# A COMPARISON STUDY OF TRANSCRIPTION FACTOR– DNA BINDING MODELS

Sudha Narang
Lecturer
Department of Computer Science & Engineering
Maharaja Agrasen Institute of Technology, Rohini
Delhi, India

## Abstract

The comparison study is drawn between two widely used motif representations i.e. Positional Weight Matrices (PWM) and Consensus Sequences. In the case of motif finding, where the binding sites are not known *a priori* but the algorithm must search a large space of possible binding sites, the PWM model may be difficult to learn as the search space is very large even for the PWM of short length ($R^N$ for a PWM of length $N$, where $R$ is the space of real numbers between 0 to 1). Optimization methods used to search for the best PWM may converge to a local minimum. On the other hand the consensus sequence has a smaller search space ($15^N$ for a motif of length $N$) which is easier to search for the global optimum.

## Introduction

The control or regulation of gene expression governs how much quantity of a particular protein is produced in a cell. The regulation is primarily achieved by turning on or off the transcription of genes. Most protein coding genes are transcribed by RNA polymerase II. However eukaryotic RNA polymerase cannot initiate transcription on its own. It requires the assistance of other proteins called transcription factors for initiating transcription. Any protein that is needed for the initiation of transcription, but which is not itself a part of RNA polymerase, is called a transcription factor (TF). Thus TFs play an important role in regulating transcriptional initiation, and hence gene expression.

Many TFs bind to DNA at specific sites, from where they collaborate with RNA Polymerase and with other TFs. A TF recognizes its specific binding site on DNA by the nucleotide sequence or pattern. The exact nucleotide sequence that is recognized varies for different TFs. It is of fixed length, usually ranging between 5-20 bp. A noteworthy feature is the ambiguity of the binding sequence. A TF can bind to a number of similar looking sequences with different binding affinities. Some positions in the binding sequence are highly conserved. Base substitutions in these positions can reduce or completely eliminate the TF binding. Whereas some other positions in the binding sequence are relatively less conserved and can be mutated without affecting the binding affinity. This ambiguity is useful as it allows different degrees of interaction with the TF at different DNA sites according to the binding affinity, which in turn results in different expression levels of various genes regulated by the same TF.

The nucleotide preferences of a TF at different base positions are described by a ***motif***. A motif is a model that essentially captures the common features of the binding sequences of a TF. Many motif representations are available in the literature [references]. However, in practical usage two motif representations are frequently encountered:

i) Positional weight matrices (PWM) **[Stormo et al.]**, and,

ii) Consensus sequence **[Wasserman et al.].**

In this study, experiments on a large number of transcription factors have been performed to study which model can more effectively represent their binding preferences under various situations. Based on this study, some general conclusions are drawn concerning the suitability of these two models. The conclusions of this study are meaningful to any bioinformatics study concerning motif representation or motif finding.

## Method

This section describes the methods used in this study to learn and test the different motif representations. The PWM and consensus representations are described first. Then a description of the cross-validation experiments performed to estimate the goodness of these two models is given. The criteria used to evaluate the methods, *viz.* sensitivity or true positive rate (TPR), false positive rate (FPR) and receiver-operating characteristics (ROC), are also explained.

### *The PWM Representation*

The positional weight matrix (PWM) **[Stormo et al. (1982), Stormo (2000)]** is a numerical representation of the binding preferences of a TF. It records the base preferences at each position of the binding sequence of the TF. Let the binding sequence of the TF be of length *l*. The PWM for this TF is a matrix of dimensions $4\times l$ whose each cell $w_{b,j}$ records the relative preference or weight of the base $b \in A,C,G,T$ at the position $j \in 1,2,\ldots,l$ of the binding sequence. For instance consider the 15 bp long binding sequences of the transcription factor NF-Y shown in Figure 1(a). The frequency matrix $F=\left[ f_{b,j} \right]$ for these sequences is shown in Figure 1(b), where $f_{b,j}$ is the relative frequency of base $b$ at position $j$ in the sequences. The positional weight matrix $W=\left[ w_{b,j} \right]$ is shown in Figure 1(c). The weight $w_{b,j}$ is calculated as

$$w_{b,j} = \ln\left( f_{b,j} / p_b \right), \qquad (1)$$

where $p_b$ is the background frequency of the base *b* in the genome[1].

