Supplementary materials to the following article:

Deciphering Principles of Transcription Regulation in Eukaryotic Genomes

Dat H. Nguyen and Patrik D'haeseleer

Department of Genetics, Harvard Medical School, Boston, MA 02115

All correspondences should be sent to Dat H. Nguyen
(dnguyen@genetics.med.harvard.edu)

Supplementary 1:

Figures 3a and 3b in the main text show that the expression correlation between genes that contain the PAC motif quickly drops to zero the further the PAC motif is located from the start codon, whereas MED predicts a significant influence for the PAC motif over a much wider range. Applying cluster analysis on the genes containing PAC spaced between 600 and 750 bp upstream, we find that their expression profiles follow two somewhat anti-correlated patterns (Fig. SF1a). Similarly, the genes containing PAC spaced between 750 and 1000 bp upstream cluster into three distinct expression patterns (Fig. SF1b), two that are anti-correlated and a third one with low expression levels. This clearly demonstrates that lack of co-expression does not necessarily imply lack of co-regulation, because average pairwise correlation is a poor measure to infer motif strength accurately as mentioned in the main text. In contrast, for a very tightly co-expressed gene ensemble (Fig SF1c), average correlation does accurately reflect the degree of co-regulation caused by the set of motifs in the ensemble.

Supplementary 2: Apparent PAC/RRPE synergetic behavior on gene expression

Expression data for PAC-only, RRPE-only, and PAC/RRPE-only gene ensembles are presented in Fig SF2. In this figure, we show both an average gene expression for each gene ensemble for each condition derived from experimental data (SF2a) and the corresponding one derived from MED using Eq. 1 (SF2b). These figures show a consistent feature that when a gene contains both PAC and RRPE motifs, its expression signal is larger in magnitude compared to a gene that contains either motif alone. Note that the term "only" here refers to the fact that the gene ensemble of one particular motif contains only such motif and nothing else. This choice of gene ensembles (instead of PAC, RRPE and PAC/RRPE ensembles) eliminates side effects of other motifs on gene expression pattern of the corresponding gene ensemble, thereby allowing us to pinpoint the root cause of apparent enhancement in expression level of PAC/RRPE-only containing gene ensemble compared to PAC-only and RRPE-only containing gene ensembles.

In Fig. SF2c, the strengths of the PAC and RRPR motifs in the gene ensemble containing both of them are plotted as a function of RRPE motif's position averaged over

all PAC positions. The two curves in this Figure SF2-c are complement to the two black diamond curves shown in Figure 3a in the main text.

Supplementary 3: PAC/RRPE motif data

Figure SF3-A shows the position distribution of PAC and RRPE in associated promoters of genes containing each. In Figure SF3-B, the distribution of expression correlation coefficients for a set of instances of PAC- and RRPE-containing gene ensembles is shown.

Supplementary 4: Proof of uniqueness (see supplementary 11 for definition of mathematical notations)

As shown in Eq. 2 in the main text, the first part of the MED algorithm decomposes the gene expression data matrix **E** (m-genes by n-conditions) in order to obtain motif matrix **M** (m-genes by k-motifs) and the matrix of proxy regulator activity **A** (k-motifs by n-conditions) according to the following equation:

$$E \approx M \bullet A \qquad [2]$$

For any invertible k x k matrix **X**, Eq. 2 above can always be written as follows:

$$E \approx M \bullet A = M \bullet X \bullet X^{-1} \bullet A = M^{*} \bullet A^{*} \qquad [S1]$$

In order to prove that matrices **M** and **A** are unique, it is sufficient to prove two things. First we prove that the MED formalism restricts **X** to be an identity matrix, provided that (1) M and A are full rank, and (2) the matrix M is sparse and rectangular (i.e. there must be at least k-1 zero-matrix elements in each column in general, and the number of rows is larger than the number of columns). Secondly, when rows of matrix **A** are linearly dependent or near-linearly dependent, we prove that MED formalism can still produce unique motif matrix **M** regardless of rank deficiency of matrix **A**, due to the second term of Eq. 3 or 4.

Proof

1. Matrix X is an identity matrix as an outcome of the MED formalism:

In Eq. 1 in the main text:

$$E_{gc} \approx \sum_{j \in \Omega_g} M_{gj} A_{jc}$$

if $j \notin \Omega_g$ (i.e. motif j does not exist in the promoter of gene **g**) then the corresponding matrix element M_{gj} of the motif matrix M must be zero and remain so and as it is always the case in MED. As a result, for each matrix element M_{gj} of matrix M whose value is zero, the corresponding matrix element M^{*}_{gj} of matrix M^{*} must also be zero after the

linear transformation with matrix X. Therefore, for a zero element M_{gj} in the column **j**th of the motif matrix M, we have:

$$M^*_{gj} = \sum_{l=1}^{j-1} M_{gl} X_{lj} + \overset{0}{\overbrace{M_{gj} X_{jj}}} + \sum_{l=j+1}^{k} M_{gl} X_{lj} = \sum_{l=1}^{j-1} M_{gl} X_{lj} + \sum_{l=j+1}^{k} M_{gl} X_{lj} = 0 \qquad \text{[S2]}$$

