CS-433: Project-2 report
Sequence-dependent clustering of DNA in Protein-DNA Xray crystal data and its comparison with clustering in cgDNA+ model

Anas Ibrahim[1], Vincent Parodi[1], and Rahul Sharma[1]

[1]Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland
{*anas.ibrahim,vincent.parodi,rahul.sharma*}*@epfl.ch*

***Abstract*—In this report, we performed hierarchical-clustering analysis on two data sets of DNA dimers in tetramer contexts obtained from cgDNA+ model and Xray protein-DNA crystal database. After feature reduction using a) removal of rigid modes and b) principal component analysis, we found four clusters in both the data sets corresponding to four chemically distinct base-pair steps (the combinations of purine and pyrimidine bases). These clusters also reflect experimental known features such as exceptional flexibility of pyrimidine-purine steps. Furthermore, in each of the clusters, there are less clearly defined four sub-clusters which corresponds to specific sixteen dimer base-pair steps. Lastly, PCA similarity factor between the direction of deformations in two data sets was found to be ≈ 0.98.**

I. Introduction

Deoxyribonucleic Acid, DNA is a long flexible molecule which contains genetic information required for development, function, and reproduction of living organisms. DNA is made of a repetitive unit called nucleotide containing four different types of bases (**A**denine, **G**uanine, **T**hymine, and **C**ytosine) which allows different chemical and mechanical properties in DNA fragments with different combinations of these bases. Furthermore, A/G and T/C are purine (R) and pyrimidine (Y) class of bases, respectively and are chemically more similar to each other. In this work, we solely focused on the sequence-dependent mechanical properties of DNA which play a pivotal role in protein recognition[1] and indirect readout[2].

The most challenging aspect in a systematic investigation of sequence effect in DNA properties is its vast sequence space. For example, a dodecamer (a DNA of length 12 base-pairs) has 4^{12} possible combinations and this number explodes astronomically for larger sequence. More importantly, the properties of a given fragment of DNA also depends of its flanking sequence along with the sequence of the fragment. For instance, $X_1X_2TAX_3X_4$, the shape and stiffness of TA also depends on the base-type of $X_1, X_2, X_3,$ and X_4 i.e. non-local sequence-dependence in DNA. However, the properties of a particular DNA fragment is dominated by its sequence and non-local contributions diminish exponentially for sequence away from the fragment which allows studying the behaviour of DNA at mesoscopic lengths.

Still, the vast complexity in sequence space can't be explored using traditional techniques like experiments and **M**olecular **D**ynamics simulations. LCVMM group developed a coarse-grained model of DNA trained on atomistic MD (using machine learning algorithms), cgDNA+[3] which predicts the

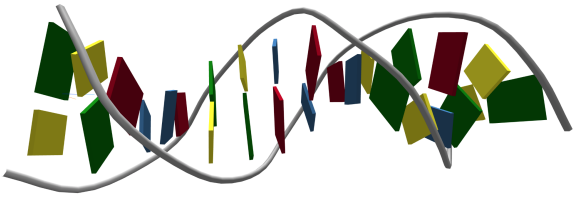

Fig. 1. A coarse-grained cartoon representation of DNA in cgDNA+ model for a sequence of length 12. Each frame represents a rigid-base and different bases are plotted using different colours. Image created using cgDNAWeb [4]

average shape and stiffness for a given sequence of DNA very efficiently and allows to make statistical conclusions in the vast sequence space of DNA. This model accurately captures the behaviour of DNA observed in MD simulations (outside the training library). The model uses more than 20,000 parameters and was trained on more than ten millions of MD snapshots. Such a model can't be trained from limited experimental data (which only contains a few thousand snapshots).

