Prediction of Insertion-Site Preferences of Transposons Using Support Vector Machines and Artificial Neural Networks

Maryam Ayat and Michael Domaratzki

Bioinformatics Lab, Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
{ayatmary,mdomarat}@cs.umanitoba.ca

Abstract. Transposons are segments of DNA that are capable of moving from one location to another within the genome of a cell. Understanding transposon insertion-site preferences is critically important in functional genomics and gene therapy studies. It has been found that the deformability property of the local DNA structure of the integration sites, called V_{step}, is of significant importance in the target-site selection process. We considered the V_{step} profiles of insertion sites and developed predictors based on Artificial Neural Networks (ANN) and Support Vector Machines (SVM), and trained them with a Sleeping Beauty transposon dataset. We found that both ANN and SVM predictors are excellent in finding the most preferred regions. However, the SVM predictor outperforms the ANN predictor in recognizing preferred sites, in general.

1 Introduction

Transposons, or jumping genes [1], are short mobile DNA sequences that can insert themselves into the genome of the cell (i.e., host genome) and replicate. They are used in transferring genes of interest into the genome of the target cell and have applications in discovering function of genes (especially those that cause cancer) as well as in gene therapy (e.g., therapy of genetic disorders in humans). However, the applicability of a transposon for these uses depend highly on the target-site selection properties, which are not well understood. Predicting hotspots, or most preferred insertion sites of transposons helps in determining the risks of adverse effects that a transposon insertion may have.

There may be many factors that affect preferences in transposon integration [2], but among the studied factors the local DNA structure has a more effective role in the target-site selection process, as Liu et al. [3] showed for the Sleeping Beauty transposon (SB). Liu et al. found that there is a relationship between the natural deformability property of target sites, which is described by a parameter called V_{step} [4], and the mechanism of the target-site selection of the SB transposon. The composite parameter V_{step} represents the physical relationships of any two planar base pairs in term of their relative displacements and angular orientation in the 3D-structure of DNA [2]. It is a measure of dimer deformability: the higher the V_{step} value, the more deformable the dimer step is, where steps

N. El Gayar et al. (Eds.): ANNPR 2014, LNAI 8774, pp. 183–192, 2014.

Table 1. V_{step} values for dimer steps

dimer	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
V_{step}	2.9	2.3	2.1	1.6	9.8	6.1	12.1	2.1	4.5	4.0	6.1	2.3	6.3	4.5	9.8	2.9

refer to dinucleotides along a DNA sequence. Table 1 shows the V_{step} values for all possible dimer steps. For the details on V_{step} calculations, see [3,4].

Based on the work done by Liu et al., Geurts et al. [5] analyzed integration-site preferences of SB, piggyBac and Drosophila P-element transposons to detect V_{step} patterns for preferred sites. However, they did not succeed in finding consistent V_{step} pattern for all of the studied transposons. The main drawback of Geurts et al.'s method, in our opinion, is in the way of developing the preference rules, which is more ad-hoc than structured: first, they find the preferred sites based on observations; then, they try to infer the general form of V_{step} patterns of a transposon preferred sites by visually comparing the V_{step} diagrams of observed integration sites. To resolve this weakness, we used machine learning methods for predicting transposon insertion sites.

We considered the insertion site prediction problem as a classification problem, and constructed two types of predictors: one based on Support Vector Machines (SVMs) and the other based on Artificial Neural Networks (ANNs). Both SVMs and ANNs have applications in classification and regression problems, and have been widely used in bioinformatics [6,7]. We employed these predictors for identifying preferred regions (100 bp sequences) in a host genome based on the V_{step} profile of the individual insertion sites (12 bp sequences). To evaluate each predictor, we used a five-fold cross-validation on a SB transposon dataset. Finally, we compared the results of SVM and ANN predictors to each other as well as to Geurts et al.'s results.