[1] In Figure 1(c), the weights $w_{b,j}$ are calculated as $\ln\left( \left( f_{b,j} + 0.1 \right) / 0.25 \right)$ where 0.1 is a pseudocount and $p_b$ is assumed to be 0.25 for all $b \in \{A, C, G, T\}$ considering equal proportion of the bases A,C,G and T in the genome.

| | Position → | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Seq 1 | C | T | T | G | G | C | C | A | A | T | C | A | G | A | A |
| Seq 2 | T | T | C | A | G | C | C | A | A | T | C | G | G | A | G |
| Seq 3 | C | G | C | G | G | C | C | A | A | T | C | A | G | C | G |
| Seq 4 | T | T | T | A | G | C | C | A | A | T | C | A | G | C | T |
| Seq 5 | C | C | T | G | G | C | C | A | A | T | C | A | G | C | G |
| Seq 6 | C | C | C | G | G | C | C | A | A | T | C | A | G | C | G |
| Seq 7 | G | T | T | A | G | C | C | A | A | T | C | A | G | C | A |
| Seq 8 | A | T | C | A | G | C | C | A | A | T | G | A | G | C | T |
| Seq 9 | C | C | C | A | G | C | C | A | A | T | C | A | G | A | G |
| Seq 10 | C | T | C | A | G | C | C | A | A | T | G | G | G | C | G |

(a)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.1 | 0 | 0 | 0.6 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0.8 | 0 | 0.3 | 0.2 |
| C | 0.6 | 0.3 | 0.6 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0.8 | 0 | 0 | 0.7 | 0 |
| G | 0.1 | 0.1 | 0 | 0.4 | 1 | 0 | 0 | 0 | 0 | 0 | 0.2 | 0.2 | 1 | 0 | 0.6 |
| T | 0.2 | 0.6 | 0.4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0.2 |

(b)

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | -0.32 | -1.01 | -1.01 | 0.93 | -1.01 | -1.01 | -1.01 | 1.39 | 1.39 | -1.01 | -1.01 | 1.19 | -1.01 | 0.37 | 0.09 |
| C | 0.93 | 0.37 | 0.93 | -1.01 | -1.01 | 1.39 | 1.39 | -1.01 | -1.01 | -1.01 | 1.19 | -1.01 | -1.01 | 1.07 | -1.01 |
| G | -0.32 | -0.32 | -1.01 | 0.60 | 1.39 | -1.01 | -1.01 | -1.01 | -1.01 | -1.01 | 0.09 | 0.09 | 1.39 | -1.01 | 0.93 |
| T | 0.09 | 0.93 | 0.60 | -1.01 | -1.01 | -1.01 | -1.01 | -1.01 | -1.01 | 1.39 | -1.01 | -1.01 | -1.01 | -1.01 | 0.09 |

(c)

**Figure 1: A small sample of binding sites for the transcription factor NF-Y.**

Given a new sequence, the PWM can be used to evaluate the binding affinity of the TF to this sequence. The binding affinity is represented by the *matrix score*. The matrix score of a sequence *S* of length *l* is calculated as

$$Matrix\ score = \frac{\sum_{j=1}^{l}\left(w_{S_j,j} - w_{\min_j,j}\right)}{\sum_{j=1}^{l}\left(w_{\max_j,j} - w_{\min_j,j}\right)},$$

where $W=\left[w_{b,j}\right]$ is the PWM of the TF, $S_j \in \{A,C,G,T\}$ is the base at position *j* in the sequence *S*, $\min_j$ is the base which has the minimum weight at position *j* of the PWM, and $\max_j$ is the base which has the maximum weight at position *j* of the PWM. The matrix score is a real number within the range [0,1]. The process of calculating the match score of a sequence *S* against a given PWM *W* is usually referred to as *matching* the sequence *S* with *W*.

