Determination of unzipping step-size distribution using Hidden Markov Models

Author: Laia Montraveta Jiménez.
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Fèlix Ritort
(Dated: 19 de enero de 2015)

Abstract: In this paper we explain the unzipping experiment as well as the basic theory of hidden Markov models (HMM). Using a specific type of HMMs, variable step-size HMM, in which the quantized distance variable is represented by a large number of states of the Markov model we try to determine the size distribution of the steps in the unzipping process.

I. INTRODUCTION

Using atomic force microscope, magnetic tweezers and optical tweezers it is possible to apply external forces to biomolecules. Since these tools let us experiment with forces of some pN and very low energies, measuring biomolecules'mechanical response we are able to determine molecular free energy and kinetic rates with good accuracy. One of the experiments carried out with optical tweezers is Unzipping.

Unzipping consists on pulling a DNA molecule for each strand measuring the force and the distance between the ends, which increases with the number of opened base pairs (bp) (see Fig. 1). Representing force vs distance curve (FDC) we obtain a very characteristic sawtooth pattern near 15pN (see Fig.2). The strong dependence of this pattern can be used to determine the DNA sequence[1]. Moreover, it can be used to find the specific places where proteins and enzymes are fixed to the DNA. It is really interesting since it is very useful in a lot of fields, for example, finding the selectivity of some anticancer drugs.

In this experiments we find cooperative unzipping-zipping regions (CUR). This regions are zones where several base-pairs of different length are involved in the transition, behaving like an all or nothing. This complicates the determination of the individual sequence.

The length of the CUR that separates contiguous intermediate states along the unzipping pathway have been determined adopting a Bayesian approach where at each experimental data point of distance-force they assign the intermediate state with more probabilities of containing it [10].

In these paper we are going to find the CUR length treating the problem as a variable step-size Hidden Markov Model (VS-HMM). Markov models assume that when we have different possible states, what state we go to next depends only on what state we are in. If we only have observations which are probabilistic functions of the state we are in, it is known as a Hidden Markov Model (HMM). HMMs have been widely used for speech recognition and in computational molecular biology, among others [2], [3], [7]. VS-HMM is a HMM improved in order to describe at the same time the molecular state and the position of a processive molecular motor [4].

We supppose treating the problem as a VS-HMM is going to make the study quicker and more automatic, giving the experimental data and obtaining directly the probability of every transition length. Moreover, I am going to use it as a initial point of my next project (*Mathematics Treball fi de Grau*) where we want to implement an algorithm that let us determine, using VS-HMM, the places where proteins and enzymes bind the DNA.

II. UNZIPPING EXPERIMENT

The data we are going to analyze is obtained from DNA unzipping experiments that have been done with a short DNA hairpin (490 bp).

DNA hairpin is formed by a single-stranded DNA sequence where the first n bases on one end are complementary to the last n bases on the other end taken in reverse order. In the middle of the strand, we have some more bases not complementary that are going to form the loop region. This structure is present in DNA and RNA molecules in vivo as well as in vitro. The experiments are done with hairpins because that allows us to do, undo and redo optical trapping experiments with the same molecule [6].

The experimental setup is formed by de DNA molecule we want to unzip, two doble-stranded DNA (dsDNA) called handles and two dielectric beads.

Each end of the DNA molecule is coupled to one of the handles and each handle is attached to a dielectric bead. One of these beads is sucked with a micropipette and the other is confined in an optical trap with potencial $V_b(x)$ generated by the laser beams. The position of the micropipette is considered fixed while the hairpin is pulled by moving the optical trap along the x axis at a constant speed v and the distance between two beams is measured (see Fig. 1).

As you can see in Fig. 1, R_{b1} and R_{b2} are the beads radius, x_h is the extension of dsDNA handles, d the hairpin end-to-end distance, x_n the released ssDNA length and x_b the position of the bead with respect to the center of the optical trap.

We can measure the distance between beads

$$\lambda(f,n) = 2x_h(f) + 2x_n(f,n) + d + x_b(f) + R_{b1} + R_{b2} \quad (1)$$

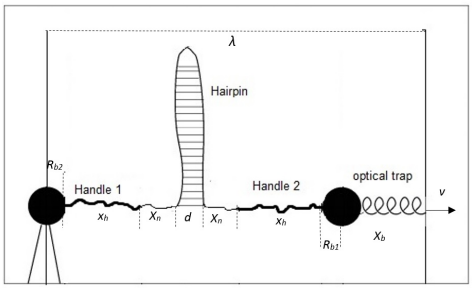

Fig. 1: Scheme of the experimental unzipping setup (not to scale). Configurational parameters are included in the picture: x_h, x_n, d, x_b, R_{b1}, R_{b2}, as the projection along the x axis of each element. Moreover, when the optical trap is moved we measure a force $f = k_b x_b$.