2 Materials and Methods

2.1 Dataset

We used an SB transposon integration dataset for training and testing our predictors from the Hackett lab [8]. The main preference of the SB transposon is TA dinucleotides sites in a host genome. We possessed a 7758 bp plasmid pFV/Luc sequence, the actual SB transposon TA integration sites, and the number of hits per integration site in the host sequence. Therefore, our dataset consisted of all TA sites of the 7758 bp plasmid pFV/Luc sequence along with the insertion frequencies. In the 7758 bp plasmid pFV/Luc sequence, there were 489 TA sites with 193 total number of insertions in 97 sites. Similar to Geurts et al. [5], we used the V_{step} profile of a window of 12 bp, including 5 bp flanking each side of a target TA dinucleotide. A V_{step} vector has 11 elements, as there are 11 consecutive dinucleotides in a 12 bp subsequence.

We normalized the V_{step} values and scaled them to the range $[0, 1]$ using the min-max normalization technique. Also, we normalized the integration frequencies to the range $[0, 1]$, since their actual values were very small (less than 0.05).

2.2 Performance Measures

For each predictor, we measured sensitivity (SN), specificity (SP), and the overall accuracy (ACC). They are defined as:

$$SN = \frac{TP}{TP + FN},$$

$$SP = \frac{TN}{TN + FP},$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN},$$

where TP, FN, TN, and FP refer to the number of true positives, false negatives, true negatives, and false positives, respectively.

To evaluate the strength of a classification, we also generated a Receiver Operating Characteristic (ROC) curve and computed the area under the ROC curve (AUC). An AUC close to 1 indicates a strong test, and an AUC close to 0.5 represents a weak test.

2.3 Support Vector Machine

We used an SVM [9] with a Gaussian Radial Basis Function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma(\mathbf{x}_i - \mathbf{x}_j)^2}$. SVM is used in its basic form for binary classification, and a version of SVM for function estimation from a set of training data is Support Vector Regression (SVR) [10]. In an epsilon-SVR model the margin of error tolerance, i.e., ϵ, should be set as a parameter. Our designing model for predicting preferences of transposon insertion sites has two phases. In the first phase, we construct the best binary SVM for predicting preferred individual insertion sites, and in the second phase, we construct an SVR with the same architecture as the best SVM resulting from the first phase. This SVR predicts the insertion distribution along the sequence bins (i.e., regions).

To implement our SVM predictior, we used the SVM package LIBSVM [11] in MATLAB environment. In following, we illustrate the SVM architecture for finding preferred individual sites and regions, respectively.

Preferred Individual Sites. In this case, the SVM-based tool predicts if a given 12 bp insertion site is a preferred SB transposon integration site or not. The input data for the SVM is the V_{step} vector for an insertion site of 12 bp, and the output is the label of the class, i.e., +1 or -1 corresponding to a preferred or not-preferred insertion site. We used a binary SVM with a Gaussian kernel for classification. In this architecture, there are two parameters: the soft margin parameter C and the kernel parameter γ. We applied a grid-search [12] on C and γ for selecting parameters. Using this method, we tried various pairs of (C, γ), within the range $1 \leq \log_2 C \leq 5$ and $-3 \leq \log_2 \gamma \leq 1$, and chose the one with the best 5-fold cross-validation accuracy which we measured by the area under the

ROC curve. We also set the parameter C for both positive and negative classes by different weights due to having unbalanced data. The final Gaussian-kernel SVM has the following configuration: the best parameters $(C, \gamma) = (2.0, 0.25)$, and weights for C in positive and negative classes $(w_{+1}, w_{-1}) = (20, 1)$.

To find the best results, we also tried different definitions of preferred insertion sites (i.e., positive class) in terms of the number of hits (i.e., the integration frequency). The best results of SVM were obtained when we assumed a preferred insertion site as a site with more than two integrations.

