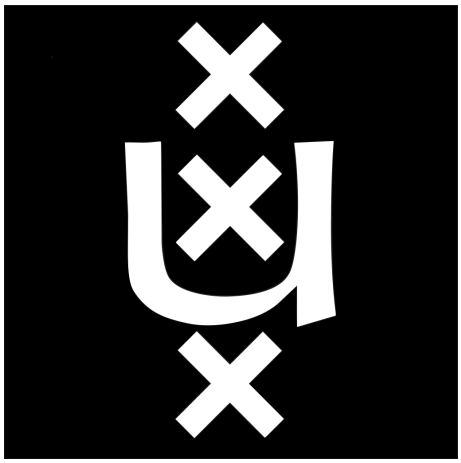

Optimal DNA alignment and the Lieb Liniger model

Connecting DNA strands, directed polymers and the Lieb Liniger model

Student Seminar Theoretical Physics

University of Amsterdam

6 May 2018

Roshell Haaker Marieke Kral

Contents

Why DNA strings are a path integral

When cells divide DNA is copied: the copy matches the ancestor up to some possible mutations, deletions or insertions of the DNA sequence. Different species are categorized based on their DNA sequence, making mutations of DNA important out of an evolutionary perspective. This makes the detection of mutual correlations between different DNA strands of interest. However, although DNA is build of only four bases, these bases make up a long sequence, e.g.: the sequence of the human genome consists of about 3 billion of these bases. Different attempts have been made to create an algorithm that detects the overlap between different DNA strands and in that regard it is key to find a method that can be used to tell us how much overlap there is between DNA strands. This is the field of the sequence alignment, which tackles the question how the similarity between sequences can be quantified and how dissimilarities can arise.
One thing to take into account there is that the evolution process involves insertions and deletions, resulting in a shift in the positions of the bases of the strands. Therefore one has to allow for gaps in the alignment so that the correlating regions can still align. Another question that arises is whether the aligned regions actually represent the same functionality, which is the question of biological relevance. This all boils down to a search of correct parameters to represent the accuracy of the alignment.
In this digest we will shine light on the field of sequence alignment and specifically on the method explored by Hwa and Lässig (1996). Although one method would be to empirically deduce the best parameters of the alignment from sequence pairs of which the functionality is known, they explore a different approach to the parameter selection problem. They use methods of statistical physics, starting with a rendition of the similarities of sequences as a path integral. This will lead to physics known from other fields of condensed matter, i.e. the Lieb Liniger Model, which in the end is able to give some insight in the alignment of DNA sequences.

1.1 Visualizing DNA strands

So how do they tackle to problem of quantifying the similarities between DNA sequences? If we look at two sequences $P = \{P_i\}$ and $Q = \{Q_j\}$ and define an ordered set of pairings (P_i, Q_j) and also define gaps $(P_i, -)$ and $(-, Q_j)$ we can distinguish three situations: a match $(P_i = Q_j)$,

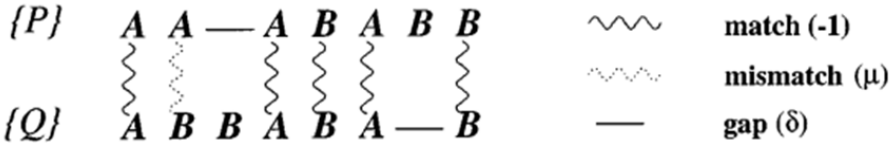

Figure 1.1: An example of a possible alignment, indicating matches, mismatches and gaps. Image from Hwa and Lässig (1996)

a mismatch ($P_i \neq Q_j$) and unpaired elements or gaps, see fig 1.1. In this case only two letters, A and B, are considered but for DNA strands this can easily be generalized to the 4 bases of DNA strands A, C, G or T as was done in (Drasdo et al., 1998). This depicted in fig. 1.2
The optimal alignment of the two strands is determined by minimizing a cost function or energy function E. A cost is assigned to a match, mismatch or gap and the total cost is calculated as the sum over all costs. The costs are chosen in a way to favour matches over mismatches and gaps. A gap contributes -1 to the energy, a mismatch $\mu > 0$ and a gap $\delta > 0$. The extent to which the alignment captures the mutual correlations is called the fidelity.
The basic idea is to start with two known strands, the "ancestors". Next a part of the elements from the ancestors are changed is a specific way to produce daughter sequences P_i and Q_j. The sequences are aligned by minimizing the energy function and the fidelity is quantified as the fraction of correctly discovered pairs ($P_i = Q_j$). For long stands two regimes can be distinguished that are separated by a critical transition: a low similarity phase of effectively zero fidelity and a high similarity phase where the fidelity has a finite value.

