Seventh IEEE International Conference on Data Mining - Workshops

Statistical Approaches to Identifying Androgen Response Elements

Li Li
North Carolina State Univeristy
lli7@ncsu.edu

Steffen Heber
North Carolina State Univeristy
sheber@ncsu.edu

Qiang Zhang
The Hamner Institutes for Health Sciences
qzhang@thehamner.org

Melvin E Andersen
The Hamner Institutes for Health Sciences
mandersen@thehamner.org

Abstract

DNA-binding transcription factors play an integral role in regulating gene expression. Transcription factor binding sites (TFBS) in the gene promoter regions can be predicted by using computational methods, such as Support Vector Machine (SVM), Hidden Markov Model (HMM), and Random Forest (RF), all of which summarize sequence patterns of experimentally determined TFBSs. Androgen receptor (AR), a ligand-dependent transcription factor, plays an important role in male reproductive functions by regulating gene transcription through directly binding to androgen response elements (ARE) in target gene promoters. The aim of this study is to use data mining tools to identify and characterize AREs based on sequence information. Three statistical methods were explored to strengthen the prediction of putative AREs in the human genome. Cross-validation results indicated that all of the three models provided good sensitivity and specificity in identifying AREs, with an accuracy of at least 80%. It is the first time that HMM, SVM and RF have all been applied to constructing ARE prediction models.

Introduction

Computational methods for modeling and identifying DNA regulatory elements have been developed over the past two and a half decades [1]. Predicting transcription factor binding sites (TFBS) remains important in that not only it can detect rarely expressed genes but also analyze genes with unknown function. TFBSs are usually short, between 5-15bp in length. Potential binding sites can occur very frequently by chance in large genomes such as the human genome. Hence, it can be difficult to predict TFBS using simple sequence searching tools like BLAST. However, many other methods have been developed to predict TFBSs for a specific transcription factor using a collection of known binding sites from available resources [2]. Our proposed study in this paper is relevant to the latter approach.

Androgen receptor (AR) belongs to the steroid hormone receptor superfamily of ligand-inducible transcription factors. It plays key roles in male sexual differentiation and pubertal sexual maturation, and is essential for the maintenance of male reproductive function and behavior through mediating the effect of androgens [3]. AR contains distinct ligand binding, DNA-binding, and NH_2-terminal domains, responsible for transcriptional activation, subcellular localization, and dimerization. Upon ligand binding, it undergoes a conformational change resulting in its binding to androgen response elements (ARE) in the promoter region of its controlled genes. The DNA-receptor complex then exerts, in most cases a positive, or negative effect on gene transcription by recruiting either coactivators or corepressors to their target genes [4]. The sequence of ARE is highly conserved and is composed of two hexanucleotide half-sites separated by a three-nucleotide spacer. Identification and characterization of consensus AREs is of importance for understanding the mechanisms of androgen-specific gene transcription and AR-associated diseases, such as prostate cancer.

Biologists study gene regulation and protein-DNA interaction based on laboratory experiments such as chromatin immunoprecipitation and electrophoretic mobility shift assays. These wet-lab experiments allow the identification of putative AREs; however, they are not only time consuming and labor intensive, but also costly. On the other hand, computational methods can be used to identify TFBS, albeit at the expense of accuracy with the benefit of being easier to conduct with available resources [5]. Some state-of-the-art statistical methods are employed to characterize the binding preferences of transcription factors, and to identify their putative sites in genomic sequences. For instance, Hidden Markov Model (HMM) [6], Support Vector Machine (SVM) [7] and Random Forest (RF) [8] have been found to be effective and robust for classification. In this paper we utilized these three statistical approaches with the goal of building models to identify AREs. The classifiers were developed based on the biological sequence information acquired from the genome browser database. The models were then

0-7695-3019-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDMW.2007.81

trained to recognize sequence features and conservation patterns that distinguished between known regulatory regions and nonfunctional sequences. Finally sensitivity and specificity were evaluated to analyze the performance of the classifiers.

Methods

Data collection

We gathered a collection of experimentally validated AR binding sites from the biomedical literature. The sites and the upstream sequences were manually assembled to ensure data consistency. We retained in the data only those genes confirmed by ENSEMBL database (www.ensembl.org). To avoid overfitting, the sites for the same factors in promoters from orthologous genes were discarded. A set of 40 putative AR binding sites was shown in Supplemental Table (www4.ncsu.edu/~lli7). All of the binding sites were from Human. For the negative AREs, we randomly selected upstream sequences of housekeeping genes whose sequences contain no known AREs.

