Enhanced Matrix Model for Finding Sequence Motif

Pankaj Agarwal[1] & Laxmikant Vishwamitra[2]

[1]Professor & Head, Department of Computer Science, IMS Engineering College, Ghaziabad, U.P
Pankaj_agwl@rediffmail.com

[2]Asst. Professor, MCA Department, Radha Govind Engineering College, Meerut, U.P

Abstract

This paper presents a probabilistic approach to locate motifs in a given set of molecular sequences. Motif is a short sequence of DNA or RNA (or amino acids) which often consists of 5- 16 nucleotides. The methodology presented here is based on probabilistic approach (Likely Hood Ratio) with three wild card locations to express motif locations and their poor, moderate and good results on the basis of matches found is observed. The proposed method allows three displacement locations (i.e. 1, 9, & 11) unlike other motif finding techniques that allow only two displacement locations.

I. INTRODUCTION & BACKGROUND

The two basic methods for computational gene prediction are S*equence similarity search* and *Ab initio*. The sequence similarity search is based on finding similarity in gene sequences between ESTs (Expressed Sequence Tags) or other genomes and functional regions are more conserved evolutionarily than nonfunctional regions. While, ab initio uses gene structure as a template to detect genes and relies on two types of sequence information: signal sensors and content sensors [1]. A combinatorial optimization framework for motif finding that is flexible enough to model several variants of the problem and is not limited by the motif length [3]. The different approaches that build models to distinguish binding targets of a TF (i.e. positive sequences) from non-targets of a TF (i.e. negative sequences) [4]. The threshold for positional weight matrix compares favourably with broadly used Match method. It uses four digits [0, 1, 2, 3] of quaternary system for four nucleotides [A, C, G, T] and applied the random sampling scheme to perform the computation and sample size [5]. The versatile combinatorial optimization framework for motif finding that couples graph pruning techniques with a novel integer linear programming formulation. It combines graph theoretic and mathematical programming approach. It includes incorporating substitution matrices and phylogenetic distances [6]. The Boolean Matrices as Model for Motif Kernels [7] introduces BMA (Boolean Matrix Algorithm) which is more realistic than data models based on consensus strings and Hamming distance. Algorithm works on signal-noise ratio and selects the set of motif candidates by the minimal score of all its members. A weight matrix for a pattern of length n is defined as a matrix of numbers W*i*,x where *i* is in {1,2,...,n} and x is in {A,T,G,C} for DNA. The scores of the string x*1* ... x*n* is given by: W*1*,x*1* + W*2*,x*2* + ... + W*n*,x*n*. [2,8].

Matrix Models

In matrix, numbers containing scores for each residue or nucleotide at each position of a fixed-length motif.

There are two types of weight matrices.

A position frequency matrix (PFM) records the position-dependent frequency of each residue or nucleotide. PFMs can be experimentally determined from SELEX experiments or computationally discovered by tool such as MEME using hidden Markov models. While, position weight matrix (PWM) contains log odd weights for computing match score. A cut-off is needed to specify whether an input sequence matches the motif or not. PWMs are calculated from PFMs.

Position Weight Matrix Model:
Site 1 A G A T G G A T G G
Site 2 T G A T T G A T G T
Site 3 T G A T G G A T G G
Site 4 A G A T T G A T C G
Site 5 T G A T G G A T T G
Site 6 T G A T G G A T T G

Site 7 A G A T G G A T T G

IUPAC (International Union of Pure Applied Chemistry) consensus:
W G A T G G A T N G (where W = A or T)

PWM represents frequencies of each base at each position in the motif

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

Information content IC

The least variable positions likely are important for specifying the protein-DNA interaction. Therefore high information content = low sequence variation at that position.

Information Content at position i:

$$ICi = 2 + \sum_{b=G,A,T,C} P_b(i) * \log2(P_b(i))$$

Where $P_b(i)$ is the probability of base $_b$ at position i

(If using $\log_2$, the information is in 'bits')

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

IC	1.0	2.0	2.0	2.0	1.1	2.0	2.0	2.0	0.5	1.3	15.9 bit

Maximum IC if P of some base is 1.0: = 2 + [(1.0 * 0) + 0 + 0 + 0] = 2

Minimum IC if P is 0.25 for all bases: =2 + [0.25*(-2)]*4 = 0

Is the sequence A G A T T G A T C T a match to this matrix?