A natural question arises whether the sequence-dependent mechanical behaviour observed in the model is also present in experiments. A direct one-to-one comparison of experiments with the model is rather complicated (due to multi-variate Gaussian data-points) and therefore, we have just addressed whether we observe similar mechanical behaviour in two data sets through clustering. Due to the scarcity of experimental data, we did our analysis at tetramer level (i.e. DNA sequence of length 4) comparing the average shape and covariance of the middle dimers in all possible 256 tetramer contexts which allowed to explore the non-local effect of this tetramer context. In this project, we have addressed the following questions: a) given a data set (either model or experiment), is there any sequence-dependence in DNA mechanics at all? Also, are there any patterns or clusters in the data set? b) if yes, are these clusters physically sensible and consistent in two data sets?

II. Methods

A. Configuration of DNA

We have compared the average shape and covariance of DNA in a coarse-grained configuration as shown in fig. 1 which is parameterised using relative internal coordinates[5]. Six degrees of freedom (three rotations and three translations) are defined between each base-pair (for example, between red and blue frames) called intra-base pair coordinates and

then for consecutive base-pairs further six degrees of freedom are defined called inter-base pairs coordinates. This way to completely describe the configuration of a DNA of length N base-pair, $12N-6$ internal coordinates are required.

B. Databases

In this work, we have compared two data sets, M (obtained from cgDNA+ model) and E (obtained from experimental database). In each of the data sets, there are 256 Gaussian pdfs correspond to dimers in 256 possible tetramer contexts (to study non-local sequence effect) and each Gaussian pdf is defined by a mean (18×1) and a covariance matrix or inverse stiffness matrix (18×18) which corresponds to the average shape and covariance of the middle dimer junction in the tetramer.

The experimental data is obtained from protein-DNA crystal database available at PDB Database server[6]. Note that in the experimental data, DNA is present in bound form with protein while the cgDNA+ model is trained on MD simulations of naked DNA in solution. Therefore, this comparison between two data sets holds valid under the assumption that distortion of DNA due to protein binding is in different random directions which effectively cancels out for an ensemble of protein-DNA complexes.

C. Data pre-processing

The cgDNA+ model data for every dimer the in tetramer context is a well-defined Gaussian pdf. In contrast, the experimental data obtained from the server contains some outliers and therefore, if any of the internal coordinates is beyond 3 standard deviations then that particular data point was rejected. Moreover, we have only chosen those data points which have the two strands of DNA in bound form. Lastly, in case of redundant entries of DNA-protein complex, we have taken the one with the highest Xray technique resolution.

D. Definitions and protocols

1) Divergence between Gaussian pdfs: To quantify divergence between two Gaussian distributions (ρ_1 and ρ_2) with mean vectors (μ_1 and μ_2) and inverse covariance matrices (K_1 and K_2), we have used Kullback-Leibler divergence (D_{KL}) and Mahalanobis distance (M_D),

$$\begin{aligned} D_{KL}(\rho_1,\rho_2) &= \frac{1}{2}\left[K_1^{-1} : K_2 - \ln\left(\frac{|K_2|}{|K_1|}\right) - I : I\right] + M_D^2 \\ M_D^2(\rho_1,\rho_2) &= \frac{1}{2}(\mu_1-\mu_2)^T K_2 (\mu_1-\mu_2), \end{aligned} \tag{1}$$

where colon is Euclidean inner product for square matrices. We have also used other metrics such as Hellinger distance and corresponding results are provided with code but not included in the report due to limited space.

2) Clustering algorithms: For clustering, we have used K-means and hierarchical clustering algorithms and employed various linkage methods to define the distance between two clusters in hierarchical clustering. In the report, we have presented our results using average linkage algorithm which defines the distance between two clusters as the average of distances between all pairs of objects in two clusters. Rest of the results are provided with the code.