For the analysis of the alignment of the sequences we can look at a two dimensional grid formed by the elements of P_i and Q_j. Alternatively the cells of the grid can be labbeled by (r,t) where $r \equiv i - j$ and $t \equiv i + j$ see fig 1.3. A pairing is depicted as a diagonal bond, a mismatch as a diagonal dashed bond and a gap as an horizontal $(P_i, -)$ or vertical line $(-, Q_j)$. The figure shows how similarities between the sequences can be visualized as a path that is directed along the t coordinate: once there is a match or mismatch in the sequence a step is taken on the grid, moving forward along the diagonal. When there is gap, although there is still movement, one deviates from the original path in the horizontal or vertical direction.

1.2 Quantifying the path

Because the mutual similarities of the daughter strands are known, it is possible to draw a target path R_0 on the lattice. This is the optimal path that we would like to find. The optimal

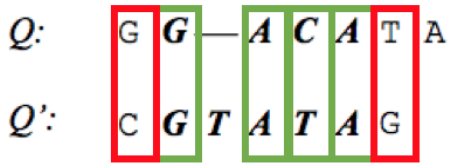

Figure 1.2: An example of a possible alignment. Green indicates a match, red a mismatch and the rest are either unpaired elements or gaps. Image adapted from Drasdo et al. (1998)

alignment path r_0 is the one we find after minimizing the energy, so the outcome path of the method. If the method works well, R_0 and r_0 will overlap for most of the path, if it doesn't there will be few overlaps. Visually this would mean that on the grid a path would be formed around the diagonal. Thus, if the path deviates from the diagonal target path, the alignment is less 'optimal'. If one wants to quantify the similarities between the sequences, one can thus regard the displacement from the diagonal path.
In the low-similarity phase the path is super-diffusive and the fluctuations around the target path $\delta r(t) \equiv r_0 - R(t)$ scale as $|\delta r(t)| \sim t^{2/3}$. That is depicted in the right panel of fig 1.4 In the high-similarity phase the fluctuations around the target path are finite $|\delta r(t)| \sim \xi_\perp$ and the path is localized around the target path, see fig 1.4 left panel. The path r_0 is a faith-full representation of the target R_0. The fidelity of the alignment is given by the overlap of the two, given y the number of intersections per unit of t.

Another possibility would be to look at the length of the path. Can the length of the path tell us anything about how 'good' the alignment of a path is? First we need to define more precise what the length of the path is. If we take a look at how our coordinate system is defined we see that the length in terms of the number of matches N_+, mismatches N_- and gaps N_g can be quantified as $L = 2N_+ + 2N_- + N_g$ (Drasdo et al., 1998). We thus notice that different lengths could respond to a different ratio in mismatches, matches and gap. Working with the energy function, which assigns scores to matches, mismatches and gaps, is therefore the way forward.

Hwa and Lässig (1996) set the energy function up as follows. The product of the elements P_i and Q_j are binary and can take values -1 or +1 depending on whether there is a match or a mismatch. Every step on the grid lowers the energy, no matter whether you are dealing with a match or a mismatch, and that energy corresponds to the energy value associated to a gap. One can therefore also view a gap as subtracting no additional energy to the energy function.

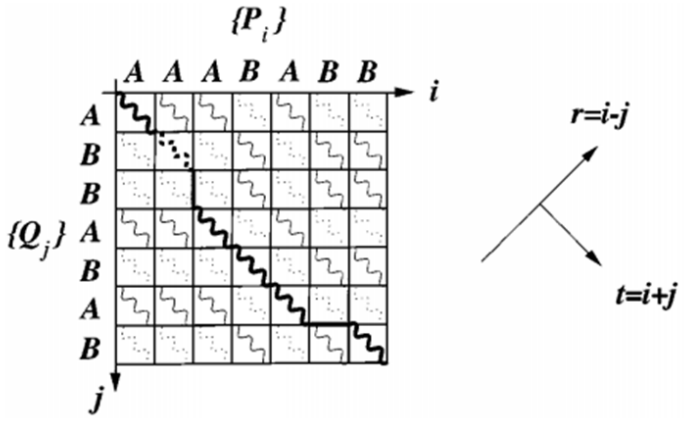

Figure 1.3: A match in the sequences results. Image from Drasdo et al. (1998)

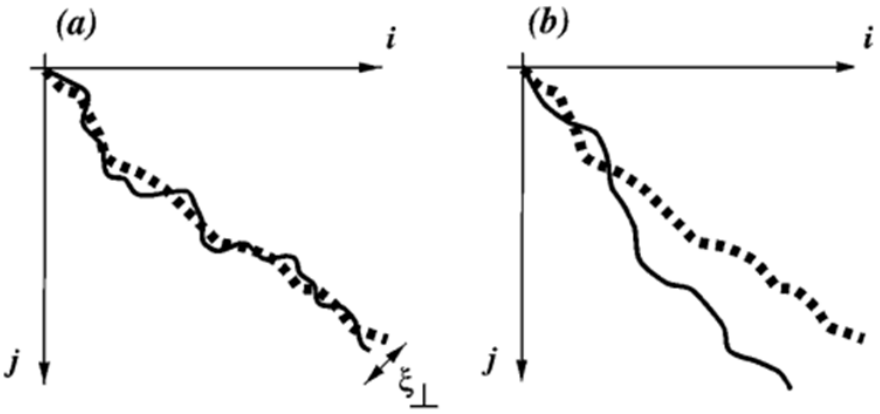

Figure 1.4: Example of alignments depicted as paths. The target path is depicted with squares, the continuous line shows a possible actual path. a) shows the high similarity phase, b) shows the low similarity phase. Image from Hwa and Lässig (1996)

