The BaMM algorithm

The Supplementary Methods section provides equations that demonstrate that the binding strength of a sequence X (quantified by negative Gibbs energy) is approximately equal to the log-odds score S(x) = log2 ( Pmotif(X) / Pbg(X), so this score is used to calculate the relative binding strength of any proposed motif. In general, the authors use this result to support their claim that a purely statistical approach to motif binding is equivalent to the statistical physics approach.

PWM’s calculate the log-odds score for each potential motif by summing the log of the probability of occurrence of each nucleotide in each position in the motif, divided by the sum of the log of the probability of occurrence of the same nucleotide in any position in the background. This calculation means that each nucleotide is considered independently of other nucleotides in the sequence. The example provided in the paper is of two possible motifs, each of which occurs 50% of the time in the sequences:

GATC

GTAC

leading to a simple PWM of

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
| A | 0 | 0.5 | 0 | 0 |
| C | 0 | 0 | 0 | 1 |
| G | 1 | 0 | 0 | 0 |
| T | 0 | 0.5 | 0 | 0 |

The problem is that, using the PWM model, the query sequence GTTC will be scored as 1 x 0.5 x 1 x 1 = 0.5. The query sequence GTAC will also be scored as 1 x 0.5 x 1 x 1 = 0.5. Yet the second query sequence is a motif, whereas the first is not. The PWM does not contain the information that a T in 2nd position should be followed by an A in 3rd position whereas an A in 2nd position should be followed by a T in 3rd position.

The BaMM algorithm uses a Markov Model to address this problem. A kth order Markov Model will take note of the previous k positions, but runs into the problem that the number of parameters increases exponentially with the order of the model – i.e. for a sequence of length W there will be W X 3 X 4k parameters (the “3” refers to the three states of match, insertion or deletion). For each possible sequence an even larger number of sequences of length W needs to be found to provide robust probability information.

Previously developed algorithms that use Markov Models have developed various methods to reduce the parameters. The authors claim that these methods have been arbitrary, often result in loss of information, and are not suited to *de-novo* motif discovery.

BaMM uses an interpolated Markov Model to find motifs:

The algorithm begins with k=5 and finds all possible 5-mers in the set of sequences. It uses expectation maximisation to calculate transition and emission probabilities and develop a Markov Model based on these initial 5-mers. Then lengthens each k-mer by 1. The probability of each new sequence (the posterior probability) is transformed, using Bayes theorem, into the product of the likelihood of the added nucleotide and the prior probability (the probability of the previous (k-1) sequence).

Expectation maximisation is used to estimate the parameters of the new model that provide the maximum probability of the lengthened sequence. The E/M steps use pseudocounts, fixed as prior strengths: α0 = 1 and αk = β × γk − 1 for k ≥ 1, with hyperparameters β = 20 and γ = 3. The authors explain that the increasing strength of pseudocounts with increasing k reflects the prior belief that dependencies should quickly decline with distance, and that the strong regularization will prevent overtraining on all datasets.

Maximizing the posterior probability yields

p(k)j (xj |xj−k: j−1) = [n j (xj−k: j ) + αk pj(k−1) (xj |xj−k+1: j−1) ] / [ nj−1(xj−k: j−1)+αk ]

xj=last nucleotide in new k-mer (the new additional nucleotide)

xj-k= first nucleotide in new k-mer

xj−k: j−1=previous (k-1)-mer without the addition of xj

n j (xj−k: j )=number of new k-mers

xj−k+1: j−1= the previous (k-1)-mer also using one less nucleotide at the start.

For frequently occurring k-mers xj − k: j − 1 the counts dominate over the pseudocounts and we can accurately estimate the conditional probabilities from the counts. For k-mers with few counts the pseudocounts dominate and the probability reverts to the estimate at order k−1 (represented in the equation by xj−k+1: j−1).

Because their probability matrix is derived from the Bayesian Markov Model approach, the authors call the matrix a BaMM rather than a PWM or PSSM.