If the matrix score for the sequence *S* exceeds a certain threshold *t*, the sequence is said to be a valid binding sequence of the TF. The threshold *t* is calculated based on p-value which states the likelihood of obtaining a score higher than *t* by chance. The p-value is computed with the help of a background sequence set. The background sequence set represents the average composition of the genome. It could consist of sequences chosen randomly from the genome, or sequences generated by a Markov model trained on the genome [reference for RSA tools]. By matching a large number of

background sequences (~1,000,000 sequences) to the PWM, the chance distribution of the match score is obtained. The nature of the match score distribution varies according to the PWM, but in general it resembles the distribution shown in Figure 2. The area under the distribution beyond *t* gives the chance probability of obtaining a match score higher than *t*. This chance probability is the p-value of the score *t*. Or stated simply,

$$P-value(t)=\frac{\text{No. of background sequences with match score } \geq t}{\text{Total number of background sequences}}$$

We choose the threshold *t* for the PWM corresponding to a standard p-value cutoff of 0.001 or 0.0001.

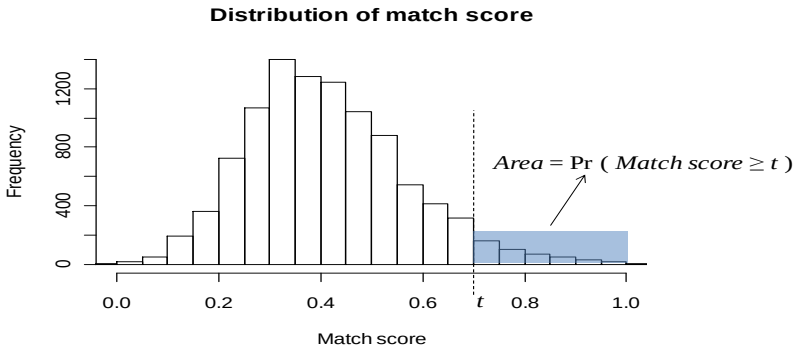


**Figure 2: Match score distribution**

***The Consensus Representation***

The consensus sequence model only specifies the set of valid binding sequences of a TF using a regular expression or other notations. It does not provide any numerical measure of the binding affinity. In other words, it only specifies which sequences are bound and which sequences are not bound by the TF. Two consensus notations are popular. The first notation uses the IUPAC nomenclature of single letter codes (Figure 3) to represent the allowed bases at any particular position in a binding sequence. The second notation states the most common (or highest affinity) binding sequence of the TF and specifies the maximum number of base substitutions that are allowed in the binding sequence **[Pevzner and Sze (2000)].** This notation assumes that all positions are equally open to base substitution. In this study, we have used the first notation which uses IUPAC nomenclature, as it is closer to the biological understanding.

There are several ways of generating an IUPAC consensus sequence from the data of known binding sites of a TF **[Day and McMorris (1992)]**. For instance in one of the approaches a base is considered significant at a position if it occurs in any one of the binding sites. In another approach a base is considered significant at a position only if it occurs in more than 25% of the binding sites **[Daniels and Deininger (1991)]**. The former approach is more inclusive while latter approach is more accurate. We have chosen the latter approach in this study due to its better accuracy. An illustration of generating the IUPAC consensus sequence in this manner corresponding to the binding site data of Figure 1(a) is shown in Figure 3.

Given a new sequence, the consensus model can determine by direct comparison with the consensus sequence whether it is a valid binding sites for the TF.

| | Position → | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Seq 1 | C | T | T | G | G | C | C | A | A | T | C | A | G | A | A |
| Seq 2 | T | T | C | A | G | C | C | A | A | T | C | G | G | A | G |
| Seq 3 | C | G | C | G | G | C | C | A | A | T | C | A | G | C | G |
| Seq 4 | T | T | T | A | G | C | C | A | A | T | C | A | G | C | T |
| Seq 5 | C | C | T | G | G | C | C | A | A | T | C | A | G | C | G |
| Seq 6 | C | C | C | G | G | C | C | A | A | T | C | A | G | C | G |
| Seq 7 | G | T | T | A | G | C | C | A | A | T | C | A | G | C | A |
| Seq 8 | A | T | C | A | G | C | C | A | A | T | G | A | G | C | T |
| Seq 9 | C | C | C | A | G | C | C | A | A | T | C | A | G | A | G |
| Seq 10 | C | T | C | A | G | C | C | A | A | T | G | G | G | C | G |
| Consensus | C | Y | Y | R | G | C | C | A | A | T | C | A | G | M | G |