Joint probability: assuming each position is independent,

$$P(\text{motif}) = \prod_{\substack{b=G,A,T,C \\ i}} P_b(i)$$

Background model: $P(G,A,T,C) = 0.25$

P(sequence | matrix model) = (0.4)(1.0)(1.0)(1.0)(0.3)(1.0)(1.0)(1.0)(0.2)(0.2) = 0.0048

P(sequence | background model) = (0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)(0.25)=6.8e-24

Log-likelihood ratio LLR= $\log_2$ (P(sequence | matrix model) / P(sequence | background model))

A measure of how different the likelihood of the sequence is, given the motif model vs. the background model.

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

Is the sequence **A A A T T G A T C T** a match to this matrix?

Joint probability: assuming each position is independent,

$$P(\text{motif}) = \prod_{\substack{b=G,A,T,C \\ i}} P_b(i)$$

P(sequence | matrix model)=(0.4)(0)(1.0)(1.0)(0.3)(1.0)(1.0)(1.0)(0.2)(0.2)=**0**

If PWM was trained on a small sample set, it might have missed some examples = *over fitting* of the matrix (ie. *too* specific)

II. PROPOSED METHODOLOGY

Formula:

The Position Weight Matrix represents frequencies of each base at each position in the motif.
TIBi (Total Information Bit at position i)

$$TIB_i = 3 + \sum_{b=GATC} P_b(i) * \log_2(P_b(i)) \qquad \ldots(1)$$

Joint Probability: assuming each position is independent,

$$P_{(motif)} = \prod_{b=G,A,T,C} P_b(i) \qquad \ldots(2)$$

Where $P_b(i)$ is the probability of base b at position i.
The total information is on bits only when $\log_2$ is used. [9]

Test Data I

Table: 1 Representation of Motifs (Position-Specific Weight Matrices)

	1	2	3	4	5	6	7	8	9	10	11
Site 1	A	G	A	T	G	G	A	T	G	G	C
Site 2	T	G	A	T	T	G	A	T	G	T	C
Site 3	T	G	A	T	G	G	A	T	G	G	A
Site 4	A	G	A	T	T	G	A	T	C	G	T
Site 5	T	G	A	T	G	G	A	T	T	G	T
Site 6	T	G	A	T	G	G	A	T	T	G	T
Site 7	A	G	A	T	G	G	A	T	T	G	A
	W	**G**	**A**	**T**	**G**	**G**	**A**	**T**	**N**	**G**	**X**

The above table is the combination of nucleotides. Where W (A or T), N (C or G or T) and X (C or A or T) are Wild Card locations. In the above any six locations can be fixed on the basis of the sequence of motif for which Likely Hood Ratio is to be calculated.

Table: 2 No. of occurrences of each nucleotide

G	**0**	**7**	**0**	**0**	**5**	**7**	**0**	**0**	**4**	**6**	**0**
A	**3**	**0**	**7**	**0**	**0**	**0**	**7**	**0**	**0**	**0**	**2**
T	**4**	**0**	**0**	**7**	**2**	**0**	**0**	**7**	**3**	**1**	**3**
C	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**

The above table shows the column-wise number of occurrences of each nucleotide. Where 7 shows only one nucleotide is allowed at those locations. While, others are variable but column 5 & 10 allow only two types of nucleotides only. Here, T & G are considered.

Table:3 Finding Matches to (instances of) a PWM

G	0.00	1.00	0.00	0.00	0.71	1.00	0.00	0.00	0.57	0.86	0.00	
A	0.43	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.29	
T	0.57	0.00	0.00	1.00	0.29	0.00	0.00	1.00	0.43	0.14	0.43	
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	
*TIB=	2.01	3	3	3	2.13	3	3	3	2.01	2.41	1.44	28 Bits

The above table shows column-wise information bit and Total Information Bits (TIB). If each column is fixed (which is not possible) then the total information bit will be 33. Here loss of bits (33 – 28 = 5) i.e. the loss of 5 bits.