Relevant options in the command line:

Sequence options

--negSequenceSet <FILEPATH>

FASTA file with negative/background sequences used to learn the (homogeneous) background BaMM. If not specified, the background BaMM is learned from the positive sequences.

--reverseComp

Search motifs on both strands (positive sequences and reverse complements). This option is e.g. recommended when using sequences derived from ChIP-seq experiments.

Options to initialize one or more BaMMs from XXmotif PWMs

--minPWMs <INTEGER>

Minimum number of PWMs. The options --maxPValue and –minOccurrence are ignored. The default is 1.

--maxPWMs <INTEGER>

Maximum number of PWMs.

--maxPValue <FLOAT>

Maximum p-value of PWMs. This filter is not applied to the top minimum number of PWMs (see --minPWMs). The default is 1.0.

--minOccurrence <FLOAT>

Minimum fraction of sequences that contain the motif. This filter is not applied to the top minimum number of PWMs (see --minPWMs). The default is 0.05.

--rankPWMs <INTEGER> [<INTEGER>...]

PWM ranks in XXmotif results. The former options to initialize BaMMs from PWMs are ignored.

Options for (inhomogeneous) motif BaMMs

-k <INTEGER>

Order. The default is 2.

-a|--alpha <FLOAT> [<FLOAT>...]

Order-specific prior strength. The default is 1.0 (for k = 0) and 20 x 3^(k-1) (for k > 0). The options -b and -g are ignored.

-b|--beta <FLOAT>

Calculate order-specific alphas according to beta x gamma^(k-1) (for k > 0). The default is 20.0.

-g|--gamma <FLOAT>

Calculate order-specific alphas according to beta x gamma^(k-1) (for k> 0). The default is 3.0.

--extend <INTEGER>{1,2}

Extend BaMMs by adding uniformly initialized positions to the left and/or right of initial BaMMs. Invoking e.g. with --extend 0 2 adds two positions to the right of initial BaMMs. Invoking with --extend 2 adds two positions to both sides of initial BaMMs. By default, BaMMs are not being extended.

Options for the (homogeneous) background BaMM

-K <INTEGER>

Order. The default is 2.

-A|--Alpha <FLOAT>

Prior strength. The default is 10.0.

EM options

-q <FLOAT>

Prior probability for a positive sequence to contain a motif. The default is 0.9.

-e|--epsilon <FLOAT>

The EM algorithm is deemed to be converged when the sum over the asolute differences in BaMM probabilities from successive EM rounds is smaller than epsilon. The default is 0.001.

XXmotif options

--XX-ZOOPS

Use the zero-or-one-occurrence-per-sequence model (default).

--XX-MOPS

Use the multiple-occurrence-per-sequence model.

--XX-OOPS

Use the one-occurrence-per-sequence model.

--XX-seeds ALL|FIVEMERS|PALINDROME|TANDEM|NOPALINDROME|NOTANDEM

Define the nature of seed patterns. The default is to start using ALL seed pattern variants.

--XX-gaps 0|1|2|3

Maximum number of gaps used for seed patterns. The default is 0.

--XX-pseudoCounts <FLOAT>

Percentage of pseudocounts. The default is 10.0.

--XX-mergeMotifsThreshold LOW|MEDIUM|HIGH

Define the similarity threshold used to merge PWMs. The default is to merge PWMs with LOW similarity in order to reduce runtime.

--XX-maxPositions <INTEGER>

Limit the number of motif positions to reduce runtime. The default is 17.

--XX-noLengthOptimPWMs

Omit the length optimization of PWMs.

--XX-K <INTEGER>

Order of the (homogeneous) background BaMM. The default is either 2 (when learned on positive sequences) or 8 (when learned on background sequences).

--XX-A <FLOAT>

Prior strength of the (homogeneous) background BaMM. The default is 10.0.

--XX-jumpStartPatternStage <STRING>

Jump-start pattern stage using an IUPAC pattern string.