**(a)**

| Symbol | A | C | G | T | R | Y | M | K |
|---|---|---|---|---|---|---|---|---|
| Meaning | A | C | G | T | A/G | C/T | A/C | G/T |
| Symbol | S | W | H | B | V | D | N | |
| Meaning | G/C | A/T | A/C/T | G/C/T | A/C/G | A/G/T | A/C/G/T | |

**(b)**

**Figure 3: (a) Learning the consensus sequence from a collection of TF binding sequences, (b) single-letter IUPAC codes for representing degeneracy of nucleotides in a consensus sequence**

***Cross validation experiments***

In this study, the performance of a motif model (PWM or consensus) has been evaluated by performing cross-validation experiments. The data for this experiment is a set of experimentally validated TFBSs for a particular TF obtained from TRANSFAC or JASPAR databases. The TFBS sequences must be of the same length and aligned with respect to each other.

In the *K*-fold cross validation procedure, the total set of TFBSs is partitioned into *K* equal parts. The partitioning is performed randomly. Then *K* iterations of training and testing are performed as follows.

In the first iteration, the sequences in parts 1,2,...,(*K*-1) are together used for learning the motif model. Then the learnt model is tested on the sequences in part *K*. During testing, the model is used to classify each sequence in part *K* as TFBS or not. If the model is 100% accurate, it must classify all sequences in part *K* as TFBSs. However due to the modelling error, some sequences are not classified as TFBSs. The *true positive rate* (TPR) or *sensitivity* of the model is then computed as

*TPR = No. of sequences classified as TFBSs* / *Total no. of sequences tested*

Simultaneously, the motif model is also tested on a set of background sequences. The background sequences are supposed to not contain matches of the motif. However, we may still find matches of the PWM in the background by random chance. The *false positive rate* (*FPR*) of the model is computed as

*FPR = No. of background sequences classified as TFBSs* / *Total no. of background sequences tested*

In the second iteration, the parts 1,2,…,(*K*-2),*K* are together used for learning the motif model, whereas the part (*K*-1) is used to test the model. Again the *TPR* and *FPR* are computed. Similarly, in the $n^{th}$ iteration, the parts 1,2,…,(*K*-*n*),(*K*-*n*+2),…,*K* are used for learning the motif model, whereas the part (*K*-n+1) is used for testing the model.

After *K* such iterations, one cross validation experiment is completed. During the course of cross-validation, the entire data has been used as test set exactly once. Therefore it provides an unbiased estimate of the model's performance, i.e., the performance is not biased by the manner in which the data is partitioned into training and test sets. The *TPR* and *FPR* for the cross validation experiment is the average of the *TPR* and *FPR* values obtained over all *K* iterations.

A special case of K-fold cross-validation is the leave-one-out cross validation (LOOCV). In LOOCV, a single observation is used as the test data, and the remaining observations as the training data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sample. The iterations of cross-validation ensure that each observation is used once as the test data. Leave-one-out cross-validation is computationally expensive, however it is possible in this problem as the sequence data in this study is limited.

***Receiver-operating characteristics***

A cross-validation experiment gives an unbiased estimate of the *TPR* and *FPR* of TFBS detection by the motif model. The TPR gives an estimate of how easily the model can detect true matches, whereas the FPR gives an estimate of how easily the model reports false matches. Ideally one would like 100% TPR and 0% FPR. However, practically TPR is lower that 100% and FPR is higher that 0%, and the two are related. For example, in the case of a PWM, if the match score threshold is kept low, both TPR and FPR will be high. On the other hand, if the match score threshold is kept high, both TPR and FPR will be low. The receiver-operating characteristic (ROC) describes the relationship between TPR and FPR as the model parameters (such as PWM match score threshold) are varied. The variation is illustrated in Figure 4. The perfect model yields a point in the upper left corner (coordinate (0,1)) of the ROC space. Whereas the ROC curve of a completely random model is the 45º diagonal line. For a better-than-random model, the ROC curve lies somewhere above the 45º diagonal, and the further away this curve is from the diagonal the better the predictor's performance. Thus, the area under the ROC curve is an indicator of the model's performance. In this paper, the performance of the motif model is studied using the ROC curve and the area under the ROC curve.