3) Dimension reduction: In the provided data sets, we have 18 features, however, not all are equally sensitive to the sequence of DNA or the variation in those features is so less that it is not possible to resolve the sequence effect from underlying noise. Therefore, we decided to remove such features in two ways: a) remove rigid features: in the average covariance matrix, the diagonal entries indicate rigidity of features and highly rigid features are likely to be insensitive to sequence (which we have verified). So, we removed such features from the data sets. The key advantage of such feature reduction is the remaining features are in well-accepted canonical direction of DNA movement, and b) principal component analysis: alternate method we employed to reduce the dimension is to project the data in the direction of principal components of the covariance matrix. These principal components represent the direction in which DNA deforms the most. However, the direction of principal components in two data sets are different and therefore, we have defined PCA similarity factor [7], S which quantifies similarity between two covariance matrices or direction of DNA deformations[8] and is defined as

$$S = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k}(v_i^E \cdot v_j^M)^2 \tag{2}$$

where v_i^E and v_j^M are principal components of E and M and k is the number of principal components chosen. k is chosen heuristically based on the variance explained by k principal components.

III. Results and Discussions

A. K-means cluster

In the first step, we applied K-means clustering to find 4 (RR,RY,YR,YY) or 16 (total possible dimer steps - AA, AT, AG, AC, TA, TT, TG, TC, CA, CT, CG, CC, GA, GT, GC, GG) physically anticipated clusters. As input for k-means clusters, we transformed our Gaussian pdf input to a vector form by vectorizing covariance and combine it with mean. The clusters, thus, obtained contains mix dimers and physically less expected. Furthermore, the results didn't improve even after feature reduction. In such high dimensional space, any sort of feature reduction, make our analysis more and more abstract and difficult to understand the underlying sequence effect. Therefore, we decided to apply hierarchical clustering in which various parameters can be chosen more intuitively.

B. Hierarchical clustering in original data

In fig. 2, we have plotted dendrogram for the model data using hierarchical average linkage algorithm in which there

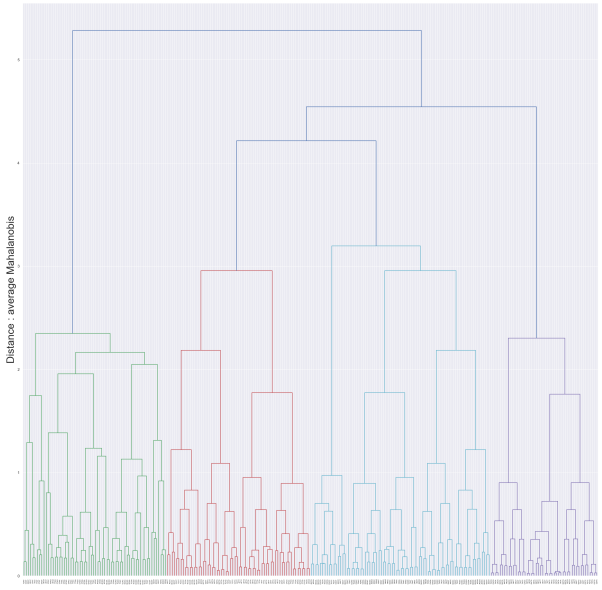

Fig. 2. Dendrogram for hierarchical clustering of cgDNA+ model data.

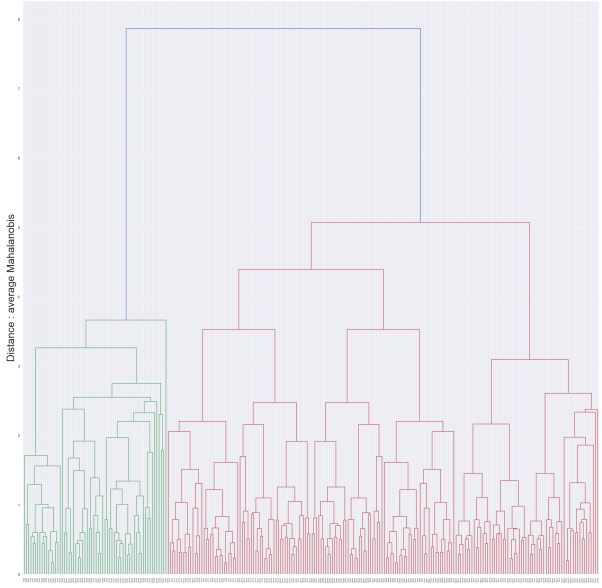

Fig. 3. Dendrogram for hierarchical clustering of experimental data after pre-processing.