Feature selection

Feature selection is an essential data processing step prior to applying machine learning methods. It is important to select features which are most relevant to AR binding site classification. The features used in this study include k-mers, CpG islands and conservation scores. CpG islands are common near transcription factor start sites, and may be associated with promoter regions [9]. Conservation scores show a measurement of evolution conservation in eight vertebrates, including mammals, based on a phylogenetic footprinting approach, which has been taken to identify regulatory elements in the noncoding portion of genomes [10]. Their numerical values for each gene were obtained through the UCSC genome browser (http://genome.ucsc.edu/). In addition to the specifically identified ARE sites, the DNA sequence flanking ARE is known to be essential for AR-mediated transcriptional activity [11]. Therefore, sequence-based features corresponding to all possible DNA sequence variants of a given length k (k-mer) in the promoter neighborhood were calculated. We extended each ARE motif sequence upstream 20bp and downstream 20bp, and collected a pool of 5-mers for each genes. To select the most discriminative k-mers specific to AR binding site, F-scores computed were similar to that of Chen *et.al* [12] to measure the discrimination of two classes. Given the training vector χ_k, k =1,..., 4^k, the number of positive and negative instances are $n_+ = n_- = 40$, and the F-score of the *i*-th feature is defined in the following equation.

$$F(i) = \frac{(\overline{\chi}_i^{(+)} - \overline{\chi}_i)^2 + (\overline{\chi}_i^{(-)} - \overline{\chi}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+} (\chi_{k,i}^{(+)} - \overline{\chi}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-} (\chi_{k,i}^{(-)} - \overline{\chi}_i^{(-)})^2}$$

where $\overline{\chi}_i^{(+)}, \overline{\chi}_i^{(-)}, \overline{\chi}_i$ are the average of the *i*-th feature of the positive, negative, and whole data sets, respectively. $\chi_{k,i}^{(+)}$ is the *i*-th feature of the *k*-th positive instance, and $\chi_{k,i}^{(-)}$ is the *i*-th feature of the *k*-th negative instance. The retrieval of these features was done with Perl and R scripts.

Construction of Hidden Markov Model

A HMM consists of a set of states, where each state can emit symbols (nucleotides) based on a probability distribution. The emission probability for a certain nucleotide is specific for each state. States are connected in a chain-like structure, where the probability of moving from one state to another is termed transition probability [13]. The architecture of the HMM model for ARE prediction was based on the biological characteristics of the AR binding sites and the training data (Fig. 1). From the background state B, it is possible to move to the match state, of which it consists of two half-site models separated by a spacer state. These interacting match states are able to emit all nucleotide symbols with probabilities according to their fitted parameters. However, a match state may be connected to a mismatch state, allowing for the possibility of sequence variability in the model. Transition events are restricted to match, mismatch, and spacer states. The last non-interacting match state demands a transition to the end state. Only the positive ARE training sequences were aligned to construct positive HMM model through the HMMER package. The negative HMM model was implemented by setting up the probability of all four nucleotides as 0.25. Given a test sequence categorized by comparing the score produced by the positive HMM to that produced by the negative HMM, the classifier reported a bit-score and an E-value for every testing sequence.

Support Vector Machine

Based on statistical learning theory [14], the central idea of the SVM classification is to find a decision surface that has a maximum distance (margin) from the nearest training data points. The implementation of the SVM approach is as follows. Consider a classification problem with training dataset pairs $\{(\chi_i, y_i)\}$ where $\chi_i \in \Re^d$, and $y_i \in \{+1,-1\}$ denotes two ARE classes: $y_i = +1$ indicates sample i being positive ARE and $y_i = -1$ indicates negative ARE. A linear separating hyperplane generated by the SVM is given by $\{\chi : f(\chi) = \chi^T \beta + \beta_0 = 0\}$ where β is a unit vector: $\|\beta\| = 1$. A classification rule induced by $f(\chi)$ is $sign[y_i(\chi^T \beta + \beta_0)]$, which gives the signed distance from a data point χ to the bounding plane. It is not difficult to find a function $f(\chi) = \chi^T \beta + \beta_0$ with $y_i f(\chi_i) > 0$ to separate two classes [15]. Support vectors are data points that lie on the margins (Fig. 2). In short, SVM classification is used to achieve maximum separation between two classes. Finally, the SVM classifier was trained based on the conservation score, CpG island score, and top-ranking 5-mer features. The R-based SVM package allows us to construct ARE-SVM classifiers in R modules.

