

RESEARCH ARTICLE

Construction of precise support vector machine based models for predicting promoter strength

Hailin Meng¹, Yingfei Ma², Guoqin Mai², Yong Wang^{3,*} and Chenli Liu^{1,2,*}

¹ Bioengineering Research Center, Guangzhou Institute of Advanced Technology, Chinese Academy of Sciences, Guangzhou 511458, China

² Center for Synthetic Biology Engineering Research, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

³ Chinese Academy of Sciences Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China

* Correspondence: yongwang@sibs.ac.cn, cl.liu@siat.ac.cn

Received October 10, 2016; Revised December 12, 2016; Accepted December 15, 2016

Background: The prediction of the prokaryotic promoter strength based on its sequence is of great importance not only in the fundamental research of life sciences but also in the applied aspect of synthetic biology. Much advance has been made to build quantitative models for strength prediction, especially the introduction of machine learning methods such as artificial neural network (ANN) has significantly improve the prediction accuracy. As one of the most important machine learning methods, support vector machine (SVM) is more powerful to learn knowledge from small sample dataset and thus supposed to work in this problem.

Methods: To confirm this, we constructed SVM based models to quantitatively predict the promoter strength. A library of 100 promoter sequences and strength values was randomly divided into two datasets, including a training set (≥ 10 sequences) for model training and a test set (≥ 10 sequences) for model test.

Results: The results indicate that the prediction performance increases with an increase of the size of training set, and the best performance was achieved at the size of 90 sequences. After optimization of the model parameters, a high-performance model was finally trained, with a high squared correlation coefficient for fitting the training set ($R^2 > 0.99$) and the test set ($R^2 > 0.98$), both of which are better than that of ANN obtained by our previous work.

Conclusions: Our results demonstrate the SVM-based models can be employed for the quantitative prediction of promoter strength.

Keywords: support vector machine model; quantitative prediction; promoter strength; machine learning

INTRODUCTION

Promoter strength or activity plays a key role in regulating the transcription of downstream genes. Living cells have evolved a number of promoters with a range of strengths to fine-tune the expression of key genes so as to achieve specific physiological function. For artificial biological systems reconstruction, various strength of promoters and

other regulatory elements are also indispensable tools for design of controllable circuits or networks. Although construction of random mutation based library has practical application [1–3], quantitative modeling strategies are still required to improve efficiency and save costs during the design of large scale networks and systems. Therefore, developing methodologies based on model calculation and prediction for designing element sequence will become a trend in the future. Certain modeling methods have been tried to achieve precise prediction or even *de novo* design of element sequence, including a serial of rational methods (e.g., biophysical modeling

This article is dedicated to the Special Collection of Synthetic Biology, Aiming for Quantitative Control of Cellular Systems (Eds. Cheemeng Tan and Haiyan Liu).

[4,5]) and irrational methods (e.g., position weight matrix modeling [6], partial least squares regression modeling [7], and machine learning based modeling [8], etc.). Recent progress on construction of such quantitative models was reviewed and discussed by our previous work [9]. Specifically, as a machine learning based method, artificial neural network (ANN) was employed to characterize the highly nonlinearity between promoter sequence and its strength [8], and a high regression correlation coefficient of $R^2 = 0.96$ was achieved for both model training and test, far outstripping those modeling methods based on linear regression (or its derivative methods). The success of this example demonstrates the application prospect of machine learning methods in predicting promoter strength. As another important machine learning method, Support Vector Machine (SVM) was thus tried to reconstruct such models in this work.

SVM is developed by Vapnik [10,11] in the 1990s based on the statistical learning theory. It applies kernel functions to map the input data into a higher dimensional feature space so as to change the non-linear problem to a linear one in the higher space. Compared to ANN, SVM is a relatively newer and more disciplined machine learning algorithm. Traditional learning methods such as ANN use Empirical Risk Minimization (ERM) criteria to minimize the error of sample points in training, easily resulting in an intractable problem of over fitting. In contrast, SVM takes the criteria of Structural Risk Minimization (SRM) rather than ERM, making it easy to overcome the local minimum and over fitting that usually occurs in ANN modeling, and thus improving the generalization ability of the model. This advantage is more prominent in the study of small samples. Besides, other advantages versus ANN mainly include [12]: i) the automatic structure selection; ii) more advantages in non-linear and high dimensional pattern recognition and function regression; and iii) a more rigorous mathematical derivation and proof. Accordingly, it can be well applied to solve both

classification and regression problems. Due to many distinct advantages, SVMs have been widely employed to different fields of artificial intelligence such as handwritten character recognition, face recognition, text classification, data mining, etc. [12].