Moreover, optical trap can be considered harmonic, with $V_b(x_b) = \frac{1}{2}k_b x_b^2$ and the corresponding force is $f(x_b) = k_b x_b$, where k_b is the stiffness of the optical trap.

This pulling experiments give the force distance curves (FDC), corresponding to the force, f as a function of λ (see Fig.2).

Different contributions to Eq.1 can be obtained by using elastic models for biopolymers:

1. Taking R_{b1}, R_{b2} and d as a constants.

2. Released ssDNA x_n. We consider $d = 0,59nm$ and assume a Freely Jointed Chain (FJC), the simplest model of polymer conformation. It treats the chain formed by rigid subunits of identical length joined by perfectly flexible hinges.

3. Handles, x_h. We assume an elastic model, Worm Like Chain (WLC). Here dsDNA is treated as a continuum elastic body, describing its configuration as function of the position vector and the contour length.

4. Optical trap. $x_b = \frac{f}{k_b}$, $k_b = 0,066\frac{pN}{nm}$.

Since in the range of unzipping forces x_h is almost constant and what we want to measure is $x_{n,t} - x_{n,t+1}$, in the following sections we are going to work with $\lambda - x_b$.

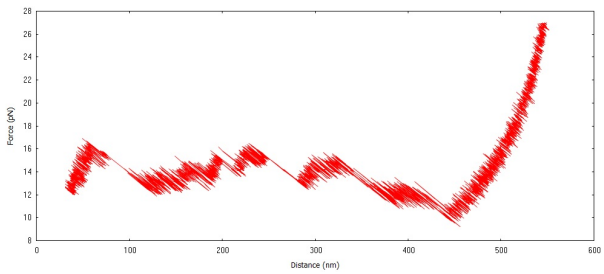

Fig. 2: FDC diagram obtained with force vs $\lambda - x_b$, the data we are going to use int he following sections. You can see the characteristic sawtooth pattern around 15 pN.

III. MARKOV CHAINS

A. discrete-time Markov Model

A discrete-time Markov Model is a collection of random variables $\{X_t\}_t$ $t \in \{0, 1, 2...\}$, taking values in the countable state space E and having the following property:

$$\forall t \geq 0 \text{ and } \forall\ i_0, i_1, \cdots, i_{t-1}, i, j \in E$$

$$P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \cdots, X_0 = i_0) = P(X_{t+1} = j | X_t = i) \quad (2)$$

We are going to consider only processes where $P(X_{t+1} = j|X_t = i)$ is independent of t, called Homogeneous Markov Chain (HMC)[8].

Each system is determined by:

- The countable state space formed by the N possible states the system could be, $E = \{S_1, S_2, \cdots, S_N\}$.
- The transition probabilities: $a_{ij} = P(X_{t+1} = j|X_t = i)$, $1 \leq i, j \leq N$.

Transition coefficients must obey standard stochastic constraints, so that $\forall 1 \leq i, j \leq N$: $a_{ij} \geq 0$ and $\sum_{j=1}^{N} a_{ij} = 1$.

B. Hidden Markov Model

The data recollected with the experiments described above, is affected by important fluctuations. We are going to treat this noise using HMM.

1. *Definition*

A Hidden Markov Model is a doubly stochastic process. It has an underlying stochastic process that is not observable, but can only be observed through another set of stochastic processes that produce the sequence of observed symbols. It has two defining properties[9]:

1. Y_t, the observation at time t, was generated by some process whose state X_t is hidden for the observer. Here we are assuming the observations are sampled at discrete and equal-spaced time intervals, so t can be an integer-valued time index.

2. The state of this hidden process satisfies the *Markov property* (Eq.2). And the observations also satisfy a *Markov property* with respect to the states: given X_t, Y_t is independent of the states and observations at all other time indexes.

2. Elements

The elements of a hidden Markov model are:

- N, the number of states in the model. The individual states are denoted $S = \{S_1, S_2, \cdots, S_N\}$ and the state at time t as q_t.
- M, the number of distinct observation symbols per state. We denote the individual symbols as $V = \{v_1, v_2, \cdots, v_M\}$.
- The state transition probability distribution $A = a_{ij}$ where

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], 1 \leq i, j \leq N. \quad (3)$$

- The observation symbol probability distribution in state j, $B = \{b_j(k)\}$, where $b_j(k) = P[v_k \text{ at } t | q_t = S_j]$, $1 \leq j \leq N$ and $1 \leq k \leq M$.
- The initial distribution, $\pi = \{\pi_i\}$, where $\pi_i = P[q_1 = S_i]$, $1 \leq i \leq N$.