That means that the energy function for one step on the grid can be written as

$$v_{r,t} \equiv -J - \Delta P_{i=((r+t)/2} Q_{j=(r-t)/2} \tag{1.1}$$

where $J \equiv 2\delta - (\mu - 1)/2$ and $\Delta = (\mu + 1)/2$ are the effective gap and mismatch costs. We thus see that a gap (or vertical/horizontal step on the grid) lowers the energy by 2δ, a match by $2\delta + 1$ and a mismatch by $2\delta - \mu$. The total energy E of a path $r(t)$ is a summation over all these individual steps.

1.3 Low-similarity phase

To model the low-similarity phase (right panel of fig 1.4) Hwa and Lässig (1996) used sequences P_i and Q_j from an ensemble of unbiased random sequences with no mutual correlations, thus

$$\overline{P_i} = \overline{Q_j} = 0 \tag{1.2}$$

$$\overline{P_i P_{i'}} = \delta_{i,i'} \tag{1.3}$$

$$\overline{Q_j Q_{j'}} = \delta_{j,j'} \tag{1.4}$$

$$\overline{P_i Q_j} = 0 \tag{1.5}$$

Furthermore, they use a random potential, the pairing potential $v_{r,t}$ for equation 1.1 with average $\overline{v_{r,t}} = -J$ and variance

$$\overline{v_{r,t} v_{r',t'}} = J^2 + \Delta^2 \overline{P_i Q_j P_{i'} Q_{j'}} = J^2 + \Delta^2 \delta_{i,i'} \delta_{j,j'} = J^2 + \Delta^2 \delta_{r,r'} \delta_{t,t'} \tag{1.6}$$

In order to derive the large scale fluctuations on the alignment paths, Hwa and Lässig (1996) look at the partition function in the path integral representation. They show that the action of the path integral take the form

$$S = \int d\tau \left[\frac{\gamma}{2} \dot{\rho}^2 + \eta(\rho(\tau), \tau) + \mathcal{O}(\dot{\rho}^4, \eta \dot{\rho}^2) \right] \tag{1.7}$$

where γ is a finite line tension, and $\eta(\rho, \tau)$ a random potential with second moment

$$\overline{\eta(\rho, \tau) \eta(\rho', \tau')} \sim \Delta^2 \delta(\rho - \rho') \delta(\tau - \tau') \tag{1.8}$$

This action and potential might not seem familiar at this point. It is in the universality class of a directed path in a two dimensional Gaussian random potential. One of the properties of that class is that fluctuations of the optimal path are super-diffusive with

$$\overline{[r_0(t) - r_0(t')]^2} \simeq A |t - t'|^{4/3} = A |t - t'|^{2\zeta} \tag{1.9}$$

where ζ is called the roughness parameter. Hwa and Lässig (1996) have also tested this model numerically and found that is in accordance with this picture.

In the rest of this digest we will show how one can arrive at this conclusion, using a mapping to the Lieb Liniger model. We will therefore solve this model to show that the typical fluctuations of the optimal path $r_0(t)$ are indeed super-diffusive with roughness parameter $\frac{2}{3}$. First however, we will turns to techniques known from the directed polymer problem.

How the path integral of a DNA string relates to directed polymers

In the previous chapter we have shown that the problem of optimal DNA alignment can be depicted as finding the optimal path in a 2D-grid, where this optimal path is determined by the energy function. This implies that if we go to continuum limit we can write down the following, general partition function:

$$Z(x,t) = \int_{x(0)=0}^{x(t)=y} Dx e^{-\frac{1}{T}\int_0^t d\tau \left(\frac{\kappa}{2}\left(\frac{dx}{d\tau}\right)^2 + V(x(\tau),\tau)\right)}, \tag{2.1}$$

where the first term in the exponent describes the kinetic part and the second term the potential. The wonderful thing now is, that the partition function describing a directed polymer in a random potential is actually the same as this partition function.

What is a directed polymer? It is an elastic polymer stretched between two fixed point while there is a random environment. This polymer too can be seen as the path of a particle moving from one point to the other, with no loops, as was the case for the DNA sequences.
What is 'a random environment'? This refers to the fact that the potential in the field where the polymer is stretched is an uncorrelated, centered Gaussian potential. Thus the properties of this potential are:

$$\overline{V(x,t)} = 0 \tag{2.2}$$

$$\overline{V(x,t), V(x',t')} = \delta(t-t')R(x-x') \tag{2.3}$$

This indicates that our potential is ultra local: there is no interaction between different 'particles' (or steps in the path) in our grid.

