced that the tangle S is the two string braid σ_1^{-3} whose closure is the positive trefoil knot, while the tangle R is the simple negative crossing σ_1 . Further, they then showed that $N(S + R + R + R + R)$ must be a 6_2 knot. This knot has in fact been observed as a product.

Since the substrate tangle T appears in only one equation, it is impossible to determine it uniquely by a rigorous mathematical argument. However, the results are consistent with the assumption that T is the tangle in our K^0 diagrams (the ‘zero tangle’).



More information on “topological enzymology” and other applications of knot theory to chemistry and physics are discussed in Chapter 7 of *the Knot Book*. Knot theory also has applications to other branches of mathematics, and one such application, to graph theory, is treated by Adams in Chapter 8.