In the field of life sciences, SVM is also a powerful tool to build effective predicting models in bioinformatics and computational systems biology, such as protein structure and stability prediction [13,14], RNA secondary structure prediction [15], bacterial transcription start sites prediction [16], virtual screening for drug discovering [17–20], drug metabolism prediction [21], disease prognosis and prediction [22,23], as well as promoter recognition and structure analysis [24–32]. However, SVM has not been reported to use in predicting the strength of promoter or even the regulatory elements. Due to its many advantages compared to ANN, SVM is therefore supposed to be able to build a precise model for prediction of promoter strength after being trained by a small dataset. To this end, we tried to construct a high-performing SVM model for prediction of promoter strength in this work. After multi-parameter optimization, model training and test, we finally obtained the best predicting model, which can precisely output a desired strength value from the promoter sequence (Figure 1).

RESULTS

Model construction and training

The complex relationship between promoter sequence (x) and its strength (y) is considered to be mapped by an SVM regression function $y = f(x)$, $x \in R^l$, $y \in R$. To achieve this goal, SVM models were tried to construct according to Vapnik *et al.* [10,11]. The SVM toolbox [33] running on Matlab platform was employed to build, train and test the SVM models for promoter strength prediction. The performance of constructed SVM models was evaluated by the following two indexes:

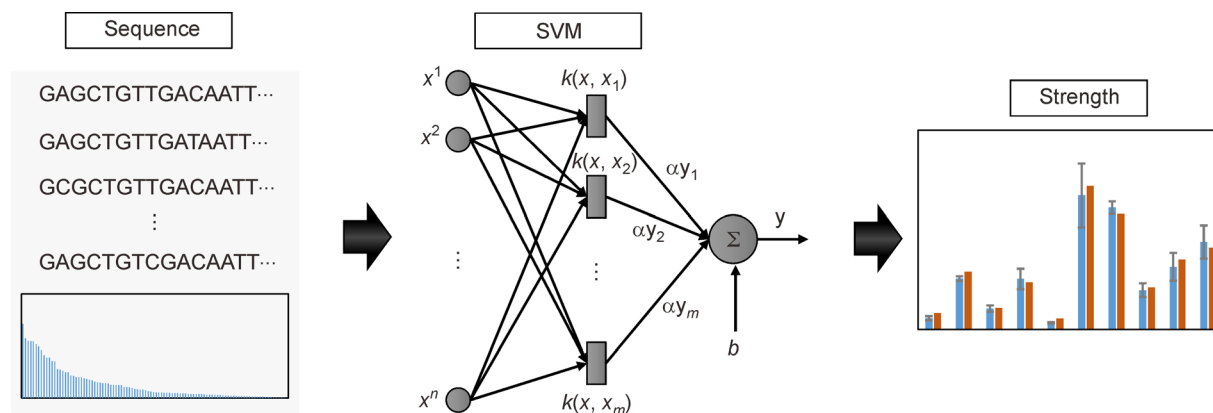


Figure 1. A sketch map of promoter strength prediction based on SVM models.

i) Mean squared error (*MSE*)

$$MSE = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2,$$

and ii) Squared correlation coefficient (R^2)

$$R^2 = \frac{(n \sum_{i=1}^n f(x_i) y_i - \sum_{i=1}^n f(x_i) \sum_{i=1}^n y_i)^2}{(n \sum_{i=1}^n f(x_i)^2 - (\sum_{i=1}^n f(x_i))^2)(n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2)},$$

where $f(x_i)$ and y_i are the strength values of prediction and experiment, respectively. Several kernel functions including the polynomial function, sigmoid function, and radial basis function (RBF) were tried one-by-one in preliminary experiments, and found that the RBF is most suitable for fitting the data.

A mutation library containing 100 promoter sequences and their corresponding strength values [8] was randomly divided into a training set and a test set for training and test of the SVM models, respectively. Different amount of sequences (from 10 to 90) were tried one-by-one to determine the minimum size of training set for reaching the best prediction performance. Each size was independently and randomly sampled for five times, and maximum R^2 and minimum *MSE* values for prediction of the test set were calculated. As a result, R^2 increases and *MSE* decreases with an increase of the size of training set (Figure 2). The best prediction performance was achieved at the size of 90 sequences, similar to that of ANN [8]. Therefore, the model was trained based on 90 sequences and tested on other 10 sequences.

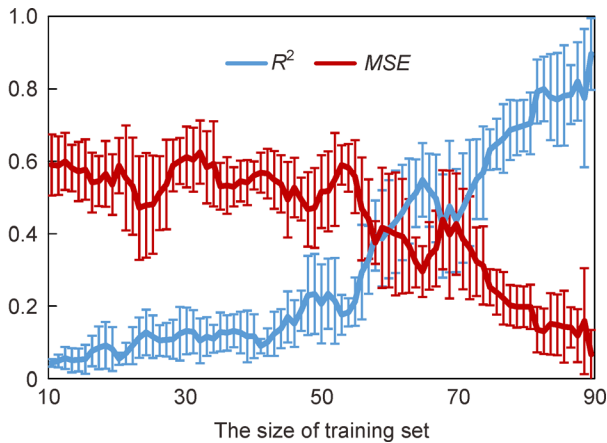


Figure 2. The prediction performance of SVM models as a function of the size of training set. The size of training set ranges from 10 to 90 sequences. Each size was independently and randomly sampled for five times to train SVM models, and maximum R^2 and minimum mean squared error (*MSE*) values for prediction of the test set were calculated.