Preferred Regions. Assuming that we are given the V_{step} scores of a sequence, which is divided into bins of size 100 bp, the SVM predicts the most preferred SB transposon insertion bins (i.e., it predicts the insertion distribution along the sequence bins). In this case, we took advantage of support vector regression. We constructed an epsilon-SVR, with a Gaussian kernel and the same parameters we had found in the binary SVM, to model the relationship between the V_{step} vectors and integration frequencies of insertion sites. We also set the tolerance criterion, ϵ, to 0.001. Then, we ran a 5-fold cross validation over all insertion sites, and obtained the predicted frequency for each insertion site. Afterward, we computed the summation of frequencies for each bin, scaled them to the range [0,1], and obtained the distribution of predicted integration frequencies.

2.4 Radial Basis Function Neural Network

We took advantage of a three-layer RBF neural network [13] for prediction. RBF networks are suited for pattern recognition problems such as this research wherein the pattern dimension is sufficiently small. Similar to the SVM solution, our designing model for predicting preferences of transposon insertion sites has two phases. In the first phase, we construct the best RBF network for predicting preferred individual insertion sites, and in the second phase, we obtain the insertion distribution along the sequence bins based on the best RBF architecture resulting from the first phase.

We constructed our ANN predictor using Open Desire package [14]. In following, we illustrate the ANN architecture for finding preferred individual sites and regions, respectively.

Preferred Individual Sites. In this case, our RBF neural network predicts if a given 12 bp insertion site is a preferred SB transposon insertion site or not. The input data for the ANN is the V_{step} vector for an insertion site of 12 bp, and the output is the insertion frequency which is converted to a binary value 1/0 corresponding to a preferred or not-preferred site. For this purpose, we constructed a set of RBF networks with different configurations and applied a 5-fold cross validation over each to find the best neural network. Finding the best RBF neural network requires searching for the optimal number of hidden units, as well as the parameter σ (,or $\gamma^{-0.5}$) in the radial basis function and the threshold values. We used a destructive method in design. We started with a

network with a maximal number of hidden units and connections, and gradually deleted hidden units to reach the optimal network. Meanwhile, we tried to find the best σ by a random selection for each network. We also benefited from ROC curve analysis to find a threshold or cut-off for generating binary outputs. The best RBF network has 262 hidden units and parameter $\sigma = 2.75$.

Similar to the SVM solution, we tried different definitions of preferred insertion sites in terms of the number of hits. The best ANN was obtained from the situation in which we defined a preferred insertion site as a site with more than two integrations.

Preferred Regions. Having the V_{step} scores of a sequence, which is divided into 100 bp bins, the ANN predicts the insertion distribution along the sequence bins and recognizes the most preferred SB transposon insertion regions. Here, we used the same constructed RBF neural network for individual sites, and ran a 5-fold cross validation over all insertion sites, but we did not apply any threshold on output frequencies. Instead, we computed the summation of frequencies for each bin, scaled them to the range [0,1], and obtained the distribution of predicted integration frequencies.

3 Results

3.1 SVM Results

In Individual Sites. Figure 1 shows the ROC curve for the final SVM predictor. The cut-off point recognizes the best binary SVM which has 83% sensitivity and 72% specificity. The area under the curve is 0.85, which indicates that the SVM predictor has a good performance in finding preferred individual insertion sites.

In Regions. Figure 3(a) shows a plot for comparing distribution of observed and predicted insertion sites in the 7758 bp pFV/Luc sequence. The sequence is divided into 77 bins of 100 bp. The plot illustrates an apparent overlap between the two distributions. For example, it shows that the SVM could predict the four most preferred bins (bins #16, #17, #47, and #69) successfully. Therefore, if our concern is to predict the preferred regions in the sequence, then the epsilon-SVR will produce better results compared to the binary SVM for individual sites. Using ROC curve analysis, we found that the epsilon-SVR predictor has 100% SN, 94% SP, and AUC=0.97 in recognizing the most preferred insertion regions. We considered the most preferred insertion regions in our observed data as the bins in which the number of insertions is more than three (i.e., bins in which the scaled insertion frequency is more than 0.4). Also, we found that the epsilon-SVR predictor has 85% SN, 90% SP, and AUC=0.89 in recognizing the preferred insertion regions, which we considered them in our observed data as the bins in which the number of insertions is more than two (i.e., bins in which the scaled insertion frequency is more than 0.3). The AUC values indicate an excellent discriminatory power of the SVM in finding preferred insertion regions.