This formalism is helpful, because as it turns out, the directed polymer problem is analytically solvable. To see this we begin to apply the so-called 'replica trick'. We follow the strategy as in Nishimori (2001).

What does this mean? First we need to set up the configurational average of the nth power

of the partition function,

$$\overline{Z^n} = \int Dv_j P(v_j) Z^n \tag{2.4}$$

$$= \int Dv_j P(v_j) \int_{x(0)=0}^{x(t)=y} Dx e^{-\frac{1}{T}\sum_{j=1}^{n}\int d\tau\left(\frac{\kappa}{2}\left(\frac{dx_j}{d\tau}\right)^2 + V(x_j(\tau),\tau)\right)} \tag{2.5}$$

$$= \int \int_{x(0)=0}^{x(t)=y} Dx Dv_j P(v_j) e^{-\frac{1}{T}\sum_{j=1}^{n}\int d\tau V(x_j(\tau),\tau)} e^{-\frac{1}{T}\sum_{j=1}^{n}\int d\tau \frac{\kappa}{2}\left(\frac{dx_j}{dt}\right)^2} \tag{2.6}$$

Where $P(\nu_j)$ is a Gaussian probability distribution. Now we can use the fundamental identity

$$\langle e^{-mV}\rangle_V = e^{\frac{m^2\sigma^2}{2}} \tag{2.7}$$

where V is a zero mean Gaussian random variable with variance σ^2 and $m = -\frac{1}{T}\int d\tau$ to compute the averaging. To convince you (and ourselves) we show in detail that this identity holds.

For a Gaussian distribution with normalization N the average is computed as

$$\begin{aligned}
N < e^{-mV'} > &= \int dV' P(V') e^{-mV'} \\
&= \int dV' e^{\frac{(V'-\mu)^2}{2\sigma^2}} e^{-mV'} \\
&= \int dV e^{\frac{-V^2}{2\sigma^2}} e^{-m(V+\mu)} \\
&= e^{-m\mu} \int dV e^{-\frac{V^2}{2\sigma^2} - mV} \\
&= e^{-m\mu} \int dV e^{\frac{1}{2\sigma^2}(V+m\sigma^2)^2} e^{\frac{1}{2\sigma^2}(m\sigma^2)^2} \\
&= e^{-m\mu} e^{\frac{m^2\sigma^2}{2}} \int dV e^{\frac{1}{2\sigma^2}(V+m\sigma^2)^2}
\end{aligned}$$

where $P(V')$ is a Gaussian probability distribution, m and V as specified above and $\mu = 0$ in our case. Since the final Gaussian integral has a constant value, we can choose our normalization N such that it is equal to this value. If we also set $\mu = 0$ (now we are centred around zero) we arrive at the desired result of 2.7.

We know V is a Gaussian potential that is centered around zero and that the second moment is given by eqn. 2.3. If we also use eqn. 2.6 and eqn. 2.7 so we can rewrite $\overline{Z^n}$ as

$$\overline{Z^n} = \int Dx e^{\frac{1}{2T^2}\sum_{j=0}^{n}\int d\tau d\tau' \delta(\tau-\tau') R(x-x')} e^{-\frac{1}{T}\sum_{j=1}^{n}\int d\tau \frac{\kappa}{2}\left(\frac{dx}{d\tau}\right)^2} \tag{2.8}$$

$$= \int Dx e^{\frac{1}{2T^2}\sum_{j=0}^{n}\int_0^t d\tau R(x-x')} e^{-\frac{1}{T}\sum_{j=1}^{n}\int_0^t d\tau \frac{\kappa}{2}\left(\frac{dx}{d\tau}\right)^2}. \tag{2.9}$$

We recognize the form of this equation: a path integral with as an integrant an exponent containing an integral over time and a Lagrangian. So equation 2.9 has the form:

$$\overline{Z^n} = \int Dx e^{S_n} \tag{2.10}$$

$$S_n = \int d\tau \sum_{j=0}^{n} L, \tag{2.11}$$

with S being the action. In Feynman and Hibbs (1965) the tools are provided to find the corresponding Hamiltonian, which we use while setting $\hbar$ to 1. This method is based on considering a displacement in time. We consider a wavefunction $\psi(t)$ and a wavefunction over the same space but earlier by a time δ, so $t_\delta = t - \delta$, so we have:

$$\psi(t) = \psi(t_\delta) + \delta \frac{\partial \psi}{\partial t} = \psi_\delta + \delta \frac{\partial \psi}{\partial t}. \tag{2.12}$$

Since we want to find an H to satisfy

$$\frac{\partial \psi}{\partial t} = H\psi \tag{2.13}$$

we can find an expression for the Hamiltonian by considering the difference between the wavefunctions $\psi(t)$ and ψ_δ. So combining equations 2.13 and 2.12 we find:

$$\delta H\psi = \delta(\psi(t) - \psi(t_\delta)) \tag{2.14}$$

$$\langle\chi| H |\psi\rangle = \delta(\langle\chi| 1 |\psi\rangle - \langle\chi| 1 |\psi_\delta\rangle) \tag{2.15}$$