The compact notation $\lambda = (A, B, \pi)$ is used to designate the complete parameter set of the model.

3. Variable step-size HMM

In order to obtain the chemical-kinetic model of a molecular motor's reaction cycle, a special implementation of HMM have been presented [4]. The particularity of molecular motors is that transitions are among a small number of molecular states (the equivalent to S_i) but the observable quantity is the molecule's position, X_k. The variable position reflects the accumulation of elementary steps of random size.

Variable step-size HMM (VS-HMM) describes both the molecular state and the position of a processive molecular motor.

Now, we consider a discret-time Markov model, too, $t \in \{1, 2, \cdots, T\}$. Moreover, the position is discretized, taking equidistant points with smaller separation than the noise standard deviation. $x_t \in \{1, 2, ..., M\}$, an integer representing the position. The molecular states, which for example can describe if the motor has ligands or not, are represented by $s_t \in \{1, 2, ..., N\}$. The characteristic of VS-HMM is that the molecular state and the position are taken together forming a composite state (s_t, x_t). The transition probability (Eq. 3) from the composite state (i, u) to $(j, u + w)$, in that model is:

$$a_{ij}(w) = P(s_{t+1} = j, x_{t+1} = u + w | s_t = i, x_t = u) \quad (4)$$

Satisfying $\sum_{j,w} a_{ij}(w) = 1$.

IV. DETERMINATION OF STEP-SIZE DISTRIBUTION

We have programmed, in C++ the *forward-backward algorithm*, that allows us to compute the log likelihood $L = \log P(O|\lambda)$ and do the reestimation of the parameters. We have needed to do some approximations in order to accelerate the computational time.

A. Adaptation and initial parameters

As we have said in the introduction, we want to determine the step-size distribution on the unzipping experiments. Our data have been obtained doing unzipping experiments with hairpins without ligands. Moreover, we consider the step-size is independent of the type of contiguous bases (A, C, T or G). So that we have a VS-HMM with a single state: $N = 1$ and multiple positions $M \geq 1$, where position represents the distance at time t.

The data we have corresponds, in each time, to the distance between beads with the correction of the optical trap,($\lambda - x_b$ from Fig.1). We have seen, in section II, that all the items from eq. 1 can be considered constant except x_n. So that, we are going to treat the data without more modifications.

We represent the observed position variable as $y_t = u_t + g_t$, where u_t represents the real position and g_t is a random variable having a distribution $\mathcal{N}(0, \sigma^2)$, representing additive noise. So, the probability of the observation y have the following Gaussian probability density :

$$b(y_t, u_t) = (2\pi\sigma^2)^{-1/2} exp\left(\frac{-(y_t - u_t)^2}{2\sigma^2}\right) \quad (5)$$

In order to reduce the time of iterations, we have considered $b(y, u) = 0$ for u s.t. $|y - u| > 15$. For this, if $min = min_t([y_t])$ and $max = max_t([y_t])$ (where $[x]$ means round x) true positions, $u(t)$, are taken in $\{min - 15, min - 14, \cdots, max + 14, max + 15\}$.
The step-sizes allowed are $\omega \in \{-W \cdots 0 \cdots W\}$ where $W = (max_t|y_t - y_{t+1}|) + 30 + 1$.

Our initial parameters are $\lambda = (\pi, A, \sigma)$:

- $\pi_u = \frac{1}{115 - min}$ for $u < 100$

  $\pi_u = 0$ otherwise.
- Using the results of [10]:
  $A[\omega] = p * 0{,}056|\omega|^{-0,4} exp\left(-\left(\frac{|\omega|}{60}\right)^3\right)$ where:

  $p = 0{,}2$ if $\omega < 0$
  $p = 0{,}8$ if $\omega > 0$
  $A[0] = 1 - \sum_{\omega \neq 0} A[\omega]$
- $\sigma^2 = 17$

We design our observation by $O = y_1, y_2, \cdots, y_T$.

