Free Energy Landscape Analysis of Mesoscopic Model for Finding DNA-Protein Binding Sites

Rafael Tapia-Rojo, Juan José Mazo, Andrés González, M. Luisa Peleato, Maria F. Fillat, and Fernando Falo

1 Introduction

The physical modelization of biomolecules requires a careful choice of the scale of work depending on the problem we wish to study. All-atom simulations allow an accurate description of the system but within small time scales and severe size limitations subject to the available computational power. Coarse-grained models gather groups of atoms in point particles simplifying greatly the system but keeping the essence of the important interactions in the problem of study. At this level, the free energy landscape (FEL) of the system appears as a powerful tool to extract relevant information from a system with a high number of degrees of freedom.

We propose here a coarse-grained model for DNA-protein interaction problem. Transcription from DNA to RNA appears as a complex problem that requires regulation via DNA-protein interaction. Our model considers a particle as a generic protein that diffuses along the DNA chain with an interaction term that is coupled to local openings (bubbles). By applying an algorithm we are able to extract the FEL of the system and identify possible binding sites as those states with lower free energy. We focus mainly on the so called Transcription Starting Site (TSS), where the RNA-polymerase binds before transcription starts.

R. Tapia-Rojo (✉) • J.J. Mazo • A. González • M.L. Peleato • M.F. Fillat • F. Falo
Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza, Spain
e-mail: rafa.t.rojo@gmail.com; juanjo@unizar.es; andresglez2005@gmail.com; mpeleato@unizar.es; fillat@unizar.es; fff@unizar.es

Á. Corral et al. (eds.), *Extended Abstracts Spring 2013*, Trends in Mathematics 2,
DOI 10.1007/978-3-319-08138-0_15

2 The Model

We base our model on a modified Peyrard–Bishop–Dauxois (PBD) model [1, 4]. PBD model reduces the complexity of DNA to a set of N point particles that represent the base pairs of the chain. The only degrees of freedom are the coordinates $\{y_n\}$ which stand for the distance between each base pair. The total Hamiltonian of the model accounts for two phenomenological interactions, the intra-base and inter-base potentials,

$$\mathcal{H} = \sum_{n=1}^{N} \left[\frac{p_n^2}{2m} + V(y_n) + W(y_n, y_{n-1}) \right],$$

where $p_n = m\, dy_n/dt$ is the linear momentum of the n-th base pair and m is its reduced mass. The potential $W(y_n, y_{n-1})$ describes the inter-base pair or *stacking* interactions and its model by an anharmonic interaction. The intra-base pair potential $V(y_n)$ takes the form of a Morse potential (usual in chemical bonds) with an entropic barrier to account for solvent interaction.

Inspired in the one-dimensional diffusion of DNA-binding proteins, we include now a new degree of freedom to the traditional PBD model, a Brownian particle that slides along the DNA chain [5]. This particle interacts with the DNA through the phenomenological potential

$$V_{\text{int}}(X_p, \{y_n\}) = -\frac{B}{\sqrt{\pi\sigma^2}} \sum_n \tanh(\gamma y_n) e^{-(X_p - na)^2/\sigma^2},$$

which depends on the particle position X_p and the DNA instantaneous configuration $\{y_n\}_{n=1}^{N}$. This potential is just a sum of gaussian wells, each centered at the n-th base (na) and whose amplitude depends on the opening of the base pair. The hyperbolic tangent term just saturates the interaction strength. In this sense, the particle interacts more intensely with open regions of the sequence. In addition, the base pairs are also affected by the particle, so that they are more likely to be opened if the particle is within its range of interaction.

3 Methods

Langevin dynamics simulations: The model is simulated by integrating numerically $N + 1$ Langevin equations (N base pair plus the particle) using an stochastic Runge–Kutta algorithm. Each of the DNA sequences we study is simulated in five different realizations each one covering 40 μs, with a preheating time of 1 μs, reasonable times from a biological perspective. The simulation temperature is $T = 290$ K. We use periodic boundary conditions for the diffusing particle and

fixed boundary conditions for the sequence, adding 10 *CG* base pair clamps of at the end of each sequence to provide "hard-boundaries" and avoid end effects.

