# Mathematical Modelling of DNA
# Course notes

November 19, 2001

## Contents

# 1 Introduction

## 1.1 Objectives of Course

Naturally, the course is certainly supposed to be of interest to people whose primary motivation is to learn about the mathematical modelling and computation of DNA. The focus of the course will be on the structural properties of DNA at the length scale of a few hundred, to several hundred base pairs. Much of our study will be centered on the motif of DNA mini-circles. Less obviously, the course is designed also to be of interest to people who are interested in mathematical modelling generally. That is the application of DNA is taken as a case study of the process of gaining understanding through:

- taking experimental data,
- developing a mathematical model,
- analyzing the model,
- computing with the model, and
- comparing computational predictions with further experimental data

To this second end, the course will introduce several mathematical and computational techniques that are known to be of importance in many different contexts.

### 1.1.1 What is the DNA Mini-Circle Problem?

A mini-circle is a relatively short, and therefore stiff, piece of DNA in which the double helix has closed on its own tail to form a twisted circle.

The properties of such mini-circles can be measured experimentally, and the mini-circles can even be observed directly using cryo-Electron-Microscopy. They form a very convenient motif for studying structural properties of DNA experimentally.

They also have lots (!) of very interesting mathematical features. Mini-circles will be the concrete problem that we will use as a specific context in which to introduce:

1. general approaches to modelling the structural, or tertiary, properties of DNA and other macro-molecules.
2. Several, rather general, mathematical and computational techniques.

Most of the first semester will focus on ways of describing the minimum energy, equilibrium shapes of mini-circles.

Toward the end of the first semester, and continuing throughout the second semester we will consider models of the dynamics of DNA in a solvent, which will lead us to statistical mechanics theories for polymers, Monte Carlo simulations, and various stochastic partial differential equation models of the Brownian and Langevin dynamics of DNA.

### 1.1.2 What are the Mathematical Ideas that Will Arise?

In equilibrium problems:

- non-dimensionalization and scaling
- the one dimensional calculus of variations
- the Hamiltonian formulation of self-adjoint two-point boundary value problems for ordinary differential equations
- bifurcation theory and the roles of symmetries and integrals
- the theory of the second variation and stability of equilibria
- the role of isoperimetric constraints
- the geometry and topology of Link, Twist and Writhe
- quaternion parametrization of SO(3) (i.e. proper rotation matrices)

In Statistical Mechanics Problems:

- Maxwell-Boltzmann probability distributions in phase and configuration space
- expectation values in polymer chain models

### 1.1.3 What are the Computational Ideas That will Arise?

From equilibrium models:

- numerical methods for two-point boundary value problems
- collocation as a space discretization
- numerical parameter continuation

- numerical symmetry breaking
- averaging and fitting continuum constitutive relations to discrete data

From polymer chain statistical mechanics models:

- numerical implementations of Monte Carlo methods

### 1.1.4 What are the Structural Biology Ideas That Will Arise?

- All-atom models
- Wedge-angle models
- Bead models
- Polymer chain models (freely jointed, freely rotating, twisted worm-like)

### 1.1.5 What are the Structural Biology Ideas That Will Not Arise?

There are many, very important and interesting questions in mathematics, computation and statistics related to sequencing the Human Genome, i.e. determining the list of base pairs that makes up the DNA in humans (or other organisms). There will probably not be time to mention such issues. We will be concentrating on models related to determination of the three dimensional or tertiary structure of DNA.

## 1.2 Generalities on DNA structure

### 1.2.1 The Watson-Crick Model of DNA (1953)

DNA composed of four bases (or nucleotides) A, T, C, and G which pair according to the pairing rules : A-T and C-G

DNA is composed of two complementary strands (or sequences) e.g.

$$\begin{array}{lccccccccccccccccc}
\text{strand 1:} & C & = & A & = & T & = & G & = & T & = & C & = & T & = & A & = & G \\
 & | & & | & & | & & | & & | & & | & & | & & | & & | \\
\text{strand 2:} & G & = & T & = & A & = & C & = & A & = & G & = & A & = & T & = & C
\end{array}$$

Sequence on first strand arbitrary, and then sequence on the other strand forced by pairing rules.