Random Forest

A random forest is a collection of identically distributed trees. The algorithm works in the following way [16]. First, for each iteration in RF, an n_{tree} bootstrap sample is randomly drawn from the training data. Secondly, a classification tree is induced from the bootstrap sample to maximize size without pruning. At each node, instead of searching through all the variables for the optimal split, a tree only searches through a random sample m_{try} of the variables. Finally, the two steps are repeated for a number of times before the predictions are made by majority vote of the trees for classification. The performance of RF depends on the correlation between trees and the strength of each individual tree. The idea is to maintain the strength of the trees while reducing their correlation with each other. The simulation and analyses were carried out with R, using package randomForest.

Performance Assessing

An important step in classification is to assess the accuracy of the model performance in a statistically significant way. The procedure of k-fold cross-validation involved partitioning the training data into k disjoint subsets of approximately equal size. One of the k subsets was used as the test set and the remaining k-1 subsets were put together to form a training set. Then, the average sensitivity and specificity across all k trials were computed. The results of SVM- and RF-based predictions were compared with those achieved with the HMM model.

Results

Model Performance

The HMM model was validated using the same training data set and tested in a 10-fold cross-validation. Forty known positive and forty negative sequences were applied for testing the prediction performance of the HMM models. The sensitivity and

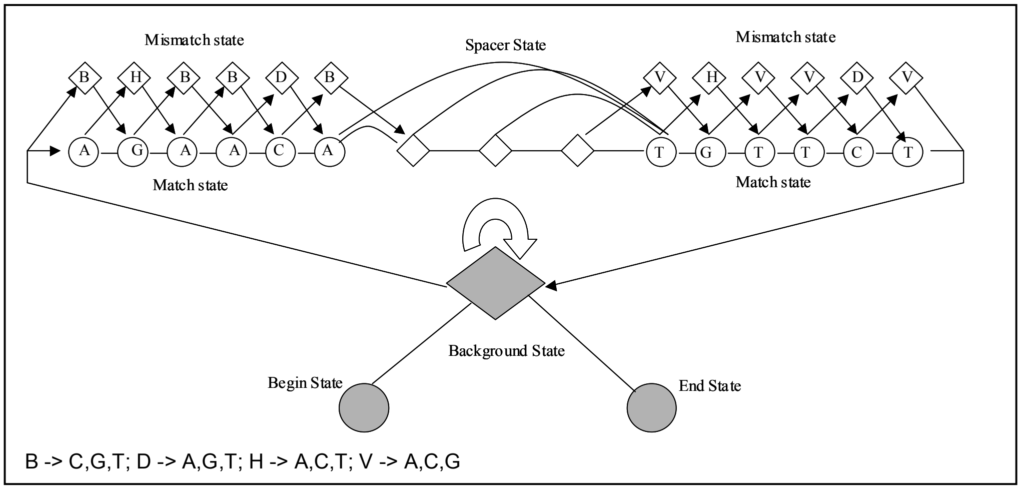

Fig.1. Graphical representation of HMM framework for AR binding sites prediction. Arrows denote possible transitions between states. From the background state, two principal transitions are possible: remain in the background state or move to the match state chains of half-site. The chain of states consists of two half-site models, separated by a set of spacer states that can generate 0-3 different spacer configurations. B represents either C, G, or T; D represents either A,G, or T; H represents either A,C, or T; V represents either A,C, or G.

