Kernel Machines Made of DNA Molecules

Yung-Kyun Noh[⋆‡], Daniel D. Lee[⋆], Cheongtag Kim[†], and Byoung-Tak Zhang[‡]
[⋆]GRASP Lab., Department of Electrical and Systems Engineering,
University of Pennsylvania, Philadelphia, PA 19104
[†]Quant Psych Lab. Department of Psychology,
Seoul National University, Kwanak-gu Seoul, 151-746, Korea
[‡]Biointelligence Lab. School of Computer Sci. and Eng.
Seoul National University, Kwanak-gu Seoul, 151-746, Korea
{nohyung, ddlee}@seas.upenn.edu, ctkim@snu.ac.kr, btzhang@bi.snu.ac.kr

Topic: learning in biological systems, Preference: Oral/Poster
Presenting author: Yung-Kyun Noh

Recent advances in biotechnology have leveraged laboratory techniques to generate synthetic biological representations of information. Here, we show how these algorithms on real biological materials can be used to build a learning system that discriminates sequence patterns. We show how a kernel classification algorithm can be implemented *in vitro* in order to discriminate a new DNA sequence based upon pre-labeled training sequences. In particular, similarities between DNA sequences can be interpreted as elements of a positive definite kernel matrix.

Definition (DNA kernel): When N different species of single stranded DNA ($ssDNA(i)$ for $i \in \{1, ..., N\}$) and complementary single stranded DNA ($cssDNA(i)$ for $i \in \{1, ..., N\}$) have the same number of molecules, their hybridizations generate a DNA kernel K matrix whose $K_{ij} \equiv |dsDNA(i,j)|$ element is defined by the amount of double stranded DNAs ($dsDNA(i,j)$), consisting of hybridized $ssDNA(i)$ and $cssDNA(j)$. ■

We investigate when this kernel matrix is positive definite and can be interpreted as a geometrical mapping into feature space. To do this, we analyze the hybridization process to see how the amounts of double strands are related. Recent thermodynamical models of DNA hybridization use binding energies, lengths of binding sites and temperature to predict hybridization. We show under certain conditions the hybridizations process can generate a positive definite kernel, and then, the well-established kernel concepts provide natural geometrical interpretation on operations out of similarity information.

Like other kernel methods, we represent everything as the linear expansion of data in the feature space, and the spanning parameters are the parameters to be learned. In our setting, they are represented by the amount of single stranded DNA sequences. Discrimination will be done by counting to which labeled strands testing strands attached more. This can be described by the following equation.

$$y_{new} = sign\left(\sum_i \alpha_i y_i K(new, i)\right) \tag{1}$$

where α_i is the population of i'th sequence. Our only assumption is the amount of double strands is bilinear to the amount of each single, which will not be violated in transient state until double strands are not saturated.

Learning implementation to optimize parameters is iterative hybridization and selection. Selection is a process of selecting hetero-labelled double strands. In other words, we select and keep double strands whose constituents' labels are different, and throw away others. With

some scaling and other simple selecting processes, we can show this iterative process eventually make the population sets $\alpha_{\{-1\}}$ of class -1 sequences and $\alpha_{\{1\}}$ of class 1 sequences approach to the solution of the following optimizing criterion.

$$\begin{aligned} &\arg \max_{\alpha_{\{\mathbf{i}\}} \in \Re^{N_i}, i=-1,1} \quad \mathbf{w_{\{-1\}}}^T \mathbf{w_{\{1\}}} \\ &s.t. \quad \mathbf{w_{\{-1\}}} = \Phi_{\{-1\}}^T \alpha_{\{-1\}}, \quad \mathbf{w_{\{1\}}} = \Phi_{\{1\}}^T \alpha_{\{1\}}, \quad \text{for} \quad \alpha_{\{-1\}}, \alpha_{\{1\}} \geq 0, \\ &\quad \mathbf{w_{\{-1\}}}^T \mathbf{w_{\{-1\}}} = 1, \quad \text{and} \quad \mathbf{w_{\{1\}}}^T \mathbf{w_{\{1\}}} = 1 \end{aligned} \tag{2}$$

where, $\Phi_{\{\mathbf{i}\}}$ is the data matrix in feature space and N_i is the number of data in class $i \in \{-1, 1\}$. As a result, it turns out that the learning process is finding two most closest vectors one of which is in the convex cone of each class in the feature space. They become the representatives of two classes and we can also interpret discrimination as measuring to which class's representative test datum is closer by angle.

Different scheduling of hybridizing temperature yields different distribution of double strands and control the sparsity of kernel matrix which works as a kernel parameter. It's application to several sets of biological data showed the performance is similar to the best performance of SVMs.

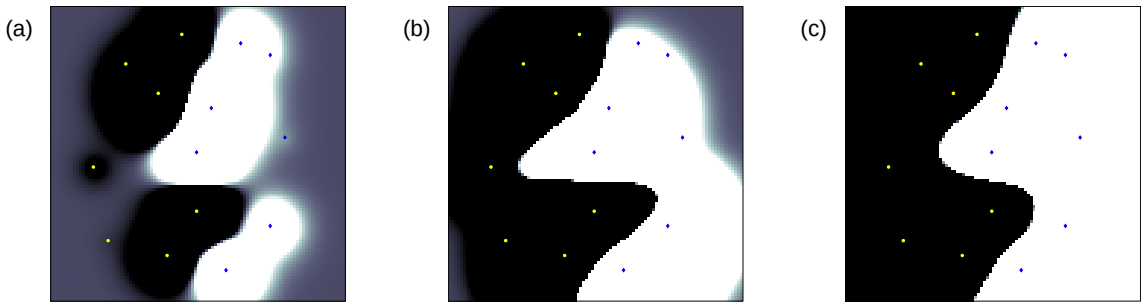

Figure 1: **[Discrimination of 2-D space]** Discrimination of 2-D space when binding energy is set anti-proportional to the distance between two points. The figures show the sparseness of kernel matrix is controlled by temperature schedule. Temperature control for kernel formation is (a) 80^oC constant, (b) 80^oC to 50^oC, and (c) 80^oC to 20^oC. The range of binding energies between training samples are scaled to have between $-5.9 \sim -8.0$ (kcal/MBP), which are within the range of real binding energy.

This work is supported in part by the Molecular Evolutionary Computing (MEC) Project by MICE and the Pioneer Program of KOSEF.

References

J.Kim, J.Hopfield, and E.Winfree. Neural network computation by in vitro transcriptional circuits. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 681–688. MIT Press, Cambridge, MA, 2005.

J.S. Kim, J.W. Lee, Y.K. Noh, J.Y. Park, D.Y. Lee, K.A. Yang, Y.G. Chai, J.C. Kim, and B.T. Zhang. An evolutionary monte carlo algorithm for predicting DNA hybridization. *BioSystems*, 91(1):69–75, 2008.

A Zien, G Rätsch, S Mika, B Schölkopf, T Lengauer, and K.R. Müller. Engineering support vector machine kernels that recognize translation initiation sites. *BIOINF: Bioinformatics*, 16, 2000.