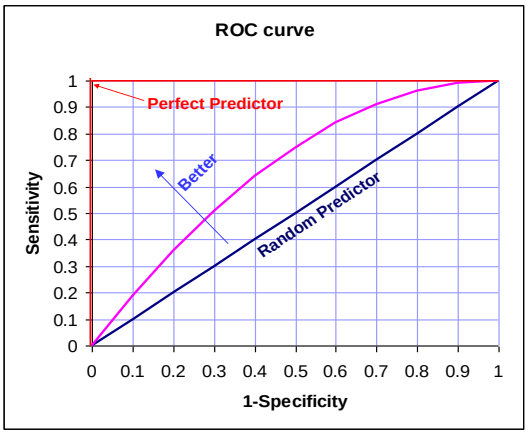


**Figure 4: The Receiver Operating Characteristics (ROC) curve.**

**Data**

The TFs of several different species from yeast to human are collected from JASPAR http://jaspar.cgb.ki.se , an open access database for eukaryotic transcription factor binding sites. In total around 40 TFs are studied for different species. The known sets of binding sites for these TFs are available in the JASPAR database. Only TFs with a minimum of 10 binding sites were selected for the evaluation.

To model the performance (FPR), a negative (background) set of sequences of the same species are also required. These could be modelled by selecting random sequences from the genome. However a better way is to use random sequences from the Regulatory Sequence Analysis Tool (RSAT) server (refer http://rsat.ulb.ac.be/rsat/random-seq_form.cgi). RSAT holds pre-computed background for the whole genomes of several different species in the form of Markov models of orders 1-8. The random sequences generated by the server using these Markov models are thus representative of the complete genome on average. 4000 random sequences were collected for each species as the negative background set in this study.

**Results**

The performance of both PWM and Consensus is evaluated by plotting the ROC curves of the data collected from JASPAR. Some significant cases are shown below (here red color signifies PWM and green color signifies Consensus):

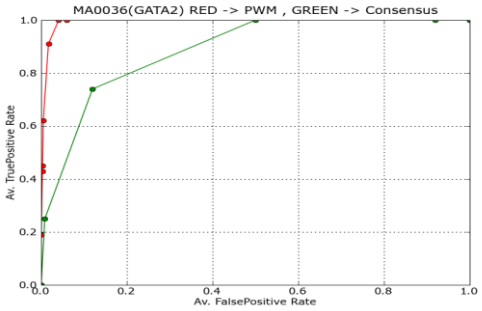


**Figure 5(a)**

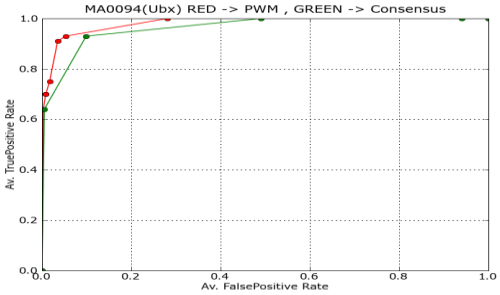


**Figure 5(b)**

ROC curve for TF GATA2 ( homosapien ) is shown in Figure5(a) and for Ubx (Drosophila ) in Fig5(b) and in both cases it is observed that PWM performed better.

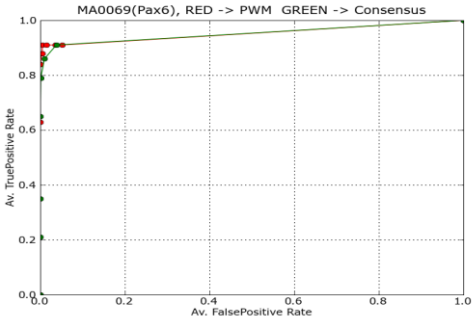


**Figure 5(c)**

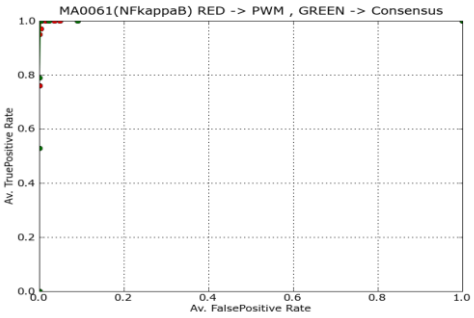


**Figure 5(d)**

ROC curve for TF Pax6 (homosapien) is shown in Figure5(c) and for NFkappaB ( homosapien ) in Figure5(d) and in both cases similar performance is observed.