P(Sequence | Background Model) = Joint Probability (by formula (2)) = $(0.25)^{11} = 2.38*10^{-7}$
P(Sequence | Matrix Model), if we take Minimum Probability of occurrence (by formula (1))

LHR = (matrix model/Background Model) = 13.16
P(Sequence | Matrix Model), if we take Maximum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) = 18.16
Seq T G A T T G A T T T A
LHR=13.57

Further more experiments have been made on the basis of occurrences of nucleotide.

Test Data II

Table: 4 First Representations of Motifs (Position-Specific Weight Matrices)

	1	2	3	4	5	6	7	8	9	10	11
Site 1	A	G	A	T	G	G	A	T	G	G	C
Site 2	T	G	A	T	T	G	A	T	G	T	A
Site 3	T	G	A	T	G	G	A	T	G	G	A
Site 4	A	G	A	T	T	G	A	T	C	G	T
Site 5	T	G	A	T	G	G	A	T	T	G	C
Site 6	T	G	A	T	G	G	A	T	T	G	C
Site 7	A	G	A	T	G	G	A	T	T	G	A
	W	G	A	T	G	G	A	T	N	G	X

The above table is the combination of nucleotides. Where W (A or T), N (C or G or T) and X (C or A or T) are Wild Card locations. In the above any six locations can be fixed on the basis of the sequence of motif for which Likely Hood Ratio is to be calculated.

Table: 5 Finding Matches to (instances of) a PWM

G	0	7	0	0	5	7	0	0	4	6	0
A	3	0	7	0	0	0	7	0	0	0	3
T	4	0	0	7	2	0	0	7	3	1	1
C	0	0	0	0	0	0	0	0	0	0	3

The above table shows the column-wise number of occurrences of each nucleotide. Where 7 shows only one nucleotide is allowed at those locations. While, others are variable but column 5 & 10 allow only two types of nucleotides only. Here, T & G are considered.

Table: 6 No. of occurrences of each nucleotide

G	0.00	1.00	0.00	0.00	0.71	1.00	0.00	0.00	0.57	0.86	0.00	
A	0.43	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.43	
T	0.57	0.00	0.00	1.00	0.29	0.00	0.00	1.00	0.43	0.14	0.14	
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	
*TIB=	2.01	3	3	3	2.13	3	3	3	2.01	2.41	1.56	29.7 Bits

The above table shows column-wise information bit and Total Information Bits (TIB). If each column is fixed (which is not possible) then the total information bit will be 33. Here loss of bits (33 – 29.7 = 3.3) i.e. the loss of 3.3 bits.
P(Sequence | Background Model) = Joint Probability (by formula (2)) = $(0.25)^{11} = 2.38*10^{-7}$
P(Sequence | Matrix Model), if we take Minimum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) = 9.27
P(Sequence | Matrix Model), if we take Maximum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) = 17.24

Seq. G G A T G G A T C T C
LHR=15.43

Test Data III

Table: 7 First Representation of Motifs (Position-Specific Weight Matrices)

	1	2	3	4	5	6	7	8	9	10	11
Site 1	A	G	A	T	A	G	A	T	A	A	C
Site 2	A	G	A	T	C	G	A	T	C	A	G
Site 3	T	G	A	T	G	G	A	T	T	A	A
Site 4	T	G	A	T	T	G	A	T	A	A	G
Site 5	T	G	A	T	A	G	A	T	A	G	G
Site 6	G	G	A	T	C	G	A	T	T	G	T
Site 7	G	G	A	T	T	G	A	T	G	C	A
	W	G	A	T	G	G	A	T	N	G	X

The above table is the combination of nucleotides. Where W (A or T or G), N (A or C or T or G) and X (A or C or T or G) are Wild Card locations. In the above any six locations can be fixed on the basis of the sequence of motif for which Likely Hood Ratio is to be calculated.