--XX-jumpStartPWMStage <FILEPATH>

Jump-start PWM stage reading in a PWM from file.

--XX-localization

Calculate p-values for positional clustering of motif occurrences in positive sequences of equal length. Improves the sensitivity to find weak but positioned motifs.

--XX-localizationRanking

Rank motifs according to localization statistics.

--XX-downstreamPositions <INTEGER>

Distance between the anchor position (e.g. the transcription start site) and the last positive sequence nucleotide. Corrects motif positions in result plots. The default is 0.

--XX-batch

Suppress progress bars.

Options to score sequences

--scorePosSequenceSet

Score positive (training) sequences with optimized BaMMs.

--scoreNegSequenceSet

Score background (training) sequences with optimized BaMMs.

--scoreTestSequenceSet <FILEPATH> [<FILEPATH>...]

Score test sequences with optimized BaMMs. Test sequences can be provided in a single or multiple FASTA files.

Output options

--saveInitBaMMs

Write initialized BaMM(s) to disk.

--saveBaMMs

Write optimized BaMM(s) to disk.

--verbose

Verbose terminal printouts.

Terminal Printout:

jackie@jackie-Latitude-E6320[jackie] BaMMmotif '/home/jackie/Documents/BIOC7002ResearchProject/Data/ResearchProjectRFiles/NullSequences/allSequences/motifsNullSequences/BammMany' '/home/jackie/Documents/BIOC7002ResearchProject/Data/ResearchProjectRFiles/NullSequences/allSequences/1kbnullSequences.fasta' --reverseComp --maxPValue 1e-50

BaMM!motif version 1.0

Check composition in sequence sets:

posSet: A: 30.00% C: 20.00% G: 20.00% T: 30.00%

posSet statistics: 17226, maxMultSeq: 1, minLength: 4007, maxLength: 4007, total: 69024582

ELONGATE IUPAC SEED PATTERNS

elongate fivemers: [===============================] 100.0%

elongate palindromic sixmers: [===============================] 100.0%

elongate tandemic sixmers: [===============================] 100.0%

TOP 10 Motifs:

1 : R Y R C C T G T A A T C C C A G C sites: 4846 log(E-Value): -4.69e+04 E-Value: 1e-20371 Pos Pval: 9.94e-01

2 : R Y R C C T G T A A T C C C A G C sites: 4845 log(E-Value): -4.69e+04 E-Value: 1e-20366 Pos Pval: 9.94e-01

3 : G C T G G G A T T A C A G G Y R Y sites: 4846 log(E-Value): -4.61e+04 E-Value: 1e-20017 Pos Pval: 9.48e-01

4 : G C T G G G A T T A C A G G Y R Y sites: 4846 log(E-Value): -4.61e+04 E-Value: 1e-20017 Pos Pval: 9.48e-01

5 : A Y R C C T G T A A T C C C A G C sites: 4042 log(E-Value): -4.05e+04 E-Value: 1e-17597 Pos Pval: 9.91e-01

6 : C C Y G C C T C R G C C T C C C R sites: 4464 log(E-Value): -4.04e+04 E-Value: 1e-17537 Pos Pval: 9.70e-01

7 : G C T G G G A T T A C A G G Y R T sites: 4042 log(E-Value): -4.00e+04 E-Value: 1e-17365 Pos Pval: 9.00e-01

8 : T T G G G A G G C Y G A G G C R G sites: 4163 log(E-Value): -3.97e+04 E-Value: 1e-17252 Pos Pval: 9.99e-01

9 : T T G G G A G G C Y G A G G C R G sites: 4163 log(E-Value): -3.97e+04 E-Value: 1e-17252 Pos Pval: 9.99e-01

10 : T T G G G A G G C Y G A G G C R G sites: 4163 log(E-Value): -3.97e+04 E-Value: 1e-17252 Pos Pval: 9.99e-01

Merge Motifs: ...... round 1: 314 . round 2: 300 . round 3: 300

TOP 10 Motifs:

1 : R Y R C C T G T A A T C C C A G C sites: 4842 log(E-Value): -4.73e+04 E-Value: 1e-20523 Pos Pval: 9.89e-01

2 : C C Y G C C T C R G C C T C C C R sites: 4582 log(E-Value): -4.16e+04 E-Value: 1e-18052 Pos Pval: 9.44e-01

3 : T G T Y R S C C A G G C T G G W S sites: 4058 log(E-Value): -3.68e+04 E-Value: 1e-15970 Pos Pval: 7.24e-01

4 : R G C Y R G G Y R T G G T G G Y sites: 4351 log(E-Value): -3.33e+04 E-Value: 1e-14451 Pos Pval: 9.90e-01

5 : T T G C A G T G A G C Y R A G A T sites: 3386 log(E-Value): -3.29e+04 E-Value: 1e-14285 Pos Pval: 9.95e-01

6 : T G G T C T Y R A W C T C C T G A sites: 3469 log(E-Value): -3.26e+04 E-Value: 1e-14161 Pos Pval: 9.99e-01

7 : T T A G T A G A G A Y R G G G T T sites: 3159 log(E-Value): -3.22e+04 E-Value: 1e-13988 Pos Pval: 9.10e-01

8 : R C C A C T G C A C T C C A G C sites: 3537 log(E-Value): -3.10e+04 E-Value: 1e-13446 Pos Pval: 9.87e-01

9 : G G A G G Y R G A G S T T G C A G sites: 2941 log(E-Value): -2.78e+04 E-Value: 1e-12089 Pos Pval: 9.16e-01

10 : G G A G A A T B R C T T G A A C C sites: 2564 log(E-Value): -2.77e+04 E-Value: 1e-12014 Pos Pval: 9.11e-01

Number of motifs: 300

ITERATIVELY REFINE PWMS

iterate Motifs: [===============================] 100.0% Merge Motifs: . round 1: 149 . round 2: 149

iterate Motifs: [===============================] 100.0% Merge Motifs: . round 1: 149

iterate Motifs: [===============================] 100.0%

\*\*\*\*\*\*\*\*\*\* Final results \*\*\*\*\*\*\*\*\*\*

Model Number 1 from 149:

===============================================================================================================================================

R Y R C C T G T A A T C C C A G C :

5322 sites in 17226 sequences

=> occurrence: 31%

E-Value: 1e-10406

pos Pval: 9.97e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 72.57 3.66 21.03 2.77 2.77 2.79 3.64 2.76 93.36 79.80 2.72 2.76 2.81 2.77 93.51 3.54 2.86

C 1.93 58.30 2.09 92.36 92.33 1.86 1.90 2.00 1.92 1.83 1.92 92.40 92.28 91.92 1.83 1.83 92.26

G 22.62 2.24 73.63 1.88 1.86 1.83 91.66 1.93 1.93 15.58 1.83 1.85 1.93 1.83 1.88 91.83 1.93

T 2.86 35.78 3.23 2.96 3.01 93.50 2.77 93.29 2.77 2.77 93.51 2.98 2.96 3.46 2.76 2.77 2.93

Model Number 2 from 149:

===============================================================================================================================================

A G C T A C T Y R G G A G G C T G :

5365 sites in 17226 sequences

=> occurrence: 31%

E-Value: 1e-9487

pos Pval: 9.66e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 92.68 4.04 3.01 3.21 91.29 2.87 3.16 4.30 30.44 4.48 4.33 92.53 4.74 4.55 2.93 3.28 4.38

C 1.98 2.37 91.03 2.44 2.30 90.28 2.97 58.15 2.71 1.91 1.98 1.95 1.86 2.25 91.15 2.22 1.90

G 2.29 90.62 2.03 1.88 3.30 2.08 2.10 3.51 62.49 90.64 90.69 2.36 90.42 90.15 2.08 2.05 90.82

T 3.03 2.94 3.91 92.45 3.08 4.74 91.75 34.02 4.33 2.94 2.98 3.15 2.96 3.03 3.82 92.43 2.87

Model Number 3 from 149:

===============================================================================================================================================

A C T G C A C T C C A G C C T G G :

5254 sites in 17226 sequences

=> occurrence: 31%

E-Value: 1e-8675

pos Pval: 1.00e+00

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 91.64 2.95 2.96 3.95 3.07 91.75 2.83 2.95 2.93 2.91 92.99 3.97 3.10 3.03 2.96 5.30 4.92

C 2.23 78.26 2.90 1.97 89.30 2.82 89.97 2.57 90.85 89.68 1.95 2.14 89.95 91.32 2.38 2.09 2.14

G 3.08 2.19 2.26 90.99 1.99 2.40 2.06 1.97 2.56 1.88 2.14 90.39 2.33 1.99 1.93 89.56 89.94

T 3.03 16.58 91.85 3.07 5.63 3.02 5.13 92.49 3.64 5.51 2.90 3.48 4.59 3.64 92.70 3.03 2.98

Model Number 4 from 149:

===============================================================================================================================================

C C A G C A C T T T G G G A G G C :

5212 sites in 17226 sequences

=> occurrence: 30%

E-Value: 1e-8616

pos Pval: 9.96e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 3.11 2.90 93.00 6.61 3.52 92.46 3.07 2.95 2.98 3.09 3.72 4.41 4.07 93.23 5.11 4.03 3.23

C 88.83 87.42 1.87 2.09 90.63 2.14 89.39 2.09 3.26 2.58 2.23 1.92 2.09 1.87 1.95 2.16 90.21

G 2.13 2.04 2.14 88.28 2.06 2.46 2.23 1.85 2.09 2.09 90.82 90.63 90.84 2.04 89.91 90.65 2.18

T 5.91 7.62 2.97 3.00 3.77 2.91 5.29 93.09 91.64 92.22 3.21 3.02 2.98 2.84 3.00 3.14 4.36

Model Number 5 from 149:

===============================================================================================================================================

A G G A G T T Y R A G A C C A G C :

4937 sites in 17226 sequences

=> occurrence: 29%

E-Value: 1e-8476

pos Pval: 9.90e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 90.04 4.55 4.40 92.40 4.29 16.66 2.89 3.31 30.03 92.73 3.86 91.92 3.26 3.51 93.02 4.03 3.44

C 2.03 1.85 2.24 1.96 1.92 2.83 2.22 65.08 2.97 2.05 2.22 1.96 89.32 89.83 1.92 2.33 90.09

G 4.93 90.83 90.62 2.61 90.88 1.96 1.96 3.42 63.89 2.24 91.07 3.16 2.44 1.85 2.03 76.81 2.37

T 2.98 2.76 2.72 3.02 2.89 78.53 92.91 28.17 3.09 2.96 2.83 2.94 4.97 4.78 3.00 16.81 4.08

Model Number 6 from 149:

===============================================================================================================================================

T T G C A G T G A G C Y R A G A T :

4227 sites in 17226 sequences

=> occurrence: 25%

E-Value: 1e-7802

pos Pval: 9.98e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 2.87 2.77 3.39 2.83 91.09 3.43 2.81 3.58 93.31 3.30 2.89 4.92 23.37 93.07 3.26 92.94 2.92

C 5.34 2.01 1.99 91.78 2.01 2.03 2.24 1.96 1.83 2.01 91.67 68.08 2.57 1.90 1.92 1.88 4.11

G 1.94 2.18 91.41 1.92 4.03 91.73 2.14 91.60 2.07 91.82 1.99 2.74 70.59 2.07 91.97 2.26 1.92

T 89.83 93.03 3.20 3.45 2.85 2.79 92.79 2.83 2.77 2.85 3.43 24.25 3.45 2.94 2.83 2.89 91.03

Model Number 7 from 149:

===============================================================================================================================================

T T A G T A G A G A Y R G G G T T :

4715 sites in 17226 sequences

=> occurrence: 27%

E-Value: 1e-7696

pos Pval: 9.96e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 2.76 2.97 88.54 4.40 2.88 92.49 3.47 92.82 4.26 91.66 3.30 24.61 6.10 5.38 5.00 2.93 2.80

C 2.10 3.66 2.04 2.03 2.80 1.87 2.91 1.99 2.22 1.97 58.96 2.12 1.93 2.03 2.14 2.18 2.18

G 2.18 2.10 3.64 90.43 2.08 2.76 90.70 2.33 90.52 3.20 2.45 69.91 88.96 89.45 88.44 2.53 1.93

T 92.94 91.24 5.75 3.13 92.22 2.86 2.90 2.84 2.97 3.15 35.27 3.34 2.99 3.13 4.40 92.34 93.07

Model Number 8 from 149:

===============================================================================================================================================

G G A G A A T B R C T T G A A C C :

4690 sites in 17226 sequences

=> occurrence: 27%

E-Value: 1e-7668

pos Pval: 8.54e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 4.89 5.30 92.55 3.36 84.39 92.99 2.86 3.52 25.30 2.95 5.55 2.95 3.13 93.19 83.11 3.09 3.01

C 2.67 2.12 1.85 1.95 1.99 1.99 2.86 52.48 2.82 91.07 2.86 2.05 2.08 1.85 1.97 90.14 89.45

G 87.80 89.72 2.76 91.48 10.83 2.20 1.89 16.78 67.95 1.87 11.99 2.59 91.97 2.14 11.97 1.93 1.83

T 4.62 2.84 2.82 3.19 2.78 2.80 92.37 27.20 3.90 4.08 79.58 92.39 2.80 2.80 2.94 4.82 5.69

Model Number 9 from 149:

===============================================================================================================================================

R G C Y R G G Y R Y G G T G G C W :

4905 sites in 17226 sequences

=> occurrence: 28%

E-Value: 1e-7537

pos Pval: 9.88e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 35.03 3.13 2.89 3.24 32.06 2.98 3.00 3.28 35.19 3.24 16.05 2.98 2.78 2.98 2.93 2.74 16.23

C 1.94 1.89 92.02 62.49 2.04 1.83 1.89 72.21 2.20 41.01 2.02 1.85 2.02 1.81 1.89 88.94 2.05

G 59.60 92.15 2.00 2.09 62.49 92.37 92.30 2.17 59.36 2.07 78.82 92.30 1.89 92.39 92.35 1.83 14.62

T 3.41 2.81 3.07 32.15 3.39 2.80 2.80 22.33 3.22 53.65 3.09 2.85 93.30 2.80 2.81 6.47 67.07

Model Number 10 from 149:

===============================================================================================================================================

G G T T T C A C C A T G T T R G C :

4269 sites in 17226 sequences

=> occurrence: 25%

E-Value: 1e-6375

pos Pval: 9.00e-01

MODEL:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

-----------------------------------------------------------------------------------------------------------------------------------------------

A 4.72 4.57 2.96 2.81 2.81 2.83 81.11 3.04 2.94 78.47 2.94 5.72 2.91 2.79 23.08 4.32 3.11

C 2.09 2.09 2.13 2.05 1.96 87.78 2.20 90.85 86.78 2.15 4.05 3.22 2.58 2.18 1.92 6.07 81.76

G 89.87 89.83 1.88 1.88 1.88 1.96 9.44 1.98 1.86 15.48 2.28 87.70 1.96 1.88 71.90 86.04 2.18

T 3.30 3.49 93.01 93.25 93.33 7.41 7.24 4.11 8.41 3.87 90.71 3.34 92.52 93.14 3.08 3.55 12.94

------------ time: 136678 seconds (2277.97 minutes) ------------

jackie@jackie-Latitude-E6320[jackie] [11:45AM]