Along each strand the bases are connected through a covalently bonded (i.e. very strong) sugar-phosphate backbone. Moreover the sugar-phosphate backbones have a direction (determined by the detailed shape of the sugars). In most configurations of DNA including the standard one, called B-DNA, the two backbones run anti-parallel. However parallel stranded DNA can occur. The two backbones are linked through (relatively weak) hydrogen bonds between the base pairs, in general 2 for A-T and 3 for G-C basepairs respectively. Again in standard B-DNA, the three-dimensional conformation (minimum energy shape) has the paired bases in the interior of a right-handed 'helix? formed by the two backbones.

The three-dimensional shape is determined by the interplay between relatively weak rotational degrees of freedom in the otherwise strong backbone bonds, and the relatively weak hydrogen bonding between base pairs. The geometry of the base pairing and the helical backbones lead to what are called the major and minor grooves of B-DNA. Essentially the backbones forming the helix are not at the two ends of a diameter across the helix, but are offset. This leads to very important biochemical phenomena–for example there are major groove and minor groove binding proteins.

The parameters of B-form DNA are shown in the cartoon. All of these numbers must be taken with a grain of salt: in fact salt concentration can substantially alter them :-).

Even without changing solution conditions, these numbers must be interpreted as some form of average.

DNA in solution is fluctuating at all times, so some form of time average is involved.

The details of the shape of the helix are also believed to depend on the base pair sequence–see detailed discussion later—so a space or sequence average is also involved.

### 1.2.2 Some other forms of DNA

The relation between the sugar-phosphate backbone and each of the 4 bases is nearly identical so that the B-form of DNA can occur for just about any sequence of bases on either strand.

However B'-DNA occurs when one strand has several A residues in close proximity (and the complementary strand therefore has several T residues). In B'-DNA the two paired bases are no longer close to co-planar, but instead have an angle between their planes of about 20 degrees (This is sometimes called propeller twist.) B'-DNA is still a right-handed helix, but the presence of the so-called A-tracts is known to cause the centerline of the double helix to bend. There is still controversy as to whether the bend arises at localized kinks at the points where the B-form helix switches to the B'-form (the junction model), or whether the whole axis of the B'-form helical segment is curved (the wedge angle model).

In any case A-tracts are known to cause some sort of curvature, they are known to occur in natural DNA, and they seem to be play a very important biological role.

Short DNA molecules containing a number of phased A-tracts that provide an overall bend of 110 degrees or so in the helix axis play a crucial role in the continuum mechanics models of mini-circles that will be described later.

*More esoteric forms of DNA*

*A-DNA also is right-handed, but sugars are in a different conformation than for the B-DNA, and the two grooves are less deep.*

*Z-DNA has anti-parallel strands as in the B and B' forms, but in the Z-form a left handed helix is formed (as well as there being many other differences).*

*Z-form is difficult to achieve under physiological conditions. Sequences in which purines (A or G) alternate with pyrimidines (T or C) along one strand seem to enhance the formation of the Z-form, as does a loading yielding a twist stress favoring left handed twist over right handed (see discussion of mini-circles and plasmids below).*

*Parallel stranded or ps-DNA can be formed, particularly if both strands comprise only A or T base pairs. The helix is right handed, but the hydrogen bonds that join the base pairs are not the standard ones and the base pair geometry differs from the usual one.*

*In some circumstances (e.g. very special base pair sequences) it is also possible to form triplex helices containing three backbones (of course two of the backbones must be parallel!)*

*Quadruplex structures are also possible!*

### 1.2.3 The Length Scales of DNA

(back to the standard B-form double helix)

Each base pair is about 20x$10^{-10}$ meters wide (which is the geometrical diameter of the double helix).

Each base pair is about 3.4x$10^{-10}$ meters high (which is the contribution of each base pair to the length of the double helix).

But there are approximately $10^{10}$ base pairs or 3.4 meters of DNA in each of the cells in your body! Actually the total length of DNA in each cell is between 1 and 2 meters, all of which is not in one piece. Human DNA comes in 22 homologous pairs + X and Y chromosomes. Longest single piece of DNA is about 10 centimeters.

Understanding which bits of the $10^{10}$ base pairs are responsible for what genes, is the topic of the human genome project

10 cm of DNA of width $20 \text{x} 10^{-10}$ m is still very long and skinny.

