Extracting Sequence Features to Predict Protein-DNA Interactions: A Comparative Study

(Supplementary Text)

Qing Zhou Jun S. Liu

Compiled motif matrices

The Sox-Oct composite motif was identified from both the Oct4 and the Sox2 positive ChIP regions by the software CisModule [1] with a heterogeneous Markov background [2]. Noting that this motif is identical to the Sox-Oct composite motif detected from an independent Oct4 ChIP-PET data set in mouse ESCs [3], we included it in our pre-compiled motif set. In addition, we included all the 219 high-quality PWMs from TRANSFAC release 9.0 [4] and the PWMs of four TFs with known functions in ES cells from ref. [5]-[8].

Ten-fold CVs on the Oct4 ChIP-chip data

To eliminate co-linearity among the background word frequencies in LR-Full, we removed the last category of the k-mers from the input feature vector for $k = 2$ and 3, respectively. In Step-SO, we started from the LR-SO model and used the stepwise method (with both forward and backward steps) to add or delete features in the linear regression model based on the AIC criterion (see R function "step"). The Step-Full was performed similarly, but starting from the LR-Full model.

For neural networks (implemented in R package "nnet" by Venables and Ripley), we tested its performance with all combinations of different number of hidden nodes (2, 5, 10, 20, 30) and weight decay (0, 0.5, 1.0, 2.0). The CV-cor of NN-SO reached a quite stable level around 0.46 when the weight decay was > 0.5 for different number of hidden nodes. Its optimal CV-cor (=0.468) was achieved with weight decay = 1.0 and 10 hidden nodes. The neural network with all the features encountered a severe overfitting problem, resulting in a CV-cor < 0.38 for all tested combinations of weight decay and numbers of hidden nodes. In order to alleviate the overfitting problem for NNs, we reduced the input features to those selected by the stepwise regression (about 45), and employed a weight decay of 1.0 with 2, 5, 10, 20, or 30 hidden nodes. We call this approach Step+NN, and it reached an optimal CV-cor of 0.463 with 2 hidden nodes.

We applied MARS (R package "mda" by Hastie and Tibshirani) to this data set under two settings: the one with no interaction terms ($d = 1$) and the one with two-way interactions ($d = 2$). For each setting, we chose different values for the penalty λ, which specifies the cost per degree of freedom. In the first setting ($d = 1$), we set the penalty $\lambda = 1, 2, \cdots, 10$, and observed that the CV-cor reached its maximum of 0.580 when $\lambda = 6$. We note that the performance of MARS was quite sensitive to the choice of λ. With $\lambda \leq 2$, MARS greatly overfitted the training data, and the CV-cors dropped to below 0.459. MARS with two-way interactions ($d = 2$) showed unsatisfactory performance for $\lambda \leq 5$ with CV-cor < 0.360. We then tested λ in the range of $[10, 50]$ and identified the optimal CV-cor of 0.561 when $\lambda = 20$.

Support vector regression (ϵ-SVR) as defined in [9] was applied to this data set with the implementation of LIBSVM [10] in the R-package "e1071". We tested the linear, radial basis, 3rd-order polynomial and sigmoid kernels with the cost $C = 0.1, 1, 10, 100$. It turned out that for this data set, the radial basis kernel performed better than all the other kernels and the corresponding CV-cor ranged from 0.489 to 0.547 for different values of C. The optimal result was achieved when $C = 1$ with insignificant decrease in CV-cor for $C > 1$.

As discussed in the main text, the tuning parameter for boosting is the number of iterations, which equals the number of additive model components (regression trees here) included. We applied the R package "gbm" by Ridgeway to this data set, with the shrinkage parameter (learning rate) $\nu = 0.1$ and the default settings for all the other parameters. The number of trees M was set between 50 and 500, and the optimal model was obtained when stopped at 100 trees. It was observed that the CV-cors of boosting, between 0.541 and 0.581, were quite robust for different number of trees.

For BART, we ran 20,000 iterations after a burn-in period of 2,000 iterations, as implemented in the R package "BayesTree" [11]. We tested the method with the number of trees ranging from 20 to 200. Other parameters for prior distributions were specified by the default setting in the R package. Notably, BART with different number of trees reached CV-cors between 0.592 and 0.6, which outperformed the optimal results of all the other methods. We also noticed that the performance of BART was very robust for different choices of tree numbers.