Where in the last line we written the expression in terms of the transition amplitude, which is defined as

$$\langle\chi| 1 |\psi\rangle = \int\int \chi*(x_N, t_N) K(x_N, t_N; x_i, t_i) \psi(x_N, t_N) dx_1 dx_N \tag{2.16}$$

$$K(x_N, t_N; K(x_i, t_i) = \int Dx(t) e^S \tag{2.17}$$

and in our case

$$S = \int dt \sum_{i}^{N} \left(-\frac{\kappa}{2T} (\frac{\partial x}{\partial t})^2 + \frac{1}{2T^2} R(x_i - x') \right). \tag{2.18}$$

So what will be the difference between the transition amplitude from ψ_δ and ψ? It comes from the shift δ in time and looking in the expressions above, where the time dependence can be found in the action, this difference depends on how the action changes as a function of time. One can find that in the end:

$$\delta \langle\chi| 1 |\psi\rangle - \langle\chi| 1 |\psi_\delta\rangle = \langle\chi| \frac{\partial S}{\partial t} |\psi\rangle \delta, \tag{2.19}$$

which is the result one would also retrieve in the classical case. So in this case:

$$\langle\chi| H |\psi\rangle = \sum_{i}^{N} (-\frac{\kappa}{2T} \langle\chi| (\frac{\partial x}{\partial t})^2 |\psi\rangle + \frac{1}{2T^2} \langle\chi| R(x_i - x') |\psi\rangle) \tag{2.20}$$

From basic quantum mechanics we know the following:

$$\left\langle \frac{\partial x}{\partial t} \right\rangle = \frac{\partial \langle x \rangle}{\partial t} \tag{2.21}$$

$$= \frac{-i\hbar}{m} \int \chi * \frac{\partial \phi}{\partial x} \partial x \tag{2.22}$$

$$\left\langle \left(\frac{\partial x}{\partial t} \right)^2 \right\rangle = -\frac{\hbar^2}{m^2} \int \chi * \frac{\partial^2 \phi}{\partial x^2} \partial x \tag{2.23}$$

Our equivalent of m is k/T so in the end equation 2.20 can be used to find:

$$\hat{H}_n = -\frac{T}{2\kappa} \sum_i^N \partial_{x_i}^2 - \frac{1}{2T^2} \sum_{ij} R(x_i - x_j) \tag{2.24}$$

so we have found the H_n that satisfies the equation:

$$\partial_t \overline{Z^n} = -H_n \overline{Z^n} \tag{2.25}$$

which can be labeled the Feynman-Kac equation, since Feynman and Kac famously derived this.

So far we have thus shown how n replications of the partition function of our system can be disorder averaged. Furthermore we have shown that the disorder averages $\overline{Z_n}$ satisfy 2.25 with Hamiltonian 2.24. This Hamiltonian contains an attractive interaction term $R(x)/T^2$. Maybe at this point you think, so what? In the next chapter we will tell you what!

How the directed polymer problem maps onto the Lieb Liniger model (with attrictive interaction)

As we have seen in the previous chapter, we can think about DNA strands as path integrals and they turn out to be described by the same partition function as the directed polymer problem. After disorder averaging the partition function while using the properties of the potential as described in 2.3 we showed how the disorder averaged partition function Z_n satisfies the Feynman-Kac equation 2.25. If we now define a mapping

$$\begin{aligned} x &= \frac{T^3}{\kappa}\tilde{x} \\ t &= \frac{2T^5}{\kappa}\tilde{t} \end{aligned} \tag{3.1}$$

apply this to eqn 2.25 and take the limit of $T \to \infty$ we can map our Hamiltonian onto the Hamiltonian we know from the Lieb-Liniger (LL) model. Let us show the details of this mapping.

First use 3.1 to see that $\partial_t = \frac{\kappa}{2T^5}\partial_{\tilde{t}}$. Next we apply the mapping to 2.24. Calabrese et al. (2010) mention that $R(x)$ is a decaying function of the correlation scale r_f and from that find that $\widetilde{R}(\tilde{x}) = 2T^3\kappa^{-1}R(T^3\kappa^{-1}\tilde{x})$. They use that expression to map $R(x_i - x_j)$ to $R(\tilde{x}_i - \tilde{x}_j)$ and find the following expression for H_n:

$$\begin{aligned} H_n &= -\frac{T}{2\kappa}\sum_{i=1}^{n}\partial_{x_i}^2 - \frac{1}{2T^2}\sum_{ij}R(x_i - x_j) \\ &= -\frac{T}{2\kappa}\left(\frac{\kappa}{T^3}\right)^2\sum_{i=1}^{n}\partial_{\tilde{x}_i}^2 - \frac{1}{2T^2}\frac{\kappa}{2T^3}\sum_{ij}R(\tilde{x}_i - \tilde{x}_j) \\ &= -\frac{\kappa}{2T^5}\sum_{i=1}^{n}\partial_{\tilde{x}_i}^2 - \frac{\kappa}{4T^5}\sum_{ij}R(\tilde{x}_i - \tilde{x}_j) \end{aligned}$$