B. Forward-backward algorithm

The forward variable is the probability of the observations up to time t and the molecule being in a particular state u, given the model lambda.

$$\alpha_t(u) = P(y_1, y_2, \cdots, y_t \text{ and } x_t = u|\lambda)$$

As we are working with very low numbers, we need to scale $\alpha_t(u)'s$ and $\beta_t(u)'s$ in order to prevent underflow in computation [11]. To do that, we use $c(t) = \frac{1}{\sum_u \alpha_t(u)}$. The forward variable is obtained with the following recursion:

1. $\alpha_1(u) = \pi_u b(y_1, u)c(1)$, u s.t. $|y_1 - u| \leq 15$
2. $\alpha_{t+1}(v) = b_{t+1}(v) \sum_{u=[y_t]-15}^{[y_t]+15} \alpha_t(u)A(v-u)c(t+1)$ $t = 1, 2, \cdots, T-1$

With that information we can obtain:

$$log[P(O|\lambda)] = -\sum_t c(t) \tag{6}$$

The backward variable is the probability of the observations from times $t+1$ to T, given the position u at time t and the model λ:

$$\beta_t(u) = P(y_{t+1}, y_{t+2}, \cdots, y_T|x_t = u, \lambda)$$

.

We obtain β as follows:

1. $\beta_T(u) = c(T)\ \forall u$.
2. $\beta_t(u) = c(t) \sum_{v=[y_{t+1}]-15}^{[y_{t+1}]+15} A(v-u)b(y_{t+1}, v)\beta_{t+1}(v)$ $\forall u,\ t = T-1, T-2, \cdots, 1$

Using α and β we can define another useful quantity:

$$\gamma_t(u) = \frac{\alpha_t(u)\beta_t(u)}{\sum_u \alpha_t(u)\beta_t(u)}$$

That gives us the probability of the distance to be u at time t.

Finally, we define $\xi_t(\omega)$, the probability of make a change of ω at time t:

$$\xi_t(w) = \frac{\sum_{u=[o_t]-15}^{[o_t]+15} \alpha_t(u)A(\omega)b(y_{t+1}, u+\omega)\beta_{t+1}(u+\omega)}{\sum_{u=[o_t]-15}^{[o_t]+15} \sum_w \alpha_t(u)A(\omega)b(y_{t+1}, u+\omega)\beta_{t+1}(u+\omega)}$$

C. Reestimation of model parameters

We want to obtain the distribution of ω that best fits our model, so optimize $A(\omega)$. It is not an easy problem, there is no known way to solve this problem analytically. What we are going to do, is find $\lambda = (\pi, A, \sigma^2)$ that maximises Eq. 6 locally using the Baum-Welch method [9].

Using our parameters $\lambda = (\pi, A, \sigma^2)$ we compute the new ones:

$$\pi_u^r = \gamma_1(u)$$

$$A^r(\omega) = \frac{\sum_t \xi_t(\omega)}{\sum_{t,\omega} \xi_t(\omega)}$$

$$\sigma^{r^2} = \frac{\sum_{t,u} \gamma_t(u)(y_t - u)^2}{T}$$

Now we define the reestimated model as $\lambda^r = (\pi^r, A^r, \sigma^{r^2})$. It has been proven that $P(O|\lambda^r) > P(O|\lambda)$, so, we have found a model that is more likely to have produced our observations.

If we iteratively use λ^r in place of λ and repeat the reestimations, we then can improve the probability of O being observed from the model until we reach some limiting point. However, we should note that forward-backward algorithm leads to a local maxima only, and that there can be many local maxima.

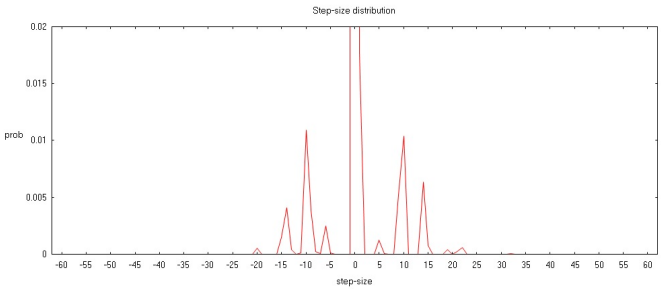

Fig. 3: One of the distribution obtained with the Baum-Welch algorithm. It is very high at 0, corresponding to no-movement and practically null for the rest.

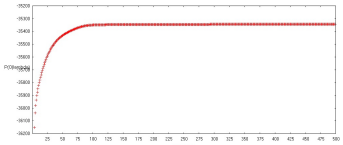

Fig. 4: Represents $\log(P(O|\lambda)$ as a function of the iteration. It starts growing very quickly an then stabilizes.

V. RESULTS

We have data obtained from three unzipping experiments and we have done 500 iterations of Baum-Welch algorithm in each one.