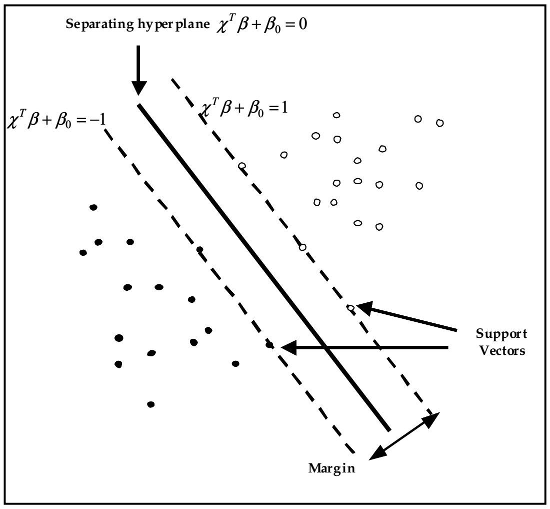

Fig.2. SVM prediction scheme. The training data features are sketched as dots in a two dimensional feature space, which are classified as +1 (open circle) and -1 otherwise (solid circle). The SVM classifier separates the classes by an optimal separating hyperplane ($\chi^T\beta+\beta_0=0$) with maximum margin from the hyperplane to the closest point. The best decision surface is determined by only small set of points called support vectors.

specificity were calculated for all possible cut-off levels (Fig.3A). Classification accuracy can be illustrated through Receiver Operator Characteristic (ROC) curves, constructed by plotting the sensitivity vs. false positive rate (1- specificity). The area under the ROC curve (AUC) has a number of desirable properties and is becoming increasingly popular as a performance measure: the higher the AUC, the lower the overall prediction error. We employed AUC for performance measurement of the HMM model and obtained an AUC of 0.75 (Fig.3B), indicating a good predictive performance.

The SVM model was trained using HMM scores, conservation scores, and k-mer features as training vectors. The radial basis function (RBF) kernel is selected to train the SVM, where the RBF kernel function is defined as: $K(\chi_i,\chi_j)=\exp(-\gamma\|\chi_i-\chi_j\|^2),\gamma>0$. The kernel function parameter γ and C, which control the complexity of the decision function versus the minimum training error, can be determined by running a two dimensional grid search. The grid-search is easily parallelized because each (C, γ) is independent, and the best performance of our SVM model was given by setting C to 8 and γ to 1 (Fig. 4A). Using the best combination of C and γ, the whole training set was trained to generate the final classifier. To evaluate the accuracy of SVM model, ten fold cross-validation was performed, resulting in the area under the ROC curve of 0.83 (Fig. 4B).

In RF, since each tree was constructed using a different bootstrap sample and about one-third of the cases were left out of the sample for estimating classification error, there was no need for cross-validation or a separate test set. The data not in the bootstrap sample, called out of bag data, was used to get a running unbiased estimate of the classification error as trees are added to the forest. The estimated error rate obtained from our RF model for ARE prediction was only about 0.15 (Data not shown).

Comparison of HMM, SVM and RF

In this study, three major statistical approaches, HMM, SVM and RF, have been built individually using the same training set to generate androgen receptor binding site profiles. The performance statistics of these models for ARE identification are illustrated in Table 1. Overall, all of the three approaches showed good performance; both sensitivity and specificity were at least 0.8. Using top ranked k-mer features alone, the SVM model only gave accuracy of 0.75, while the RF model could predict better with an accuracy of 0.88. When both the conservation score and HMM score were considered, the performance of the SVM and the RF models were improved giving better prediction than the HMM model. Additionally, we tested how many positive and negative AREs in the training dataset could be detected by these models. It is likely that the HMM approach provides the similar power for prediction, compared to the SVM and RF models.

Discussion

The availability of a large quantity of genome sequence data makes it possible to systematically analyze binding-site patterns in order to provide biological interpretations. In the present study, using the literature as a guideline, three statistical approaches were employed to build classifiers for predicting human AR target genes. The searches were based on specific models derived from experimentally verified AREs. All of the three computational approaches described the architecture of DNA regulatory elements and were capable of tapping into the tremendous amount of statistical information for classification. The HMM approach was constructed based on sequence context of positive and negative AREs, whereas the SVM encapsulated a significant amount of discriminatory information in the choice of its kernel to yield the optimal separating hyperplane, and the RF model collected tree predictors created by using