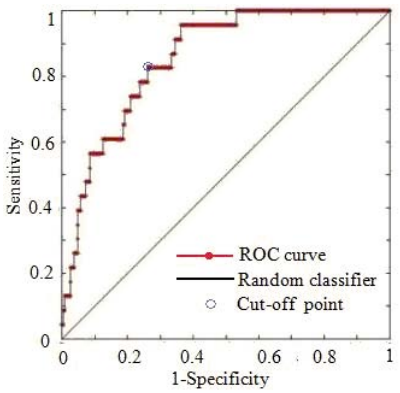


Fig. 1. ROC curve for the best Gaussian kernel SVM in individual sites, AUC=0.85

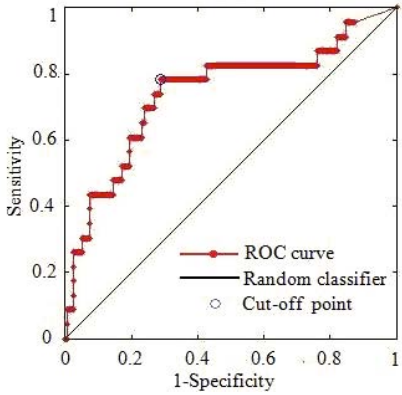


Fig. 2. ROC curve for the best RBF network in individual sites, AUC=0.71

3.2 ANN Results

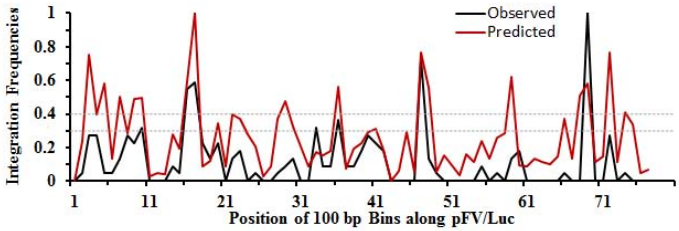
In Individual Sites. Figure 2 shows the generated ROC curve for the best network. It revealed 78% SN and 71% SP for the cut-off point. The AUC is 0.71, which shows the classifier has a fairly good discriminatory power.

In Regions. Figure 3(b) shows a plot for comparing distribution of observed and predicted insertion sites in the 7758 bp pFV/Luc sequence. The sequence is divided into 77 bins of size 100 bp. The plot illustrates an apparent overlap between the two distributions. For example, it shows that the neural network could predict the four most preferred bins (bins #16, #17, #47, and #69) successfully. Therefore, if our concern is to predict the preferred regions in the sequence, then our RBF neural network will produce better results compared to predicting individual sites. Using ROC curve analysis, we found that the RBF network predictor has 100% SN, 97% SP, and AUC=0.98 in recognizing the most preferred insertion regions (i.e., bins in which the scaled insertion frequency is more than 0.4). Also, we found that the ANN predictor has 100% SN, 72% SP, and AUC=0.90 in recognizing the preferred insertion regions (i.e., bins in which the scaled insertion frequency is more than 0.3). Both AUCs indicate an excellent discriminatory power of the network in finding preferred insertion regions.