Effectively, with every iteration we obtain a λ that fits better our observation (see Fig. 4).In the reestimation of A, we expected to have something similar to

$A[\omega] = a|\omega|^b exp\left(-\left(\frac{|\omega|}{c}\right)^D\right)$, a power law with a superexponential cut off [10]. However, we have found the distribution of Fig. 3.It is near 1 for step of $0nm$, followed by 0.1 for ± 10-step and some other little pics in $\pm 5, \pm 15, \pm 20$. It is almost zero for the rest. Moreover, we have reestimated σ^2, the variance corresponding to the noise measured in our observations. We have found a value between $9{,}5nm^2$ and 10,4 depending on the experimental data processed.

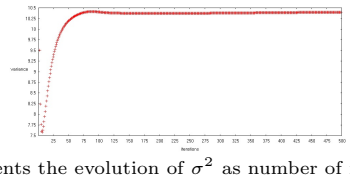

Fig. 5: Represents the evolution of σ^2 as number of iterations grow, like before, it starts growing quickly but at 150 iterations it is stabilized.

We have represented only the results obtained with one unzipping data because of clarity. The other two gives more or less the same: $A(0)$ almost 1, an decreasing pics around $A(\pm 10, \pm 15, \pm 5, \pm 20)$. The variance is around $9{,}5nm^2$ for both.

VI. CONCLUSIONS

The results obtained have been unexpected. Our goal was to find the distribution of the step-sizes in DNA unzipping experiments. We expected high probabilities for small step-sizes, rapidly decreasing to zero as the step-sizes raise following a power law with a superexponential cut off[10]. However, we have found a probability of almost zero for all possible values of these steps, except for zero that is practically one and some other little pics in $\pm 10, \pm 5, \pm 15\pm$ and $20nm$.

Taking into account previous studies, steps of less than $10bp$ are hardly detected, so we couldn't hope to find a good prediction on the interval $[-10 : 10]$. Moreover, our hairpin is shorter (490 bp vs. 2.2 kbp or 6.8 kpbs) so that we expect shorter jumps and dispose of less data (there are less jumps in total). However, we have worked with the distribution of the difference between consecutive measurements while [10] has worked with the distribution of CUR sizes (using the distance between following Gaussians that fit the histogram of opened bp). So that, we have expected to have lower probabilities. It would be great to try to do the same study taking absolute values and longer hairpins, trying to find the distribution of CUR lengths using VS-HMM. Moreover, we need to take into account we are working with continuous distribution but we have discretized the possible distance values arbitrarily, taking $1nm$ each time because it is comparable with the distance of two bp. In next studies should be great do the histograms of distance measurements and use them to determine *true position values* (the $u's$ corresponding to eq. 5).

We have obtained σ^2 between $9{,}5nm$ and $10nm$. It is a good value if we consider we are treating fluctuations but it could have had a big influence in our results. An improvement could be adjust the measurement before, with another method, and then treat the problem as a Markov model.

Acknowledgments

I would like to thank my partner, parents and friends for their support. Also thanks to, Joan Camuñas and specially Fèlix Ritort for their collaboration.

[1] Baldazzi, V., Cocco,S., Marinari,E.,&Monasson,R. (2006). *Inference of DNA sequences from mechanical unzipping: An ideal-case study. Physical Review letters*, 96(12).

[2] Lawrence R. Rabiner (February 1989). *A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE.* doi:10.1109/5.18626

[3] Eddy, S.R. (2004). *What is a hidden Markov model? Nature Biotechnology, 22(10)*, 1315-1316 doi:10.1038/nbt1004-1315

[4] Mllner,F.E., Syed, S., Selvin, P.R., &Sigworth, F.J. (2010). *Improved hidden Markov models for molecular motors, part 1: Basic theory. Biophysical Journal, 99(11)*, 3684-3695.

[5] Storm. C., Nelson, P.C. (2003). *Theory of high-force DNA stretching and overstretching Physical Review*, E 67.

[6] Ribezzi-Crivellari,M,. Wagner,M., &Ritort, F. (2011). *Bayesian approach to the determination of the kinetic parameters of DNA hairpins under tension. Journal of Nonlinear mathematical Physics.*

[7] Durbin, R. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge: Cambridge University Press.

[8] Brmaud, P.(1999). *Markov Chains: Gibbs fields and Monte Carlo simulation, and queues.* New York: Springer.

[9] Ghahramani, Z. (2001). *An introduction to hidden Markov models and Bayesian networks. International Journal of Pattern Recognition and Artificial Intelligence.* doi:10.1142/S0218001401000836

[10] Huguet, J.M., Forns,N., & Ritort,F.(2009). *Statistical properties of metastable intermediates in DNA unzipping. Physical Review letters, 103(24).*

[11] Shen, Dawei. (2008). *Some Mathematics for HMM.*