To achieve a smaller *MSE* or a higher R^2 , a range of feasible values for each parameter (including the balance factor C for loss function and the width value σ for

RBF kernel function) under different precision error ($\varepsilon = 0.01, 0.05, 0.1, 0.2$) were tried by different combinations to get the best parameters. As a result, the general trend of *MSE* is increased with a large C and σ , for both of the training (Figure 3A–3D) and test (Figure 3E–3H). Although the minimum mean *MSEs* (0.561 for training and 0.394 for test) and the minimum MAX *MSEs* (2.39 for training and 2.32 for test, Figure 3D and 3H) were both achieved when $\varepsilon = 0.2$, the best C and σ to get the lowest *MSE* occurred when $\varepsilon = 0.01$. Hence, the best parameter combination $C = 128$, $\sigma = 24.25$ and $\varepsilon = 0.01$ was further chosen to train the model again and then we obtained the best model termed ‘OptModel’. The fitting for training set by ‘OptModel’ shows a high correlation, with $R^2 > 0.99$ (Figure 4A).

Model test, prediction and evaluation

The performance of model ‘OptModel’ was evaluated by applying it to predict the data of test set. Fitting results indicate that a fine correlation was achieved ($R^2 > 0.98$, Figure 4B), and the model can accurately predict the strength of each promoter in the test set (Figure 4C), indicating the success of avoiding over learning problem. Both R^2 of training and test are better than the corresponding values ($R^2 = 0.96$ for training and test) obtained by an ANN model presented by our previous work [8]. To display more intuitively, the fitting results were further compared with that of ANN (Figure 5). Data points of SVM prediction are more concentrated along the diagonal, indicating a better fitting and a precise prediction when compared to ANN model in this case.

Next, the effect of single base mutation on the sequence strength was evaluated by the prediction of OptModel. Each base of the wildtype sequence was mutated to another one (e.g., ‘A’ was changed to ‘C’, ‘G’, or ‘T’), and the corresponding sequence strength was predicted one-by-one (Figure 6A). The highest strength 1.42 occurs at the mutation of $209_{A \rightarrow G}$, and the lowest strength 0.48 appears at $196_{A \rightarrow G}$. An average strength of 0.9 was obtained of all 672 single base mutations, which is smaller than that of the wildtype sequence. Besides, the ‘key-points’ strongly influencing the promoter strength (≥ 1.2 or ≤ 0.8) were further picked out from Figure 6A. Of these ‘key-points’, 82 cases belong to negative mutation (≤ 0.8), far more than 8 cases of the positive (≥ 1.2).