Multiply by$10^6$ then: Diameter becomes 2mm and length $10^5$ m or 100km

A thin chalk line to Geneva and back.

The total volume of the DNA double helix is still rather small so it can be packed in individual cells with plenty of space left over. But it must be very organized so that it can do its job and, for example, be exactly duplicated when the cell divides.

Much known and conjectured about the structure of this organization. In humans the DNA is wrapped on nucleosomes (groups of small globular proteins). Then the composite fiber so formed is wrapped and coiled into another fiber, and so on, with the final arrangement or this hierarchy of fiber is known as the chromosome. Perhaps the most basic function of DNA is to code the proteins that make the cells work. There is a mapping from triplets (or a codon) of base pairs to the amino acids that make up proteins. This is the genetic code. Usually one gene encodes one entire protein, and a typical length of a gene is 500 to 600 base pairs (although there is certainly much variation in this length).

The length of scale of 500 bp or so is an important one for the mechanical properties of DNA. The length will re-appear later.

Given that the total number of base pairs in the human genome is $10^{10}$, it seems that to model say a few hundred bp sequence would be a rather modest goal. However a few hundred bp is still essentially beyond the scope of practical MD simulations. Each base pair has around 60 atoms so a 200 bp sequence involves around 12000 atoms in the DNA itself.

However an explicit treatment of solvent makes the number of molecules explode, and a practical current day simulation is limited to a few nanoseconds for a 20 or 30 base pair sequence. Thus the need for better multiscale models is apparent.

There are a number of different hierarchies of models available above the atomic one.

Bases or base pairs can be modelled as rigid bodies with potentials between the sub-units defined by summing atomistic potentials over the constituent bodies.

Models can also be based on larger units than individual base pairs, e.g. Monte Carlo simulations.

The DNA can be smoothed or averaged to yield a model as an elastic line (a system with an infinite number of degrees of freedom) and then re-discretized for simulations with the discretization chosen according to purely numerical analysis criteria. Experimental data allow the following conclusions to be drawn:

- DNA is big enough to be seen with various microscopy techniques. In particular cryo-EM allows 3D shapes to be measured, which are hopefully close to the shape in solution,
- DNA quite often occurs (or can be made into) closed loops, varying from 150 bp (or less) mini-circles, to naturally occurring plasmids of a thousand base pairs,
- DNA is rather stiff. For lengths of a few hundred base pair (compare typical length of a gene fragment) DNA apparently has a well-defined shape, in which the centerline of the double helix may be far from straight (cf. Tortillon cryo-pictures).

## 1.3 The specific problem of cyclization probability

Generally accepted that unstressed DNA can have quite large natural curvatures, especially when so called A-tracts (a particular base pair sequence) are present.

Issue is how such natural curvature effects cyclization. Three different unstressed shapes

1. straight, with no intrinsic curvature,
2. bent like a C,
3. or with the shape of a S,

will require different amounts of energy to cyclize, and phasing of the end base-pairs obviously play a role in the process, as the double helix closure requires the individual strands to remain anti-parallel, which prevents the formation of a Moebius type of ribbon.

Related questions arise in DNA-protein binding which is a basic biological mechanism that can produce sharp kinks in the DNA.

For example the 156 base pair molecule shown here has an approximately 110 degree bend in its minimum-energy stress-free uncyclized state.

See figure 1.

The open configuration is generated as the minimum energy or unstressed state according to a modified Trifonov wedge angle model. Some such discrete information coming from molecular biology is the input to the continuum model.

And it cyclizes as shown in figure 2.

The solid lines are the computed output of the continuum model, with the dots being the reconstructed wedge-angle equilibrium. Difference in energies between discrete and continuous models is 0.5%.

Biochemists can experimentally measure how the equilibrium constant between cyclized and uncyclized forms depends upon differences between unstressed shapes for various short DNA molecules.

(Actual measurement is ligated cyclization rates of cyclized vs dimer products)

According to various standard statistical mechanics formul the equilibrium constant is related to the difference in free energies between uncyclized and cyclized states.

The particular question we addressed is whether continuum rod computations of the internal (elastic) energy can correctly duplicate the experimentally measured differences for DNA molecules with differing unstressed shapes.

See figure 3.

The first semester of the course will explain the static model used in these computations and the limits of the model applied here.