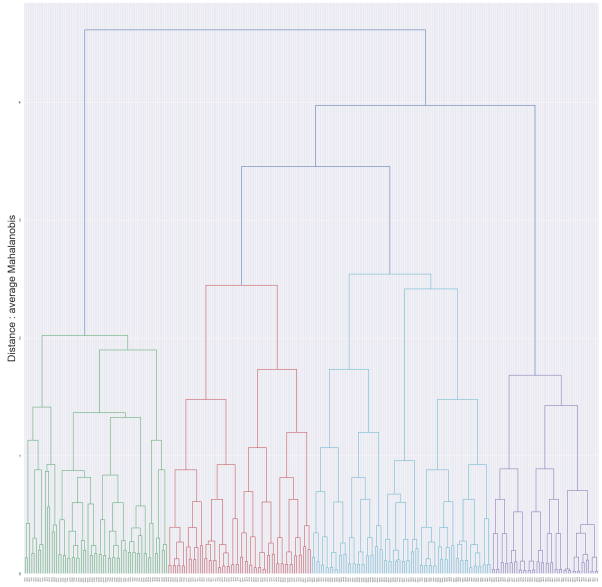

Fig. 4. Dendrogram for hierarchical clustering of model data after removing rigid-features.

are four clusters corresponding to YR, RR, YY, and RY as the middle dimer steps. This is physically sensible as chemically similar dimer-steps are clustered together. Interestingly, YR cluster is farthest from the rest of the clusters which might be explained by the fact that YR is found to be exceptionally flexible[9]. Furthermore, in each cluster, there are sub-clusters (close to each other) which corresponds to individual dimer-steps. Thus, in the model data, the result we obtained is physically anticipated. In contrast, in the equivalent dendrogram for experimental data, there is no clustering at all. The possible explanations for such observations in experiments are low statistic of experimental data and underlying noise in experimental techniques. One way to get rid of noise in the data is to remove the features which have the least variance.

C. Hierarchical clustering after removing rigid modes

As discussed in section II-D3, we removed five rigid features (smallest diagonal entries in covariance matrix) which also corresponds to the features which has the least variance in the average shape of both E and M. In M, we observed similar clustering with better sub-clusters (as shown in fig. 4). In E, there is a significant improvement in clustering (as shown in fig. 5). YR cluster is well-separated from rest of the clusters while RR, YY, and RY clusters are close to each other. Furthermore, sub-clusters in these clusters are poorly formed. One of the drawbacks of removing such rigid features is that these rigid features are often correlated with soft features.

D. Hierarchical clustering of projected data on principal components

In order to obtain the uncorrelated soft features, we have performed principal component analysis on the average covariance and cluster the projected data on these principal components (k). We found choosing smaller k (which explains less variance in the data) is insufficient to find physical clusters in both the data sets. In contrast, choosing a large value of k ($\geq$ 13) starts amplifying the noise present in the data (particularly in E) and leads to poorly formed clusters. In fig. 6 and fig. 7, we have shown the clustering of the projected data on the principal components (k=13) for M and E, respectively. The clusters in the two data sets are consistent and physically sensible. Lastly, the question remains how different are the direction of principal components or the direction of DNA deformations in the two data sets? We have quantified this using PCA similarity factor (defined in eqn. 2) which turns out to be 0.98 for $k = 12$. It confirms that in the two data sets, the deformations of DNA are in almost similar directions.