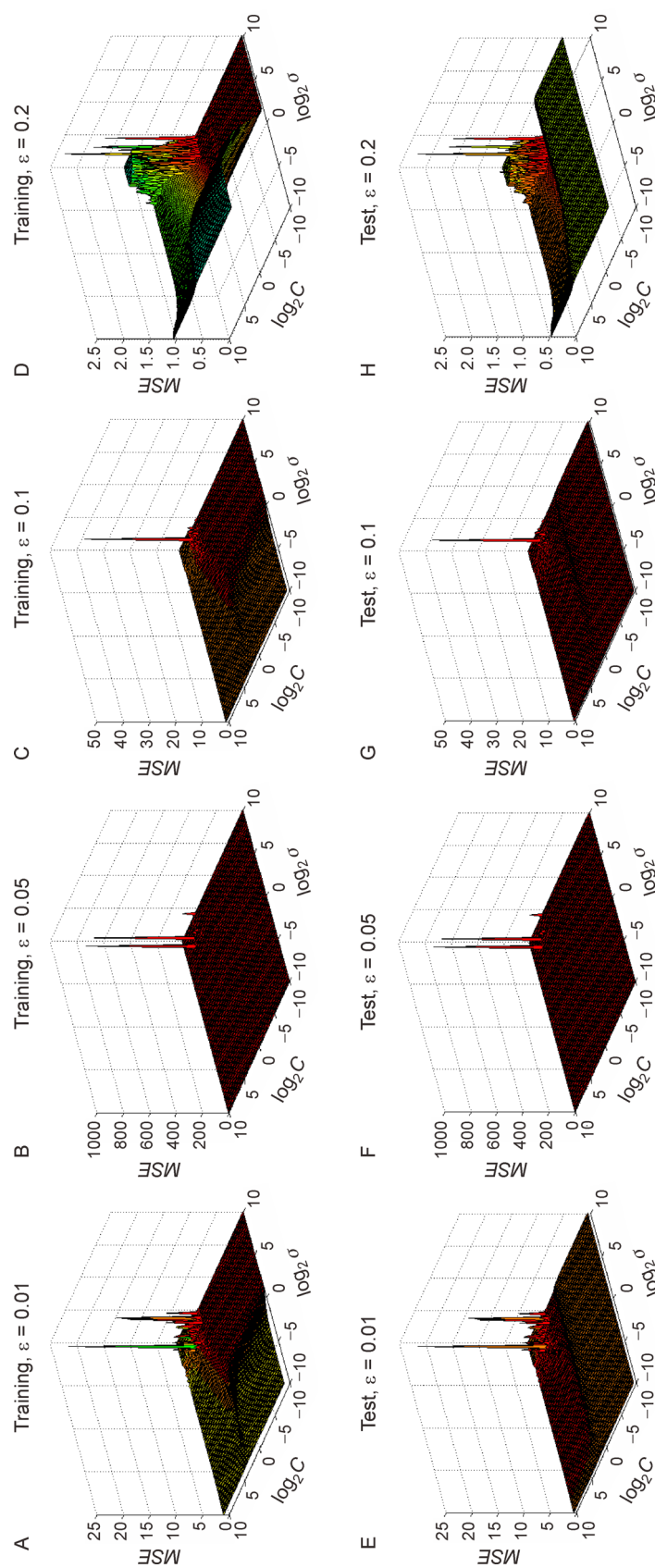


Figure 3. Parameter optimizations for model training. The balance factor C for loss function and the width σ for RBF kernel function under different precision error ϵ were optimized for searching the best parameters. MSE was shown as a function of log₂ C and log₂ σ . (A)–(D) for model training and (E)–(H) for model test.

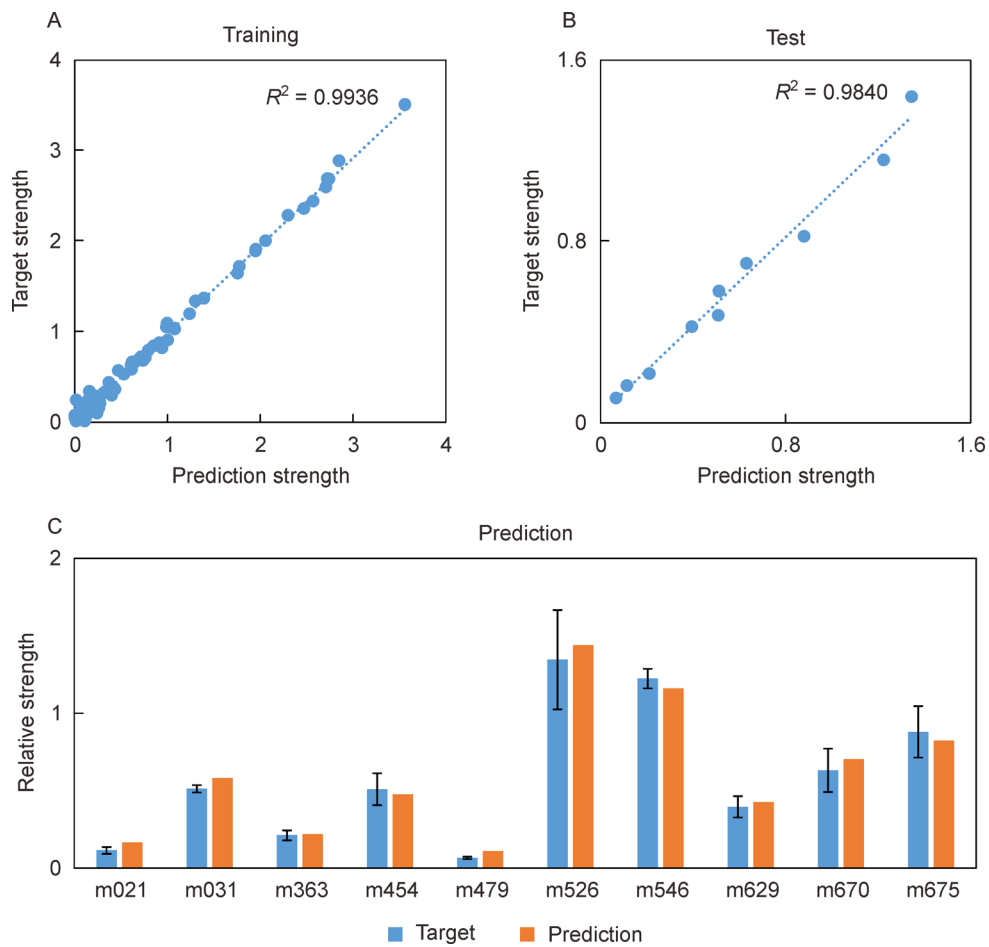


Figure 4. The best model 'OptModel' trained with the optimized parameters ($C = 128$, $\sigma = 24.25$, $\varepsilon = 0.01$) can finely predict the measured promoter strengths. (A) and (B) The predicted relative strengths fit with the measured values using the training set and test set (Supplementary Dataset S1), respectively. (C) Prediction of test set values and comparison with target values (experimental data).