Ten-fold CVs on the Sox2 ChIP-chip data

For LR-SO, LR-Full, Step-SO, Step-Full, SVM, boosting, and BART, the cross validations were performed exactly the same as we did on the Oct4 data set. For NN-SO, we tried all the combinations of the number of hidden nodes (2, 5, 10, 20, and 30) and weight decay (0, 0.5, 1, and 2), and found that NN-SO showed its optimal performance (CV-cor = 0.364) with 5 hidden nodes and a weight decay of 0.5. For various numbers of hidden nodes, optimal performance was reached with weight decay between 0.5 and 1, and the CV-cor started to drop down for weight decay = 2. For Step+NN, we fixed the weight decay to 1.0, and tested with 2, 5, 10, 20, and 30 hidden nodes, and obtained the best result when there were only 2 hidden nodes (CV-cor = 0.465). We also tried different parameter settings for MARS similarly to what we have done on the Oct4 data set. For $d = 1$, we applied MARS with $\lambda = 1, \cdots, 10$ and $15, 20, \cdots, 50$. The optimal CV-cor of 0.553 was achieved with $\lambda = 10$, and it decreased to below 0.45 for $\lambda \leq 2$. For $d = 2$, we applied MARS with $\lambda = 10, 15, \cdots, 50$ and obtained CV-cors in the range of [0.498, 0.539]. In addition, our pilot runs showed that MARS with $d = 2$ performed unsatisfactorily for $\lambda < 10$ (CV-cor < 0.45). Both boosting and BART were very robust to the number of trees included: The CV-cors ranged from 0.534 to 0.560 for boosting and from 0.561 to 0.572 for BART. The best prediction of SVM was achieved by the radial kernel with $C = 1$ and the CV-cor dropped to a level below 0.5 for the other tested values of C (0.1, 10 or 100).

Robustness of BART to negative control sequences

For both the Oct4 and Sox2 ChIP-chip data sets, the results of BART were quite robust to the selection of random sequence regions (serving as negative controls). The average change in P_{in} was about 0.06 across two independent sets of random regions, which makes no qualitative differences in identifying important sequence features.

References

[1] Zhou, Q. and Wong, W.H. (2004) CisModule: *De novo* discovery of cis-regulatory modules by hierarchical mixture modeling, *Proc. Natl. Acad. Sci. USA*, **101**, 12114-12119.

[2] Liu, J.S. and Lawrence, C. E. (1999) Bayesian inference on biopolymer models, *Bioinformatics*, **15**, 38-52.

[3] Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.*, **38**, 431-440.

[4] Matys, V., Fricke, E., Geffers, R. Göβling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374-378.

[5] Chew, J.L. Loh, Y.H., Zhang, W. Chen, X. Tam, W.L., Yeap, L.S., Li, P., Ang, Y.S., Lim, B., Robson, P. and Ng, H.H. (2005) Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol. Cell. Biol.*, **25**, 6031-6046.

[6] Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakarni, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003) The homeoprotein Nanog is required for maintenance of pluripentency in mouse epiblast and ES cells. *Cell*, **113**, 631-642.

[7] Minoguchi, S., Taniguchi, Y., Kato, H., Okazaki, T., Strobl, L.J., Zimber-Strobl, U., Bornkamm, G.W., and Honjo, T. (1997) RBP-L, a transcription factor related to RBP-Jκ. *Mol. Cell. Biol.*, **17**, 2679-2687.

[8] Gu, P., Goodwin, B., Chung, A.C., Xu, X., Wheeler, D.A., Price, R.R., Galardi, C., Peng, L., Latour, A.M., Koller, B.H., Gossen, J., Kliewer, S.A. and Cooney, A.J. (2005) Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell. Biol.*, **25**, 3492-3505.

[9] Vapnik, V. (1998) *The Nature of Statistical Learning Theory* (2nd edition), Springer-Verlag, New York.

[10] Chang, C.C and Lin, C.J. (2001) LIBSVM : a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[11] Chipman, H.A., George, E.I., and McCulloch, R.E. (2006) BART: Bayesian additive regression trees. *Technical Report*, Univ. of Chicago.