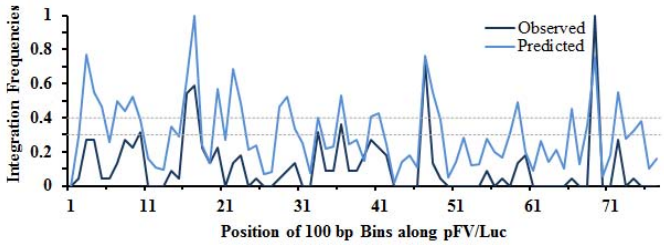
4 Discussion

4.1 SVM versus ANN

We have summarized the performance measures of the ANN- and SVM-based predictors for finding preferred individuals sites and regions in two different tables. Table 2 contains the 5-fold cross-validation result of the predictors in



(a) SVM-predicted insertion frequencies



(b) ANN-predicted insertion frequencies

Fig. 3. Distribution of observed versus distribution of SVM and ANN-predicted insertion frequencies in the 7758 bp plasmid pFV/Luc. The sequence is divided into 77 bins of size 100 bp. Dashed lines show threshold values 0.3 and 0.4, used in defining preferred and most preferred insertion bins in observed data, respectively.

individual sits. According to these results, the SVM-based predictor outperforms the ANN-based predictor in identifying preferred individual insertion sites due to having higher AUC, SN and SP. Table 3 contains the 5-fold cross-validation result of the predictors in identifying preferred 100 bp insertion regions. Based on these results, both predictors are excellent in recognizing most preferred insertion regions (similar values for AUC, SN, and SP), but the SVM performs better in identifying preferred regions (similar AUCs, but higher ACC).

Also, it is interesting that comparing the results of either of ANN or SVM predictor in individual sites and regions shows that both ANN and SVM predictors performs better in regions than individual sites. An explanation for this fact might be that the amount of preferability of an insertion site not only depends on the local sequence itself, but also depends on the region that encompasses the insertion site. Therefore, it is worth considering larger sequences of insertion sites than 12 bp as inputs for the ANN and SVM predictors, or adding some region-related features (e.g., the number of TA sites in a region) to the current models in the future.

4.2 Comparison with Related Work

Geurts et al. [5] developed rules for describing the insertion-site preferences of the SB transposon. They did not report any SN, SP or ACC, neither for individual

Table 2. ANN versus SVM predictor in identifying preferred individual sites

Predictor	SN(%)	SP(%)	ACC(%)	AUC
ANN	78	71	72	0.71
SVM	83	72	95	0.85

Table 3. ANN versus SVM predictor in identifying preferred 100 bp regions

Predictor	Prediction	SN(%)	SP(%)	ACC(%)	AUC
ANN	most preferred	100	97	96	0.98
	preferred	100	72	75	0.90
SVM	most preferred	100	94	94	0.97
	preferred	85	90	89	0.89

sites nor for regions, as they had not made any predictor based on their rules. These rules categorizes each 12 bp TA site into one of the three classes - basal, semi-preferred and preferred - based on the graphical pattern of its V_{step} profile (e.g., if a V_{step} profile of a site has 4 peaks in its diagram, it will be categorized to the preferred class). To demonstrate the successfulness of their rules (for SB transposon) in finding preferred the 7758 bp plasmid pFV/Luc insertion regions, Geurts et al. classified all the TA sites based on their rules. Then, according to the ratio of the actual number of insertions in each class to the number of TA sites of that class in the sequence, they provided a formula for calculating the total V_{step} score of a bin of given length in the sequence. Next, they divided the sequence into 100 bp bins, produced the distribution of total V_{step} scores in bins, and compared the result with the distribution of observed insertion sites. In this way, Geurts et al. succeeded to identify the three most preferred bins (bins #17, #47, and #69) in the sequence.

To be able to compare our results with Geurts et al.'s, we used the distribution of total V_{step} scores in the pFV/Luc sequence, and measured the classification power of Geurts et al.'s rules. Consequently, similar to the SVM predictor in regions, we benefited from ROC curve analysis and found that the rules have 100% SN, 89% SP, and AUC=0.97 in finding most preferred bins, and 85% SN, 91% SP and AUC=0.91 in finding preferred bins.

Based on these results, we conclude that:

1. Both SVM and ANN predictors identify the four most preferred bins, while Geurts et al.'s rules recognized the three tops. Due to the higher SP, our predictors perform better in recognizing most preferred regions, compared to Geurts et al.'s rules; and
2. The SVM predictor performs as well as Geurts et al.'s rules in identifying preferred regions.