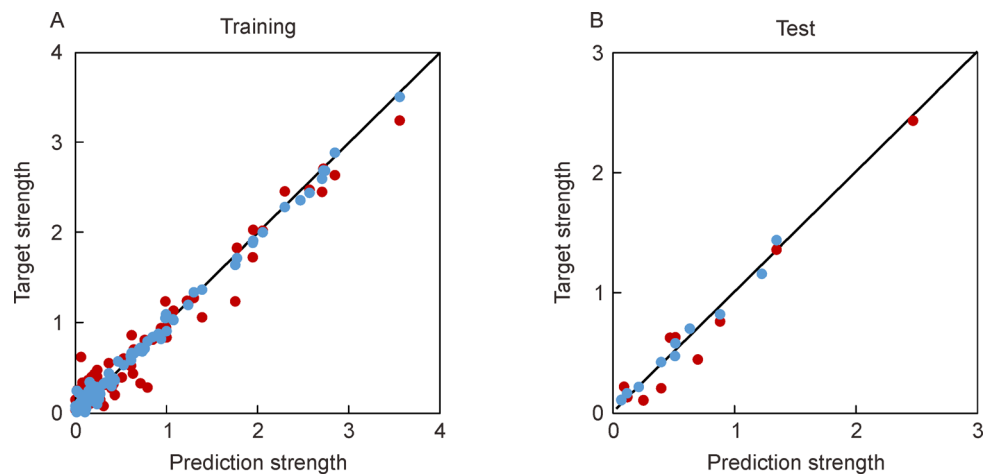
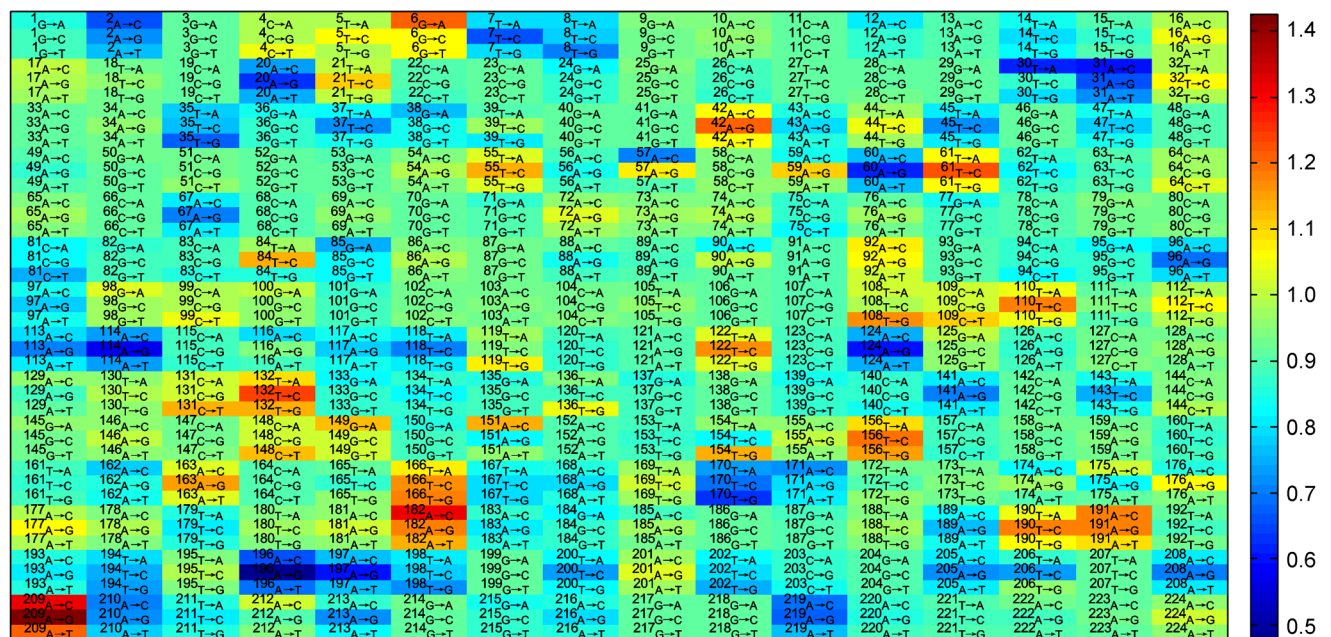


Figure 5. Fitting results comparison: SVM (blue) versus ANN (red). (A) Training. (B) Test. Data points sampled for training set and test set are different between SVM and ANN.

