**Predicting variation of DNA shape preferences in protein-DNA interaction in cancer cells with a new biophysical model**

**Kirill Batmanov1 and Junbai Wang1,\***

1. Department of Pathology, Oslo University Hospital – Norwegian Radium Hospital, Montebello 0310, Oslo, Norway

\*. Correspondence to [junbai.wang@rr-research.no]

**Supplementary Information**

**Supplementary Methods**

**Hyperparameter update formulae**

Here we present more technical details about the hyperparameter update step in the model fitting.

When computing the Hessian matrix $H$during the hyperparameter update of the shape model, we use the fact that the shape model data cost function can be represented as the full dinucleotide cost function $L\_{D}^{dinuc}$ with dinucleotide parameters linearly transformed:

$$\begin{array}{c}L\_{D}\left(w, d^{f},μ,a,b\right)=L\_{D}^{dinuc}\left(w,d,μ,a,b\right) (1)\end{array}$$

$$d=d^{f}∙D$$

$L\_{D}^{dinuc}$ was called $E(S)$ in (1), and its Hessian matrix $H^{dinuc}\left(w,d,μ,a,b\right)$ can be computed in a straightforward way using the R-propagation algorithm (2). Then $H$can be computed by transforming appropriate parts of $H^{dinuc}$in the following manner:

$$\frac{∂^{2}L\_{D}}{∂x ∂d\_{p,k}^{f}}=\sum\_{i=1}^{16}D\_{k,i}\frac{∂^{2}L\_{D}^{dinuc}}{∂x ∂d\_{p,i}}, for x\in \left\{w\_{j,a},μ,a,b\right\}$$

$$\frac{∂^{2}L\_{D}}{∂d\_{p\_{1},k\_{1}}^{f} ∂d\_{p\_{2},k\_{2}}^{f}}=\sum\_{i=1}^{16}\sum\_{j=1}^{16}D\_{k\_{1},i}D\_{k\_{2},j}\frac{∂^{2}L\_{D}^{dinuc}}{∂d\_{p\_{1},i} ∂d\_{p\_{2},i}}$$

These expressions are derived from equation (1). Second derivatives of the full dinucleotide model ($\frac{∂^{2}L\_{D}^{dinuc}}{∂x ∂d\_{p,i}}$ and $\frac{∂^{2}L\_{D}^{dinuc}}{∂d\_{p\_{1},i} ∂d\_{p\_{2},i}}$) are computed by the R-propagation algorithm in the same way as in (1). When several model weights $W\_{i\_{1}},…,W\_{i\_{n}}$ share a single regularization hyperparameter $α\_{m}$, the update formula for this alpha is as follows:

$$α\_{m}=\frac{γ\_{m}}{\sum\_{k=1}^{n}W\_{i\_{k}}}$$

$$γ\_{m}=\sum\_{k=1}^{n}1-α\_{m}H\_{i\_{k}i\_{k}}^{-1}=n-α\_{m}\sum\_{k=1}^{n}H\_{i\_{k}i\_{k}}^{-1}$$

The update for $β$ is done according to the formulae in the main text.

**Supplementary Figures**

**Supplementary Figure S1. Fitted shape preferences of MAX and BHLHE40 for different cell lines (positive part).**

The heatmaps show the preference for each shape feature at each position (the inferred $d^{f}$ matrix). Only the positive preferences are shown, that is, the preferences towards lower values of shape features. Colours faded to grey mean weaker preference. The PWM logos are shown above the shape preferences heatmaps.

****

**Supplementary Figure S2. Shape model parameters in different conditions (negative part).**

The heatmaps show the preference for each shape feature at each position (the inferred $d^{f}$ matrix). Only the negative preferences are shown, that is, the preferences towards lower values of shape features. Colours faded to grey mean weaker preference. The PWM logos are shown above the shape preferences heatmaps.

****

**Supplementary Figure S3. Changes in shape preferences.**

**A.** Variation of inferred shape preferences of ERα at different times after treatment with E2 in MCF7 cell line. **B.** Variation of inferred shape preferences of FOXA1 after treatment with either Dex or E2 in MCF7 cell line. The heatmaps show the preference for each shape feature at each position (the inferred $d^{f}$ matrix). Only the negative preferences are shown in **A**, and only the positive preferences in **B**. Colours faded to grey mean weaker preference. The PWM logos are shown above the shape preferences heatmaps.

****

**Supplementary Figure S4. ASB prediction of the full dinucleotide and shape-restricted models compared to the independent model.**

For each of the 36 ASB datasets, AUPRC of the independent model is compared to **A)** the full dinucleotide model, and **B)** the DNA shape-restricted dinucleotide model. The P value of the two sided Wilcoxon signed rank test comparing the independent and dependent AUPRC (n = 36) is given.

****

**Supplementary references**

1. Wang, J. (2014) Quality versus accuracy: result of a reanalysis of protein-binding microarrays from the DREAM5 challenge by using BayesPI2 including dinucleotide interdependence. *BMC bioinformatics*, **15**, 289.

2. Wang, J. and Morigen. (2009) BayesPI - a new model to study protein-DNA interactions: a case study of condition-specific protein binding parameters for Yeast transcription factors. *BMC bioinformatics*, **10**, 345.