5 Conclusion

In this paper, we demonstrated how machine learning methods such as SVMs and ANNs can be used for predicting insertion sites of transposons based on

the deformability property of the local DNA structure of the integration sites, or their V_{step} profiles. We constructed two predictors based on ANN and SVM methods for identifying insertion-site preferences of SB transposon in a genome, knowing that the main preference of SB trasposon is TA sites. Our model, either for SVM or ANN predictor, had two phases for predicting. In the first phase, we constructed a binary classifier for identifying preferred individual insertion sites (12 bp sites), and in the second phase, we constructed a predictor with the same architecture as the best classifier resulting from the first phase, but this time for regression purposes, or in other words, for predicting the insertion distribution along the sequence bins (100 bp regions). Using five-fold cross validation, we performed the parameter optimization process and evaluation of our SB predictors. However, measuring the performance of the final predictors by testing other host genomes remains as the next step.

We also compared our approach to Geurts et al.'s rule-based method. Our results show that both ANN and SVM predictors outperform Geurts et al.'s heuristic rules in finding the most SB preferred regions. Also, the SVM predictor outperforms the ANN predictor and is as good as Geurts et al.'s rules in recognizing preferred sites in general. However, the main preference of machine learning solutions such as ANNs and SVMs over Geurts et al.'s rule-based method is that these predictors are able to extract the rules, or the relations between inputs and outputs, themselves. It is for this reason Geurts et al.'s ad-hoc rules were not successful in identifying preferred insertion sites of the other transposons in general. Moreover, ANN and SVM predictors are scalable, so some other factors that may influence the insertion-site selection process can easily be modeled in them as new features. Therefore, it is worth constructing other transposon-specific predictors based on these methods as a future work. Such predictors can help direct experiments by helping researchers focus on potential regions of high likelihood of insertion before beginning experiments.

References

1. Pray, L.A.: Transposons: The jumping genes. Nature Education 1(1) (2008)
2. Hackett, C.S., Geurts, A.M., Hackett, P.B.: Predicting preferential DNA vector insertion sites: Implications for functional genomics and gene therapy. Genome Biology 8(S12 Suppl. 1) (2007)
3. Liu, G., Geurts, A.M., Yae, K., Srinivasan, A.R., Fahrenkrug, S.C., Largaespada, D.A., Takeda, J., Horie, K., Olson, W.K., Hackett, P.B.: Target-site preferences of sleeping beauty transposons. Journal of Molecular Biology 346(1), 161–173 (2005)
4. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M., Zhurkin, V.B.: DNA sequence - dependent deformability deduced from protein - DNA crystal complexes. Proceedings of the National Academy of Sciences of the United States of America: PNAS 95(19), 11163–11168 (1998)
5. Geurts, A.M., Hackett, C.S., Bell, J.B., Bergemann, T.L., Collier, L.S., Carlson, C.M., Largaespada, D.A., Hackett, P.B.: Structure-based prediction of insertion-site preferences of transposons into chromosomes. Nucleic Acid Research 34(9), 2803–2811 (2006)

6. Baldi, P., Brunak, S.: Bioinformatics: The machine learning approach, 2nd edn. MIT Press, Cambridge (2001)
7. Seiffert, U., Hammer, B., Kaski, S., Villmann, T.: Neural Networks and Machine Learning in Bioinformatics - Theory and Applications. In: European Symposium on Artificial Neural Networks, ESANN, pp. 521–532 (2006)
8. Hackett, P.B.: Sleeping Beauty transposon insertion data in the 7758 bp plasmid pFV/Luc, [Personal Communication] (2011)
9. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, Inc. (1998)
10. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199–222 (2004)
11. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
12. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2010), `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`
13. Haykin, S.S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (2009)
14. Korn, G.A.: Advanced Dynamic-System Simulation: Model-Replication Techniques and Monte Carlo Simulation. John Wiley & Sons, Inc. (2007)