A



B

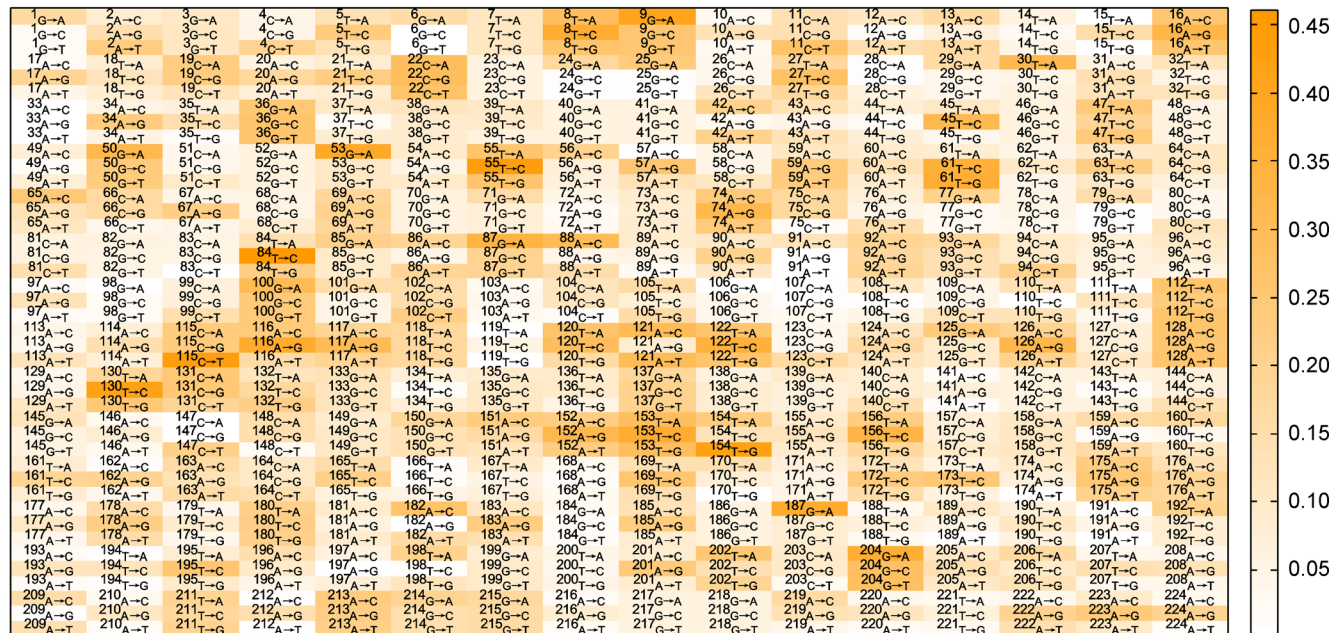


Figure 6. Effect of each single base mutation on the sequence strength predicted by 'OptModel'. (A) and difference between SVM and ANN (B). Red indicates positive mutation while blue the negative. Deeper color means more significant change of the strength. Figure in the boxes is the location number of each base, while the subscript shows the mutation of wildtype base to another one (e.g., A → G denotes A mutated to G).

Most of the mutations (541/672) tally with that of the ANN predictions [8], and only ~19% show significant difference (absolute value ≥ 0.2) between SVM and ANN predictions (Figure 6B), indicating an approximate prediction performance of these two methods.

DISCUSSION

Numerous works demonstrated that the sequences of the regulatory element and its strength/activity have a direct correlation, and many quantitative prediction models

were constructed to bridge the gap between sequence and strength [9]. Of these models, machine learning based methods such as ANN were introduced to this field in recent years, and a better prediction performance indicates the potential of this methodology in synthetic regulatory element design [8]. In this study, another important machine learning method, SVM, was tried to build similar models, and the prediction performance was proven no worse than ANN, demonstrating a promising application prospect of SVM in prediction of prokaryotic promoter strength as well. Besides, the methodology has a good generalization performance on various problems and is expected to do such work in eukaryotic promoters, and even other regulatory elements like terminators.

Machine learning technologies have been widely used in artificial intelligence (AI) and made tremendous progress; especially, rapid development of intelligent robots triggers the coming era of 'Industry 4.0'. Early this year, a powerful ANN-based machine named AlphaGo was designed by Google to play Go game, and surprisingly beat the world champion Sedol Lee with a big score. In the academic field of life sciences, introduction of machine learning based methods has greatly promoted the development of the discipline, especially modeling in bioinformatics and systems biology. Besides aforementioned applications, this work exemplified the use of SVM modeling in a new field, the prediction of promoter strength. Considering the common situation of small samples generalized by various biological experiments, as well as the advantage of SVM in learning small samples, this methodology is also expected suitable for building prediction models on such small dataset like this work.

Although the introduction of machine learning methods can effectively promote the prediction accuracy to a very high level compared to traditional methods, some limitations brought by these kinds of AI algorithms still seem difficult to overcome. It is well-known that the prediction performance of machine learning algorithms directly depends on the 'knowledge' they have learnt. For example in this study, the mutated sequences were generated by error-prone PCR technology based on the wildtype promoter, which introduced only < 30% mutation rate to the initial sequence; so the model cannot get adequate learning from this 'pseudo-random' mutation data and its generalization ability is thus weakened. The best model trained by this library may not precisely predict the strength of those sequences with $\geq 30\%$ mutation rate. Furthermore, 90 sequences for training are inappreciable compared to the sample space ($4^{224} = 7.27 \times 10^{134}$), so the information what the model can learn is very limited. Although SVM is supposed to construct high-performance models based on small samples, the prediction performance of such models

was significantly degraded when using a smaller training set (less than 90 sequences, see Figure 2). But to be sure, the more information learnt by machine learning algorithms, the more powerful the model can be. Just as the well-trained AlphaGo machine, it is quite hard for human to revenge on it henceforward. With rapid growth of experimental data, more powerful or intelligent AI models may be constructed to extensively learn knowledge from multiple datasets including strength of different regulatory elements from various kinds of species, thereupon greatly improve the precision and generalization ability, and emancipate us from repeated and laborious experiments.