The area under the ROC is observed to be high for both models. The dependency of ROC area on the motif length and the total number of binding sites in both models can be shown by the following plots.

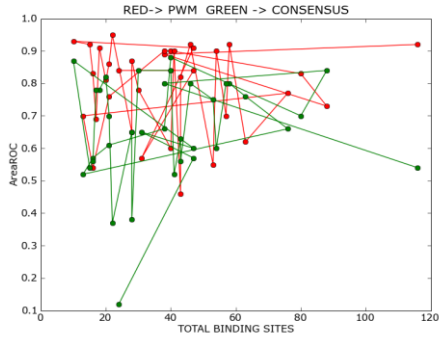


**Figure 5(e)**

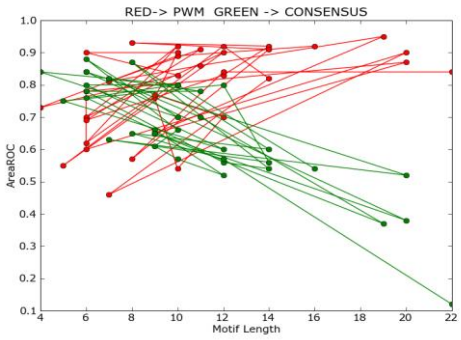


**Figure 5(f)**

Plot in Figure5(e) is drawn between Area under ROC curve obtained from various TF's studied for PWM and consensus as Y-axis and corresponding motif total number of binding sites as X-axis and Plot in Figure5(f) has a change of X-axis as Length of the corresponding motif. It is observed from the scatter plots that both models have performed better but the consensus performance decreased when the difference between length of the motif and total number of binding sites is very small. e.g. in case of TF Androgen having 22 length motif with total 24 binding sites i.e. with difference of 2 , the area under ROC for consensus is found to be 0.18 while for PWM it is 0.84.

## CONCLUSION

Both PWM and consensus sequences, are used for representing motifs. Mostly the Transcription factors with large number of binding sites can be represented better by the PWM while the Transcription factors with few and small length binding sites can be represented well by the consensus sequence. For Transcription Factors with large number or small number of binding sites, it is observed that as the no. of mismatches allowed increases the number of matching binding sites found also increases. It is an open question which of the two representations should be used in different situations or applications.

***References:***

1. Gary D.Stormo (2000) DNA Binding Sites: Representation and Discovery(vol 16,16-23)
2. Patrick D'haeseleer : What are DNA sequence motifs
3. Patrick D'haeseleer: How does DNA sequence motif discovery work?
4. Kenzie D. Maclsaac, Ernest Fraenkel : Practical strategies for discovering Regulatory DNA sequence motifs.

5. Jean Michel Claverie and Stephane Audic(1996): The statistical significance of nucleotide PWM(Vol 2 no 5,431-439)

6. Rodger Staden(1984)- Computer methods to locate signals in nucleic acid sequences

7. Naum I. Gershenzon, Gary D. Stormo1 and Ilya P. Ioshikhes: Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites.( 2290–2301 Nucleic Acids Research, 2005, Vol. 33, No. 7)

8. Albin Sandelin, Wynand Alkema, PaÈ r EngstroÈm, Wyeth W. Wasserman1 and Boris Lenhard, JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research, 2004, Vol. 32, Database issue D91±D94

9. Qing K. Chen1, Gerald Z. Hertz and Gary D. Stormo MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices Vol. 11 no. 5 1999 Pages 563-566

10. Rodger Staden: Methods for calculating the probabilities of finding patterns in sequences Vol.5, no.2. 1989 Pages 89-96

11. Tatsuhiko Tsunoda and Toshihisa Takagi. Estimating Transcription factor Bindability on DNA. Vol 15 no's 7/8 1999 Pages 622-630

12. William H.E. Day and F.R. Mc Morris : A consensus program for molecular sequences(1992, Vol 9 Pages 653-656)

13. Regulatory Sequence Analysis Tool(RSAT)