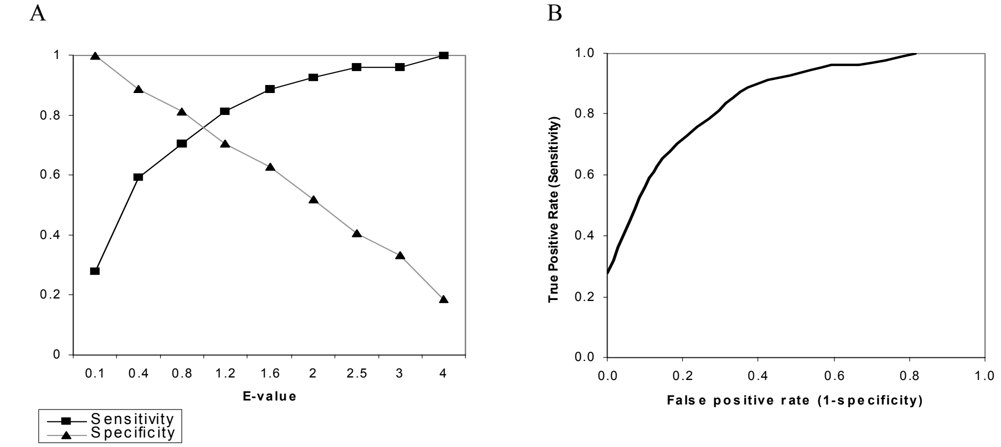

Fig.3. A. Plot of sensitivity, specificity vs. E-value based on 10-fold cross-validation. B. Receiver-operating-characteristic curves (ROC) for the prediction model. It shows the proportion of true positives selected by the HMM versus false positives. The performance is shown by the area under the ROC curve.

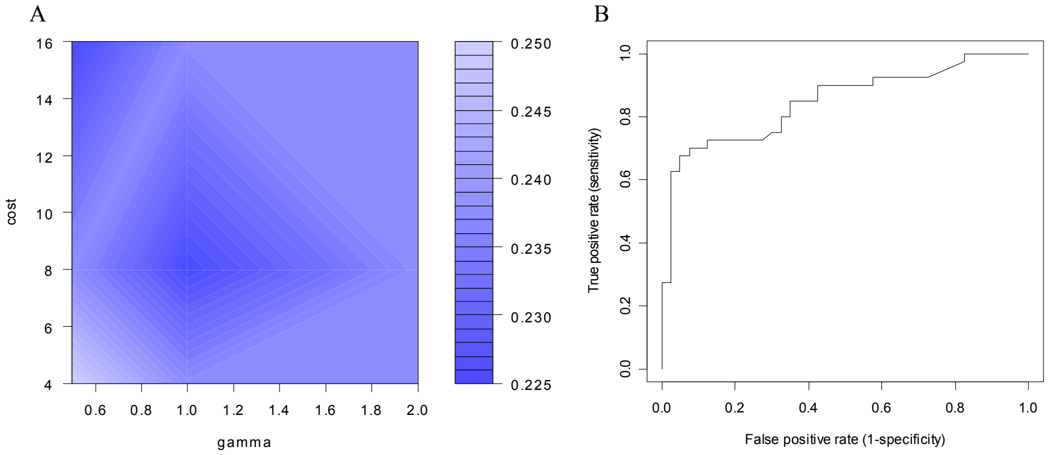

Fig.4. A. 10-fold cross validation classification rate of SVM trained using different parameter combinations, C= 4 to 16 and γ = 0.6 to 2.0. The right scale shows the misclassification error rate. The best parameter pair was obtained using C=8 and γ =1.0. B. ROC plot of SVM model for ARE prediction.

	HMM	SVM		Random Forests	
		k-mer	All features	k-mer	All features
Sensitivity	0.8	0.75	0.82	0.8	0.82
Specificity	0.8	0.88	0.83	0.9	0.9
Accuracy	0.8	0.76	0.81	0.85	0.86
Positive Prediction	35/40	34/40	35/40	32/40	33/40
Negative Prediction	37/40	29/40	38/40	36/40	36/40

Table 1. Summary of the performance for the HMM, SVM, and RF models.

bootstrap samples of the training data and random feature selection in tree induction. The results showed that all of the three methods provided similar performances with overall 80% accuracy. In contrast to the HMM model, the SVM and RF approaches are more encouraging in predicting AREs because both sensitivity and specificity seem favorable. Differences in feature selection between HMM, SVM and RF may contribute to discrepancies between the search methods. When SVM and RF map a high dimensional feature space, different information is taken into account, and more predictable accuracy is achieved. Even features ranked lowest still contain considerable information and are somewhat relevant. By removing features that are irrelevant and adding those that are closely correlated, the accuracy of the prediction model can be improved.