IV. Summary

In this report, we have investigated the clustering of DNA dimers in tetramer contexts using hierarchical clustering. The different clusters represent different sequence-dependent mechanical behaviour of DNA. In the model data, we found physical clusters corresponding to four chemically distinct base step-type. Moreover, YR step is known to be very flexible and the same was observed in dendrograms. In each of the clusters,

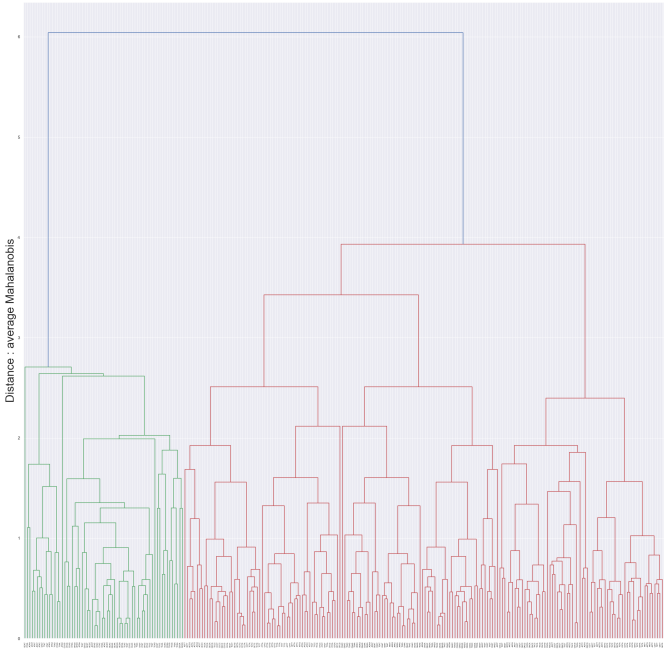

Fig. 5. Dendrogram for hierarchical clustering of experimental data after removing rigid-features.

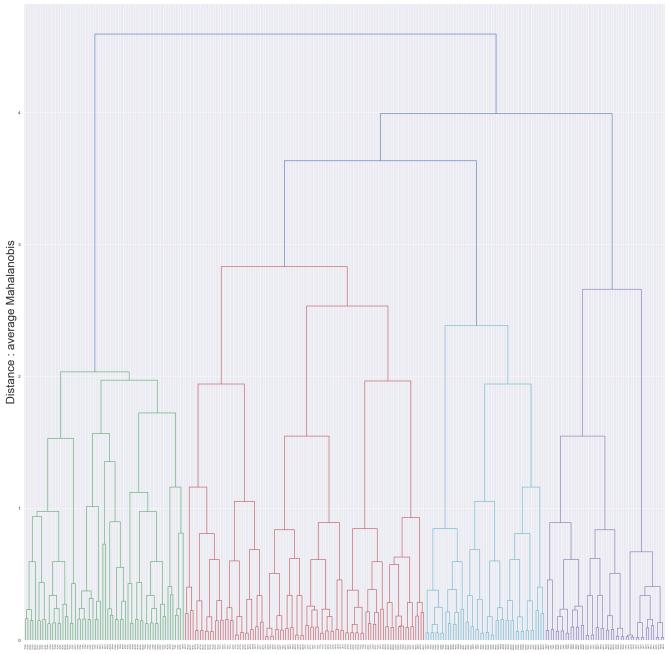

Fig. 6. Dendrogram for hierarchical clustering of transformed model data in the direction of principal components of average covariance (k=13).

we further found four sub-clusters (not perfectly formed) corresponding to four types of base-steps. However, in the experimental data, there were no such clusters probably due to the dominant noise in some of the features (in particular, the ones with the least variation in the sequence space. Therefore, we decided to remove such rigid features and found the same four clusters in the experimental data but the sub-clusters were less clear. We also performed PCA on the average covariance matrix and performed our clustering analysis in the direction of principal components. We found cleaner clusters with better sub-clustering. Lastly, we found that PCA similarity factor between the direction of deformations of DNA in two data sets ≈ 0.98. In conclusion, we have showed that the behaviour of DNA in the two data sets are very close and cgDNA+ model can be routinely used to make statistical conclusions in vast sequence space of DNA.