Analysis: We aim to find the most important conformational states in the dynamics in order to obtain a biological interpretation. To do so we apply an algorithm that allows us to obtain the Free Energy Landscape (FEL) of the system [3] so that the most relevant states can be identified and quantified. We start applying Principal Component Analysis to the trajectories to reduce greatly the dimensionality of the system but keeping most of the information. Next we translate the five reduced trajectories and the particle trajectory into a Conformational Markov Network (CMN). This coarse-grained picture is constructed by discretizing the conformational space explored by the system in order to define the nodes (microstates) of the network. Next, the links between the nodes are set according to the jumps between the microstates in the trajectories. In this sense, nodes are weighted (P_i), and the links directional and weighted (P_{ij}). In order to define the conformational states of the system, we split the CMN into its basins of attraction, i.e., regions in which the probability fluxes (P_{ij}) converge to a common state (attractor) of the network. To do so we apply the stochastic steepest descent algorithm, developed in [3]. Each basin corresponds to a coarse-grained macrostate of the system. From the basin network we can build FEL of the system represented as a hierarchical tree diagram (dendrogram), by assigning to each node a free energy according to its weight $F_i/kT = \log P_W - \log P_i$, where P_W is the weight of the weightiest node. This magnitude is used as a control parameter, increasing it step by step from the weightiest node, so that new nodes arise, together with their links. Most relevant states can now be identified as those more populated and, in addition, the topology of the network informs us about possible transitions between them.

Study of cyanobacterial genome: In this work, we focus on a concrete genome, analyzing promoters from Anabaena PCC 7120 [2]. Cyanobacteria constitute an interesting model as they show differentiated cells (heterocysts) that need several transcriptional changes to develop. In addition, several well characterized promoters are available so that our computational results can be clearly compared with experimental works.

4 Results

We show now provisional results of some cyanobacteria promoters we have analyzed so far. *TSS finding and base-pair opening*: We analyze promoter sequences comprising between 100 and 200 base-pairs. Figure 1 shows the base-pair opening profile for each promoter sequence with the TSS site highlighted. We find in any case a peak located around the TSS site. This means that, on average, this site is likely to be open, this is, bubbles form with high probability around the TSS.

Promoters with different characteristics have been chosen for the analysis. As Fig. 1 shows, some of the analyzed sequences contain a single TSS while others more than one. Even though in every case all the TSSs can be identified with

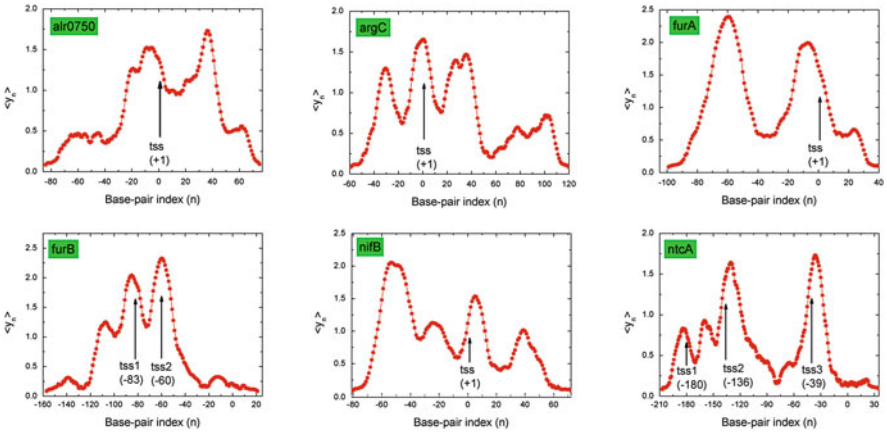

Fig. 1 Mean base-pair position for each of the six analyzed promoters. We can find a peak around the tss in every case meaning that bubbles are likely to be formed around this site