Table: 8. No. of occurrences of each nucleotide

G	2	7	0	0	1	7	0	0	1	2	3	7
A	2	0	7	0	2	0	7	0	3	4	2	0
T	3	0	0	7	2	0	0	7	2	0	1	0
C	0	0	0	0	2	0	0	0	1	1	1	0

The above table shows the column-wise number of occurrences of each nucleotide. Where 7 shows only one nucleotide is allowed at those locations. While, others are variable but column 5 & 10 allow all four (G, A, T & C) & three (G, A, & C) nucleotides respectively.

Table: 9 Finding Matches to (instances of) a PWM

G	0.29	1	0	0	0.14	1	0	0	0.14	0.29	0.43	
A	0.29	0	1	0	0.29	0	1	0	0.43	0.57	0.29	
T	0.43	0	0	1	0.29	0	0	1	0.29	0	0.14	
C	0	0	0	0	0.29	0	0	0	0.14	0.14	0.14	
*TIB=	1.44	3	3	3	1.05	3	3	3	1.16	1.62	1.16	24.4 Bits

The above table shows column-wise information bit and Total Information Bits (TIB). If each column is fixed (which is not possible) then the total information bit will be 33. Here loss of bits (33 – 24.4 = 8.6) i.e. the loss of 8.6 bits.

P(Sequence | Background Model) = Joint Probability (by formula (2)) = $(0.25)^{11} = 2.38*10^{-7}$
P(Sequence | Matrix Model), if we take Minimum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) = 8.87
P(Sequence | Matrix Model), if we take Maximum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) =15.19
Given Sequence A G A T G G A T C G A
LHR= 10.97

Test Data IV

Table: 10 First Representations of Motifs (Position-Specific Weight Matrices)

	1	2	3	4	5	6	7	8	9	10	11
Site 1	A	G	A	T	A	G	A	T	A	A	C
Site 2	T	G	A	T	C	G	A	T	C	G	G
Site 3	T	G	A	T	G	G	A	T	T	G	T
Site 4	C	G	A	T	G	G	A	T	A	C	G
Site 5	T	G	A	T	A	G	A	T	A	A	G
Site 6	G	G	A	T	C	G	A	T	T	T	A
Site 7	G	G	A	T	T	G	A	T	G	C	A
	W	G	A	T	G	G	A	T	N	G	X

The above table is the combination of nucleotides. Where W (A or C or T or G), N (A or C or T or G) and X (A or C or T or G) are Wild Card locations. In the above any six locations can be fixed on the basis of the sequence of motif for which Likely Hood Ratio is to be calculated.

Table: 11 No. of occurrences of each nucleotide

G	2	7	0	0	2	7	0	0	1	2	3
A	1	0	7	0	2	0	7	0	3	2	2
T	3	0	0	7	1	0	0	7	2	1	1
C	1	0	0	0	2	0	0	0	1	2	1

The above table shows the column-wise number of occurrences of each nucleotide. Where 7 shows only one nucleotide is allowed at those locations. While, others are variable but column 5 & 10 both allow all four (G, A, T & C) nucleotides.

Table: 12 Finding Matches to (instances of) a PWM

G	0.29	1	0	0	0.29	1	0	0	0.14	0.29	0.43	
A	0.14	0	1	0	0.29	0	1	0	0.43	0.29	0.29	
T	0.43	0	0	1	0.14	0	0	1	0.29	0.14	0.14	
C	0.14	0	0	0	0.29	0	0	0	0.14	0.29	0.14	
*TIB=	1.64	3	3	3	1.05	3	3	3	1.64	1.05	1.64	25.0 Bits

The above table shows column-wise information bit and Total Information Bits (TIB). If each column is fixed (which is not possible) then the total information bit will be 33. Here loss of bits (33 – 25.0 = 8) i.e. the loss of 8 bits.