Although existing database TRANSFAC is widely useful for prediction of TFBSs, the position-specific weight matrix (PWM) for AREs in TRANSFAC was constructed based on only 7 known AR binding sites. Their ARE motif AGWACATNWTGTTCT could only be recognized when the testing sequence is a good match. On the contrary, our approaches are flexible; allowing the variation in the sequence, such as mismatches and gaps. The ability to detect interactions between AR and its response elements offers the opportunity to study the mechanisms of gene expression mediated by AR. The conserved ARE sequence motifs could possibly be involved in the tissue specificity of transcription by mediating either a positive or negative effect on the basal promoter activity [17]. However, it remains to be elucidated whether the ARE sequence is the major molecular determinant of AR specificity. Identifying direct targets of AR may allow the detection of critical changes in AR-related disease progression [18]. For instance, prostate specific antigen (KLK3) is an indicator for the detection of prostate cancer and for monitoring disease progression [19]. Therefore, identifying specific AR-regulated genes may contribute to the systematical elucidation of the gene regulatory network mediated by androgen, which will be pivotal for the development of disease treatments. Our prediction models will be a valued resource for researchers attempting to identify genes directly regulated by AR transcription factors.

In conclusion, we have developed bioinformatics-based methods for the prediction of candidate AREs in the human genome. AR binding site analyses by the SVM, HMM and RF methods allow for quick and accurate prediction of ARE position in the unknown sequences. Identification of these novel AREs may ultimately lead to the development of better therapies for the treatment or prevention of androgen-related diseases. This study is the first in incorporating AR promoter context in a HMM, SVM and RF topology.

References

[1].GuhaThakurta, D., *Computational identification of transcriptional regulatory elements in DNA sequence.* Nucleic Acids Res, 2006. 34(12): 3585-98.

[2].Osada, R.,*et al., Comparative analysis of methods for representing and searching for transcription factor binding sites.* Bioinformatics, 2004. 20(18): 3516-25.

[3].Heinlein, CA., *et al., Androgen receptor (AR) coregulators: an overview.* Endocr Rev, 2002. 23: 175-200.

[4].Chang, CY., *et al. Development of peptide antagonists for the androgen receptor using combinatorial peptide phage display.* Mol Endocrinol, 2005. 19(10): 2478-90.

[5].Sinha, S., *et al., A probabilistic method to detect regulatory modules.* Bioinformatics, 2003. 19 Suppl 1: 292-301.

[6].Abnizova, I., *et al. Transcription binding site prediction using Markov models.* J Bioinform Comput Biol, 2006. 4(2): 425-41.

[7].Holloway, D.T., *et al., Integrating genomic data to predict transcription factor binding.* Genome Inform, 2005. 16(1): 83-94.

[8].Diaz-Uriarte, R., *et al., Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, 2006. 7: 3.

[9].Hannenhalli, S., Levy, S., *Promoter prediction in the human genome.* Bioinformatics, 2001. 17 Suppl 1: S90-6.

[10].Siepel, A., Haussler, David, *Phylogenetic Hidden Markov Models.* Statistical Methods in Molecular Evolution, 2005: 325-351.

[11].Nelson, CC., *et al., Determinants of DNA sequence specificity of the androgen, progesterone, and glucocorticoid receptors: evidence for differential steroid receptor response elements.* Mol Endocrinol, 1999. 13(12): 2090-107.

[12].Chen, Y.-W., and Lin,Chih-Jen., *combining svms with various feature selection strategies.* 2005.

[13].Eddy, S.R., *Hidden Markov models.* Curr Opin Struct Biol, 1996. 6(3): 361-5.

[14].Vapnik, V.N., *Statistical Learning Theory.* Wiley, New York, 1998.

[15].Hastie, T.T., *et al.The Elements of statistical learning: Data mining, inference, and prediction.* Springer Series in Statistics, 2001.

[16].Breiman, L., *Random Forest.* Machine Learning., 2001. 45: 5-32.

[17].Fabre, S., *et al., Identification of a functional androgen response element in the promoter of the gene for the androgen-regulated aldose reductase-like protein specific to the mouse vas deferens.* J Biol Chem, 1994. 269(8): 5857-64.

[18].Masuda, K., *et al., Androgen receptor binding sites identified by a GREF_GATA model.* J Mol Biol, 2005. 353(4): 763-71.

[19].Lai, J., *et al., PSA/KLK3 AREI promoter polymorphism alters androgen receptor binding and is associated with prostate cancer susceptibility.* Carcinogenesis. 2007 28(5): 1032-1039.