Now 2.25 is satisfied with

$$H = -\sum_{i=1}^{n}\partial\tilde{x}_i^2 - \frac{1}{2}\sum_{ij}\widetilde{R}(\tilde{x}_i - \tilde{x}_j) \tag{3.2}$$

Next they take the limit $T \to \infty$. When $T \to \infty$, $\tilde{r}_f \to 0$. From that they find $\widetilde{R}(\tilde{x}) \to 2\bar{c}\delta(z)$ with $\bar{c} = \int duR(u)$. If we also choose the interaction parameter c as $c = -\bar{c}$ we have

$$H_{LL} = -\sum_{j=1}^{n} \frac{\partial^2}{\partial x^2} + 2c \sum_{1 \leq i < j \leq n} \delta(x_i - x_j), \tag{3.3}$$

the Hamilatonian we know from the Lieb-Liniger model with attractive interaction.

How the Lieb Liniger model can be solved with the Bethe Ansatz

The Lieb Liniger model can be and has been solved (see Halpin-Healy and Zhang (1995)). From these we know that for the Lieb Liniger model with attractive interaction we can write the Bethe equations as (Doussal and Calabrese, 2012; Caux, 2018)

$$e^{i\lambda_\alpha} L = \prod_{\beta \neq \alpha} \frac{\lambda_\alpha - \lambda_\beta - ic}{\lambda_\alpha - \lambda_\beta + ic}, \qquad \alpha = 1, ..., N. \tag{4.1}$$

If we now look at a complex rapidity $\lambda_\alpha = \lambda + i\eta$ the Bethe equation for that rapidity is

$$e^{i\lambda_\alpha} L = e^{i\lambda L - \eta L} = \prod_{\beta \neq \alpha} \frac{\lambda_\alpha - \lambda_\beta - ic}{\lambda_\alpha - \lambda_\beta + ic} \tag{4.2}$$

If we consider now N to be finite but we take L to infinity there are two cases we can distinguish. In the case of $\eta > 0$ we see that the left hand side of 4.2 goes to zero. In order for the right hand side to go to zero we must choose the distance between the rapidities α and β to be c on the imaginary axis. Thus $\lambda_{\alpha'} = \lambda_\alpha - ic + \mathcal{O}(e^{-\eta L})$. If $\eta < 0$ we see that the left hand side of 4.2 goes to infinity and so the denominator on the right hand side must go to zero. So the rapidities have to be $-ic$ apart, we have $\lambda_{\alpha'} = \lambda_\alpha + ic + \mathcal{O}(e^{-|\eta| L})$. We can picture this as clusters of evenly spaced rapidities and call them strings (Kardar, 1987; Caux, 2018). See figure 4.1.

Now we can try to find the ground state energy and the eigenstates for the Hamiltonian of the LL model (see equation 3.3). From Doussal and Calabrese (2012) we know that

$$\Psi_n \propto e^{-\frac{c}{2} \sum_{i<j} |x_i - x_j|} \tag{4.3}$$

might be a solution with energy

$$E_n = \frac{c^2}{12} n(n^2 - 1) \tag{4.4}$$

so let's check that, using $H_{LL}\psi = E_n \psi$.

$$\partial_x^2 \Psi_n = \partial_x [\partial_x [-\frac{c}{2} \sum_{i<j} |x_i - x_j|] \Psi_n] = \tag{4.5}$$

$$\partial_x [-\frac{c}{2} \sum_{j \neq k} sgn(x_j - x_k) \Psi_n] = \tag{4.6}$$

$$-c \sum_{j \neq k} \delta(x_j - x_k) \Psi_n + (-\frac{c}{2} \sum_{j \neq k} sgn(x_j - x_k)^2 \Psi_n \tag{4.7}$$

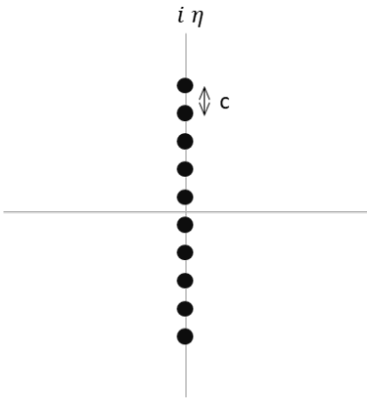

Figure 4.1: Visualization of the rapidities for an attractive interaction

. For the Lieb Liniger model with attrictive interaction, the rapidities are all aligned with spacing ic, and thus form strings.

now we sum over all j to get the first term of 3.3.

$$\sum_{j=1}^{n} \partial_x^2 \Psi_n = -2c \sum_{j<k} \delta(x_j - x_k)\Psi_n + \frac{c^2}{4} \sum_{j=1}^{n} \sum_{j\neq k} sgn(x_j - x_k)^2 \Psi_n. \tag{4.8}$$