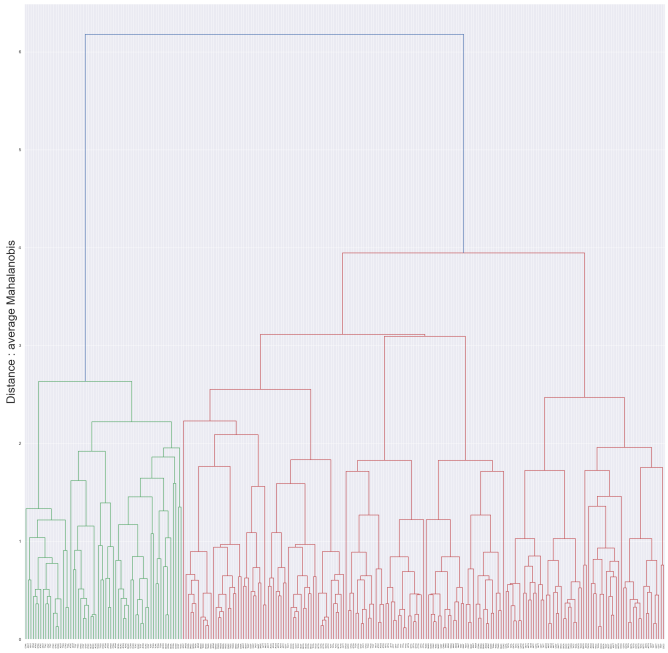

Fig. 7. Dendrogram for hierarchical clustering of transformed experimental data in the direction of principal components of average covariance (k=13).

REFERENCES

[1] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein–DNA recognition," *Nature*, vol. 461, no. 7268, pp. 1248–1253, 2009.

[2] A. A. Napoli, C. L. Lawson, R. H. Ebright, and H. M. Berman, "Indirect readout of DNA sequence at the primary-kink site in the CAP–DNA complex: Recognition of pyrimidine-purine and purine-purine steps," *Journal of molecular biology*, vol. 357, no. 1, pp. 173–183, 2006.

[3] A. Patelli, "A sequence-dependent coarse-grain model of B-DNA with explicit description of bases and phosphate groups parametrised from large scale Molecular Dynamics simulations," Ph.D. dissertation, 2019.

[4] L. De Bruin and J. H. Maddocks, "cgdnaweb: a web interface to the cgdna sequence-dependent coarse-grain model of double-stranded dna," *Nucleic acids research*, vol. 46, no. W1, pp. W5–W10, 2018.

[5] R. Lavery, M. J. H. P. D. Moakher, M., and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: Curves+," *Nucleic acids research*, vol. 37, no. 17, pp. 5917–5929, 2009.

[6] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain *et al.*, "The protein data bank," *Acta Crystallographica Section D: Biological Crystallography*, vol. 58, no. 6, pp. 899–907, 2002.

[7] K. Yang and C. Shahabi, "A pca-based similarity measure for multivariate time series," in *Proceedings of the 2nd ACM international workshop on Multimedia databases*, 2004, pp. 65–74.

[8] A. Perez, A. Noy, F. Lankas, F. J. Luque, and M. Orozco, "The relative flexibility of B-DNA and A-RNA duplexes: database analysis," *Nucleic acids research*, vol. 32, no. 20, pp. 6144–6151, 2004.

[9] G. A. A. Olson, W. K., X. J. Lu, L. M. Hock, and V. B. Zhurkin, "DNA sequence-dependent deformability deduced from protein–DNA crystal complexes," *Proceedings of the National Academy of Sciences*, vol. 95, no. 19, pp. 11 163–11 168, 1998.