METHODS

Computational platform and tools

Matlab 2013a (Mathworks Inc., <http://www.mathworks.com/>) ran on a personal computer with Microsoft Windows 10 operating system (Microsoft Inc., <http://www.microsoft.com/>). SVM Toolbox [33] was integrated into Matlab and served as the computational tool for SVM model construction, data training and prediction. All calculations and simulations were programed and run upon SVM Toolbox and Matlab environment.

Data sources and dataset preparation

A mutation library of *Escherichia coli* Trc promoter containing 100 promoter sequences and their corresponding strength values was constructed by our previous work [8]. Briefly, error-prone PCR was performed on Trc promoter region in plasmid pTrcHis2B (224 bp, including -35 box, -10 box, RBS, and other regions) to introduce random mutation; then a *gfp* gene was inserted into the plasmid as a reporter; finally the promoter strength was assayed by detecting GFP expression in *E. coli* using flow cytometry. The strength of each sequence is a relative value compared to the wildtype sequence. In this work, the library was used to train and test SVM models. Similar to the previous work, the library was randomly divided into two datasets, including a training set and a test set (Supplementary Dataset S1 shows an example sampling for training the best performance model 'OptModel').

The original sequence data coded by 'A', 'G', 'C' and 'T' were translated to digital matrix for SVM model input according to the following orthogonal rules: 'A' = [1, 0, 0, 0], 'G' = [0, 1, 0, 0], 'C' = [0, 0, 1, 0], and 'T' = [0, 0, 0, 1]. For instance, a given sequence 'AGTGCC' can be translated to a '0-1' digital series of [1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0] under this conversion rule.

Model training and test

A range of values were set for each parameter to train the SVM model. Specifically, set the balance factor $C = 2^{-10} \sim 2^{10}$, the width value of kernel function $\sigma = 2^{-10} \sim 2^{10}$, and the precision error $\varepsilon = 0.01, 0.05, 0.1, 0.5$. Other setting conditions include: 'eInsensitive' for the loss function, and 'rbf' for the kernel function. A 2-layer nested loop was employed to search for the best parameters during model training and test. The mean squared error (*MSE*) and squared correlation coefficient (R^2) were calculated as the index to evaluate the performance of model training and test under each setting of parameters. An absolute value was operated on the prediction of training and test data, since the strength values are non-negative. Finally, the best combination of parameters (C , σ and ε) was used to retrain and generate the best model.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-017-0096-3.

ACKNOWLEDGEMENTS

This work was financially supported by NSFC (Nos. 31471270, 31301017, 31670056 and 31300686), 973 Program (No. 2014CB745202), 863 Program (No. SS2015AA020936), the Guangdong Natural Science Funds for Distinguished Young Scholar (No. S2013050016987), the Science and Technology Planning Project of Guangdong Province (Nos. 2014B020201001 and 2014A030304008), Natural Science Foundation of Guangdong Province (No. 2015A030310317), the Guangzhou Science and Technology Scheme (Nos. 201508020091 and 201508020092), and Shenzhen grants (Nos. KQTD2015033117210153, JCYJ20140610152828703, KQJSCX20160301144623, CXZZ20140901004122088, JCYJ20150521144321007 and JCYJ20140901003939019).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Hailin Meng, Yingfei Ma, Guoqin Mai, Yong Wang and Chenli Liu declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