The last term can be re-written:

$$\sum_{j=1}^{n} \sum_{j\neq k} sgn(x_j - x_k)^2 = \tag{4.9}$$

$$\sum_{j=1}^{n} \left(\sum_{k>j} (-1) + \sum_{k<j} (+1) \right) \tag{4.10}$$

if we picture the particles lying in one line as in fig 4.1 it is clear that there are (n-j) k's that are bigger than j and (j - 1) k's smaller than j so we get

$$\sum_{j=1}^{n} \left(\sum_{k>j} (-1) + \sum_{k<j} (+1) \right) = \tag{4.11}$$

$$\sum_{j=1}^{n} [-(n-j) + j - 1]^2 = \tag{4.12}$$

$$\sum_{j=1}^{n} (2j - n - 1)^2 = \frac{n(n^2-1)}{3}. \tag{4.13}$$

If we plug equation 4.13 into 4.8 we see that $H_{LL}\psi = E_n\psi$, thus Ψ_n and E_n are indeed an eigenstate and eigenvalue of the Lieb Liniger model. The fact that we have shown that Ψ_n is an eigenstate of the Lieb Liniger model does not mean that $E_n$is the ground state energy. However, from the lecture notes Caux (2018) we know that this is indeed the ground state energy.

From Kardar (1987) we know that for a system of length L $\overline{Z^n(L)} \propto e^{-E_n L}$ where E_n is the ground state energy of the Hamiltonian 3.3. So for this system we can also express the disorder averaged partition function as $\overline{Z^n(L)} \propto e^{-E_n L} = e^{\frac{c^2}{12} n(n^2-1)L}$. In the next chapter we will use this form to find the roughness parameter $\zeta = \frac{2}{3}$. To understand this parameter, we first need to take a look at the free energy of this system.

How the solution of the Lieb Liniger model tells us something about the roughness of the strings

In the previous chapter, we have shown that the disorder averaged partition function can be described in terms of the ground state energy. However, the partition function is also related to free energy $\overline{F}$ of the system, by $\ln Z = -F$. It is therefore interesting to take a look at various properties of $\ln Z$ since it will tells us how the free energy of the system behaves. We can therefore write $\overline{Z^n}$ in a different form:

$$\overline{Z^n} = \overline{e^{nlnZ}} = e^{\sum_j^\infty \frac{C_j(\overline{\ln Z})n^j}{j!}}. \tag{5.1}$$

Here we expanded $\overline{Z^n}$ in terms of his cumulants C_j, as is done by Kardar (1987). $\overline{Z^n}$ is called the moment generating function and $K(n)$ the cumulant generating function. So

$$K(n) = \sum_{j=1}^{\infty} C_j(\overline{lnZ})\frac{n^j}{j!} = \overline{nlnZ} \tag{5.2}$$

Cumulants describe various properties of a function. For example, the first cumulant C_1 is also known as the mean, the second cumulant C_2 as the variance and C_3 as the third moment (skewness). C_j can be found by

$$C_j = \frac{\partial^j K(n)}{\partial n^j}|_{n=0} \tag{5.3}$$

Using our cumulant generating function, we can derive the expressions for the different cumulants. First we Taylor expand Z^n in n:

$$Z^n \approx 1 + n\ln(Z) + \frac{n^2}{2}\ln\big(Z^2\big) + \frac{n^3}{6}\ln\big(Z^3\big), \tag{5.4}$$

which we can use to expand $\overline{\ln(Z^n)}$

$$\overline{\ln Z^n} \approx \overline{\ln\left(1 + n\ln(Z) + \frac{n^2}{2}\ln(Z^2) + \frac{n^3}{6}\ln(Z^3)\right)} \tag{5.5}$$

since we know the Taylor expansion of $\ln(1+x)$ to go (up to third order) as $x - \frac{1}{2}x^2 + \frac{1}{3}x^3$. In this case, the last three terms in equation 5.4 are our equivalent to x. Writing down only the

terms up to the third order, the expansion for the cumulant generating function thus results in:

$$\overline{\ln Z^n} \approx n\overline{\ln Z} + \frac{n^2}{2}\overline{\ln Z^2} + \frac{n^3}{6}\overline{\ln Z^3} \tag{5.6}$$

$$- \frac{n^2}{2}\overline{\ln Z}^2 - \frac{n^3}{4}\overline{\ln Z^2}\overline{\ln Z} - \frac{n^3}{4}\overline{\ln Z^2}\overline{\ln Z} \tag{5.7}$$

$$+ \frac{n^3}{3}\overline{\ln Z}^3. \tag{5.8}$$

So the first three cumulants become:

$$\frac{\partial\overline{\ln Z^n}}{\partial n}|_{n=0} = \overline{\ln Z} \tag{5.9}$$