this representation, it is interesting to extract further information from the model. Specifically, different TSSs may show different activity levels depending on the amount of RNA they produce. We refer to this as the strength of each promoter. The relative height of the peaks in Fig. 1 may be a qualitative indicative of this behavior, but a quantitative study motivates the use of the algorithm described in Sect. 3. Another remarkable point in Fig. 1 is the appearance of multiple peaks apart from the ones marked as TSS, revealing the possibility of additional binding sites or just the existence of "false positives". Further discussion can be also done in this point, as the relative strength of the TSS compared to other possible binding sites found in our model can be studied and related with the known biological behavior.

FEL analysis of highlighted promoter: We choose the most interesting promoter we have analyzed so far in order to illustrate the analysis method exposed above. The interest of ntcA promoters lies in the fact that it displays more than one TSS and so it is subject to further biological interpretation. Figure 2 shows the free energy dendrogram for the ntcA promoter, with the set of basins and their accumulated weight corresponding to its TSS sites highlighted. In addition we draw the typical macrostate of the branch.

Promoter ntcA contains three different TSSs. The interpretation of the free energy dendrogram matches with the biological studies of ntcA. TSS3 and TSS1 are just transcribed when there is no nitrogen in the environment. TSS3 is a stronger site, as transcription last for a longer time and the site is still active when the heterocyst is mature. On the other hand, TSS1 is just active transitorily. TSS2 is a constitutive site, so it is always transcribed. This may explain the relative importance found in our analysis.

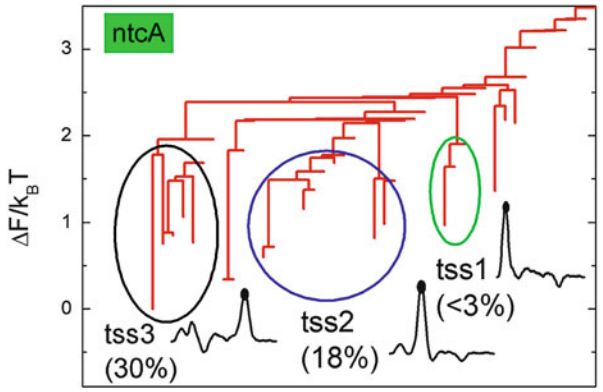

Fig. 2 Free energy dendrogram for ntcA. The TSSs appear as populated branches with relative occupations that inform about the importance of each site

5 Conclusions and Further Work

In this work we use a previously developed mesoscopic model for DNA-protein interaction to analyze the genome of a cyanobacteria organism. The goal is to identify and characterize the TSSs of each of the analyzed promoters. This site is observed to show a larger mean opening due to the formation of bubbles. This behavior is greatly influenced by the dynamics of the generic protein. In order to analyze quantitatively the sequences we apply an algorithm that translates the trajectories onto a complex network so that the free energy dendrogram can be obtained. From this analysis, most relevant sites in the dynamics can be found and related to the biological behavior of each promoter.

This work is to be continued in two main directions. First, further biological interpretation can should be done from the basin networks and the free energy dendrogram. Each of the chosen genes show different regulation features that may be related to the physical parameters found in our analysis. Additionally, more promoters are to be simulated and analyzed in order to complete the work and validate our model and method.

References

1. T. Dauxois, M. Peyrard, A.R. Bishop, Phys. Rev. E **47**, 684 (1993)
2. J. Mitschke, A. Vioque, F. Haas, W.R. Hess, A.M. Muro-Pastor, PNAS **108**, 20130–20135 (2011)
3. D. Prada-Gracia, J. Gómez-Gardenes, P. Echenique, F. Falo, PLoS Comput. Biol. **5**, e1000415 (2009)
4. R. Tapia-Rojo, J.J. Mazo, F. Falo, Phys. Rev. E **82**, 031916 (2010)
5. R. Tapia-Rojo, D. Prada-Gracia, J.J. Mazo, F. Falo, Phys. Rev. E **86**, 021908 (2012)