P(Sequence | Background Model) = Joint Probability (by formula (2))
= $(0.25)^{11} = 2.38*10^{-7}$
P(Sequence | Matrix Model), if we take Minimum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) = 7.82
P(Sequence | Matrix Model), if we take Maximum Probability of occurrence (by formula (1))
LHR = (matrix model/Background Model) = 15.14
Given Sequence G G A T G G A T C T C
LHR =9.92

ALGORITHM:
Step 1: Declare and initialize

```
          char x[7][11],s[11]
          int a[11],c[11],g[11],t[11]
          real A[11],C[11],G[11],T[11],TI[11],AA[11],CC[11],GG[11],TT[11],lhr
Step 2: input sequence in x[7][11]
Step 3: count occurrences of nucleotides A,C,T,G
Step 4: calculate probabilities of occurrences column-wise
Step 5: calculation of total information bit (at log base 2)
Step 6: input the sequence of eleven nucleotides
Step 7: for non-zero probabilities
                    begin
                      lhr <- (mm/2.38)*10000000
                      initialize float LHR
                      LHR := (log10(lhr))
                      LHR := LHR/0.3010          // for the conversion at log$_2$
                    end
Step 7: display LHR for given sequence of 11 nucleotides
Step 8: Stop
```

III. RESULTS

Experiments have been conducted & verified over 500 data sets and it is found that Likely Hood Ratio lies between 7.8 & 20.9. These experimental results give us some insight of motif finding problem. The three cases are generated, first, if the LLR is less than 10, it is non-considerable; second, if LLR is more than 10 and less than 15 then it is considered as motif that can be carried out for further calculations and the third, if LLR is between 15 and 20 is strongly recommended for the further calculation of motifs.

CONCLUSION

In this paper, when the motif length is of 11 nucleotides, and the probabilities of variable positions change, the LLR lies between 7.82 (min) and 20.92 (max), which are satisfied in above calculations. The loss of information bits does not affect the LHR. This method is considerable for the motif finding. The log to the base 2 is used as per the computational technique of the Information Theory. This method is position bound at locations 2 to 8 & 10 while, variable positions are 1, 9, and 11. This paper concludes for the motif finding with respect to their positions at the given motif length. This length may vary for finding motif in the researcher's interest. This calculation leads to the 3 displacement places. This is not compulsory to hold 2 to 8 & 10 as fixed while, 1, 9 and 11 as variables. The researchers can vary any three locations as per their own interest but the result will be same.

References

[1] A Brief Review of Computational Gene Prediction Methods Geno. Prot. Bioinfo. Vol. 2 No. 4 November 2004, Zhuo Wang, Yazhu Chen, and Yixue L]

[2] Computational Approaches to Gene Prediction, Journal of Microbiology, April 2006, p.137-144 2006 by Jin Hwan Do and Dong-Kug Choi Vol. 44, No. 2.

[3] 3, A combinatorial optimization approach for diverse motif finding applications in Algorithms for Molecular Biology 2006 by Elena Zaslavsky and Mona Singh.

[4] A boosting approach for motif modeling using ChIP-chip data, January 10, 2005, Pengyu Hong, X. Shirley Liu, Qing Zhou, Xin Lu, Jun S. Liu, and Wing H. Wong.

[5] Advance online publication, 20 November 2008 Engineering Letter, 16:4, EL_16_4_06.

[6] Elena Zaslavsky and Mona Singh A combinatorial optimization approach for diverse motif finding applications, Algorithms for Molecular Biology 2006, 1:13 17 August 2006.1186/1748-7188-1-13.

[7] International Conference on Bioinformatics, 2008 by Jan Schroder, Manfred Schimmler, Heiko Schroder, Karstern Tischer(BCBGC-08).

[8] Bioinformatics Research and Development Lecture Notes in Computer Science, 2007, Volume 4414/2007, 239-250, DOI: 10.1007/978-3-540-71233-6_19 Fast Search Algorithms for Position Specific Scoring Matrices Cinzia Pizzi, Pasi Rastas and Esko Ukkonen.

[9] Information Theory Primer With an Appendix on Logarithms http:// alum.mit.edu /www/toms/papers/primer/primer.pdf Postscipt version ftp://ftp.ncifcrf.gov/pub/delila /primer.ps, version = 2.64 of primer.tex 2010 Jan 08by Thomas D. Schneider.