$$\frac{\partial^2\overline{\ln Z^n}}{\partial n^2}|_{n=0} = \overline{\ln Z^2} - \overline{\ln Z}^2 \tag{5.10}$$

$$\frac{\partial^3\overline{\ln Z^n}}{\partial n^3}|_{n=0} = \overline{\ln Z^3} - 3\overline{\ln Z^2}\overline{\ln Z} + 2\overline{\ln Z}^3 \tag{5.11}$$

$$\tag{5.12}$$

However, we also know from the previous chapter that

$$\overline{Z^n} \propto e^{E_n L} = e^{\frac{c^2}{12}n(n^2-1)L} \tag{5.13}$$

and from equation 5.1 that $C_1 \sim n$, $C_2 \sim n^2$ and $C_3 \sim n^3$. This means that our second cumulant is zero and only the terms from 5.2 that scale with n and n^3 can be non zero. So the fluctuations around the mean will be given by the third cumulant.

We know return to the free energy of the system. We know $\overline{\ln Z} = -\overline{F}$ where $\overline{F}$ is the averaged free energy. The fluctuations in the free energy are given by the third cumulant. If we compare the expression of the third cumulant to equation 5.13 we can see that

$$C_3 \sim \overline{F}^3 \sim L \tag{5.14}$$

so we find that

$$\overline{F} \sim L^{1/3} \tag{5.15}$$

Now we want to find the fluctuations in x, the transverse direction. If we look at the simple picture from fig 5.1 we can see that if we would allow for transverse fluctuations from the original path L, we would get a new path, L_{new} of length $2y$. Now $y^2 = x^2 + (\frac{L}{2})^2$ so $L_{new} = L\sqrt{1+4(\frac{x}{L})^2} \simeq L + 2L(\frac{x}{L})^2 \rightarrow L_{new} = L + 2\frac{x^2}{T}$. So we can write

$$L_{new} - L = 2\frac{x^2}{T} \sim \Delta F \sim L^{1/3} \rightarrow x \propto L^{2/3}. \tag{5.16}$$

Thus, the typical fluctuations don't remain bounded but increase as a function of L as

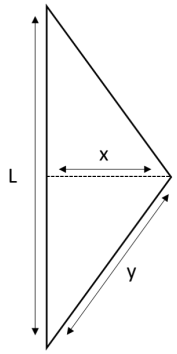

Figure 5.1: Simple picture to show the relation between a path without fluctuations (L), fluctuations in the transverse direction (x) and the new path ($L_{new} = 2y$)

$$\overline{[x(L) - x(0)]^2} \sim L^{2\zeta} \tag{5.17}$$

with roughness exponent $\zeta = \frac{2}{3}$. This should be a familiar parameter, since it was the parameter found by Hwa and Lässig (1996) as we mentioned in chapter 1.

What we now know about DNA strands

We now have showed through the Lieb Liniger model that the roughness exponent for a directed polymer is 2/3. This parameter contains information about the scale on which variance on the path of the directed polymer occurs. What does this tell us about our original problem, the DNA strands?

From chapter 2 we know that the problem of DNA strands is reminiscent of the problem of directed polymers. So, the roughness component for DNA strands is also given by the value 2/3. This actually tells us something about the right image in figure 1.4: the deviations of the path will be behave as $L^{2/3}$. This tells us that the low simularity phase, the longer the DNA sequences become, the less overlap there will be, and that decay will go as $L^{2/3}$. We thus arrive at the same result as Hwa and Lässig (1996).

Bibliography

Calabrese, P., Doussal, P. L., and Rosso, A. (2010). Free-energy distribution of the directed polymer at high temperature. *EPL (Europhysics Letters)*, 90(2):20002. 10

Caux, J.-S. (2018). The bethe ansatz part i. *Institute of Physics and Delta Institute of Theoretical Physics, University of Amsterdam*. 12, 13

Doussal, P. L. and Calabrese, P. (2012). The kpz equation with flat initial condition and the directed polymer with one free end. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(06):P06001. 12

Drasdo, D., Hwa, T., and Lssig, M. (1998). A statistical theory of sequence alignment with gaps. 6:52–8. 2, 3, 4

Feynman, R. and Hibbs, A. (1965). *Quantum Mechanics and Path integrals*. McGraw-Hill Book Company. 8

Halpin-Healy, T. and Zhang, Y.-C. (1995). Kinetic roughening phenomena, stochastic growth, directed polymers and all that. aspects of multidisciplinary statistical mechanics. *Physics Reports*, 254:215–414. 12

Hwa, T. and Lässig, M. (1996). Similarity detection and localization. *Phys. Rev. Lett.*, 76:2591–2594. 1, 2, 3, 4, 5, 17, 18

Kardar, M. (1987). Replica bethe ansatz studies of two-dimensional interfaces with quenched random impurities. *Nuclear Physics B*, 290:582 – 602. 12, 14, 15

Nishimori, H. (2001). *Statistical Physics of Spin Glasses and Information Processing An Introduction*. Clarendon Press Oxford. 6