- Blount, B. A., Weenink, T., Vasylechko, S. and Ellis, T. (2012) Rational diversification of a promoter providing fine-tuned expression and orthogonal regulation for synthetic biology. *PLoS One*, 7, e33279
- Qin, X., Qian, J., Yao, G., Zhuang, Y., Zhang, S. and Chu, J. (2011) GAP promoter library for fine-tuning of gene expression in *Pichia pastoris*. *Appl. Environ. Microbiol.*, 77, 3600–3608
- Alper, H., Fischer, C., Nevoigt, E. and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. USA*, 102, 12678–12683
- Salis, H. M., Mirsky, E. A. and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, 27, 946–950
- Lou, C., Stanton, B., Chen, Y. J., Munskey, B. and Voigt, C. A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.*, 30, 1137–1142
- Rhodijs, V. A. and Mutalik, V. K. (2010) Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, σ^E . *Proc. Natl. Acad. Sci. USA*, 107, 2854–2859
- De Mey, M., Maertens, J., Lequeux, G. J., Soetaert, W. K. and Vandamme, E. J. (2007) Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering. *BMC Biotechnol.*, 7, 34
- Meng, H., Wang, J., Xiong, Z., Xu, F., Zhao, G. and Wang, Y. (2013) Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network. *PLoS One*, 8, e60288
- Meng, H. and Wang, Y. (2015) Cis-acting regulatory elements: from random screening to quantitative design. *Quant. Biol.*, 3, 107–114
- Vapnik, V. N. (2000) *The Nature of Statistical Learning Theory*. New York: Springer-Verlag
- Vapnik, V. N. (1999) An overview of statistical learning theory. *IEEE Trans. Neural Netw.*, 10, 988–999
- Hassanien, A. E., Al-Shammari, E. T. and Ghali, N. I. (2013) Computational intelligence techniques in bioinformatics. *Comput. Biol. Chem.*, 47, 37–47
- Ho, H. K., Zhang, L., Ramamohanarao, K. and Martin, S. (2013) A survey of machine learning methods for secondary and supersecondary protein structure prediction. In *Methods and Protocols: Methods in Molecular Biology*, 932, 87–106 New York: Humana Press
- Cheng, J., Tegge, A. N. and Baldi, P. (2008) Machine learning methods for protein structure prediction. *IEEE Rev. Biomed. Eng.*, 1, 41–49
- Zhao, Y. and Wang, Z. (2008) RNA secondary structure prediction based on support vector machine classification. *Chinese Journal of Biotechnology*, 24, 1140–1148
- Towsey, M. W., Gordon, J. J. and Hogan, J. M. (2006) The prediction of bacterial transcription start sites using SVMs. *Int. J. Neural Syst.*, 16, 363–370
- Ichikawa, D., Saito, T., Ujita, W. and Oyama, H. (2016) How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. *J. Biomed. Inform.*, 64, 20–24
- Vyas, R., Bapat, S., Jain, E., Tambe, S. S., Karthikeyan, M. and Kulkarni, B. D. (2015) A study of applications of machine learning based classification methods for virtual screening of lead molecules. *Comb. Chem. High Throughput Screen.*, 18, 658–672
- Burton, J., Ijjaali, I., Petit, F., Michel, A. and Vercauteren, D. P. (2009) Virtual screening for cytochromes p450: successes of machine learning filters. *Comb. Chem. High Throughput Screen.*, 12, 369–382
- Melville, J. L., Burke, E. K. and Hirst, J. D. (2009) Machine learning in virtual screening. *Comb. Chem. High Throughput Screen.*, 12, 332–343
- Fox, T. and Kriegl, J. M. (2006) Machine learning techniques for *in silico* modeling of drug metabolism. *Curr. Top. Med. Chem.*, 6, 1579–1591
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015) Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, 13, 8–17
- Polley, M. Y., Freidlin, B., Korn, E. L., Conley, B. A., Abrams, J. S. and McShane, L. M. (2013) Statistical and practical considerations for clinical evaluation of predictive biomarkers. *J. Natl. Cancer Inst.*, 105,

- 1677–1683
24. Liang, G. and Li, Z. (2007) Scores of generalized base properties for quantitative sequence-activity modelings for *E. coli* promoters based on support vector machine. *J. Mol. Graph. Model.*, 26, 269–281
 25. Towsey, M., Timms, P., Hogan, J. and Mathews, S. A. (2008) The cross-species prediction of bacterial promoters using a support vector machine. *Comput. Biol. Chem.*, 32, 359–366
 26. Xu, W., Zhang, L. and Lu, Y. (2016) SD-MSAEs: promoter recognition in human genome based on deep feature extraction. *J. Biomed. Inform.*, 61, 55–62
 27. Sato, M. (2012) Promoter analysis with wavelets and support vector machines. *Procedia Comput. Sci.*, 12, 432–437
 28. Holloway, D. T., Kon, M. and Delisi, C. (2007) Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Syst. Synth. Biol.*, 1, 25–46
 29. Anwar, F., Baker, S. M., Jabid, T., Mehedi Hasan, M., Shoyaib, M., Khan, H. and Walshe, R. (2008) Pol II promoter prediction using characteristic 4-mer motifs: a machine learning approach. *BMC Bioinformatics*, 9, 414
 30. Carvalho, S. G., Guerra-Sá, R. and de C Merschmann, L. H. (2015) The impact of sequence length and number of sequences on promoter prediction performance. *BMC Bioinformatics*, 16, S5
 31. Hwang, W., Oliver, V. F., Merbs, S. L., Zhu, H. and Qian, J. (2015) Prediction of promoters and enhancers using multiple DNA methylation-associated features. *BMC Genomics*, 16, S11
 32. Li, Y., Lee, K. K., Walsh, S., Smith, C., Hadingham, S., Sorefan, K., Cawley, G. and Bevan, M. W. (2006) Establishing glucose- and ABA-regulated transcription networks in *Arabidopsis* by microarray analysis and promoter classification using a Relevance Vector Machine. *Genome Res.*, 16, 414–427
 33. Sandhu, R. S., Coyne, E. J., Feinstein, H. L. and Youman, C. E. (1996) Role based access control models. *IEEE Computer*, 29, 38–47