For all genes with zero-entry in the column **j**th, Eq. S2 can be written in the matrix form:

$$\begin{bmatrix} M^*_{g_1 j} \\ M^*_{g_2 j} \\ \cdots \\ M^*_{g_p j} \end{bmatrix} = \begin{bmatrix} M_{g_1 1} & M_{g_1 2} & \cdots & M_{g_1 j-1} & M_{g_1 j+1} & \cdots & M_{g_1 k} \\ M_{g_2 1} & M_{g_2 2} & \cdots & M_{g_2 j-1} & M_{g_2 j+1} & \cdots & M_{g_2 k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ M_{g_p 1} & M_{g_p 2} & \cdots & M_{g_p j-1} & M_{g_p j+1} & \cdots & M_{g_p k} \end{bmatrix} \begin{bmatrix} X_{1j} \\ X_{2j} \\ \cdots \\ X_{j-1,j} \\ X_{j+1 j} \\ \cdots \\ X_{kj} \end{bmatrix} = \vec{0} \qquad \text{[S3]}$$

if

$$\mathsf{M} = \begin{bmatrix} M_{g_1 1} & M_{g_1 2} & \cdots & M_{g_1 j-1} & M_{g_1 j+1} & \cdots & M_{g_1 k} \\ M_{g_2 1} & M_{g_2 2} & \cdots & M_{g_2 j-1} & M_{g_2 j+1} & \cdots & M_{g_2 k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ M_{g_p 1} & M_{g_p 2} & \cdots & M_{g_p j-1} & M_{g_p j+1} & \cdots & M_{g_p k} \end{bmatrix} \text{ and } \vec{\mathsf{X}}_{\bullet j} = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \cdots \\ X_{j-1,j} \\ X_{j+1 j} \\ \cdots \\ X_{kj} \end{bmatrix} \qquad \text{[S4]}$$

then the Eq. S3 becomes:

$$\mathsf{M} \bullet \vec{\mathsf{X}}_{\bullet j} = \vec{0} \qquad \text{[S5]}$$

Since the number of zero elements **p** (number of genes that do not contain motif indexed as j) in each column j of matrix M is much larger than k-1 (number of motifs in a motif set) for eukaryotic genome, and the way we estimate the initial non-zero values for matrix M in Eq. 2, the matrix M in Eq. S5 is full-rank. One possible exception is when a motif only occurs in the presence of a second motif (as may be the case with two variants of the same motif). However, in this case it is easy to see that the use of the

lambda parameter still guarantees a single optimum, similar to the proof for rank-deficient A below. Accordingly, the equality in Eq. S5 is true if and only if:

$$\vec{X}_{\bullet j} = \vec{0} \qquad \text{[S6]}$$

Therefore, all entries in column j of matrix X must be zero except X_{jj}, which can be of any value, for all j=1..k. As a result, matrix X is a diagonal matrix. Since the step (c) of the MED formalism requires that each row of matrix A has the unit norm, each diagonal element of matrix X must be 1. Therefore X is an identity matrix. This completes the proof of the first part.

2. In MED, the motif matrix M is unique even when matrix A is rank deficient:

When two or more distinct motifs are regulated by the same regulator, two or more corresponding rows of the global proxy regulator activity matrix A become linearly dependent (i.e. A is rank deficient), an event that cannot be controlled *a priori* mathematically. The same is also true when two or more regulators work together within a finite number of experimental conditions under consideration. For example, let $\vec{A}_{i_1\bullet} = \alpha_2 \vec{A}_{i_2\bullet} = \cdots = \alpha_q \vec{A}_{i_q\bullet}$. In that case, the contribution of these motifs to the expression of gene g can be written as:

$$\left[\vec{A}^T_{i_1\bullet} \quad \vec{A}^T_{i_2\bullet} \quad \cdots \quad \vec{A}^T_{i_q\bullet}\right] \bullet \begin{bmatrix} M_{gi_1} \\ M_{gi_2} \\ \vdots \\ M_{gi_q} \end{bmatrix} = \left[\vec{A}^T_{i_1\bullet} \quad \alpha_2^{-1}\vec{A}^T_{i_1\bullet} \quad \cdots \quad \alpha_q^{-1}\vec{A}^T_{i_1\bullet}\right] \bullet \begin{bmatrix} M_{gi_1} \\ M_{gi_2} \\ \vdots \\ M_{gi_q} \end{bmatrix} \approx \left[E^T_{g\bullet}\right] \qquad \text{[S7]}$$

Obviously a set of infinite solutions can result from Eq. S7, because the rank of the matrix in Eq. S7 is one and hence values can be shifted freely among the q motif strength variables M_{gj_l}, l=1..q.

However, in the MED formalism, we require that the sum of square of motif strengths in each gene be minimal or as close as possible to some user predefined values (if they are known *a priori*) as the second term of Eq. 4 or Eq. 5 represents, respectively. Mathematically, this can be written in the matrix form as follows:

$$\begin{bmatrix} \vec{A}^T_{i_1\bullet} & \alpha_2^{-1} \bullet \vec{A}^T_{i_1\bullet} & \ldots & \alpha_q^{-1} \bullet \vec{A}^T_{i_1\bullet} \\ & \sqrt{\lambda} \bullet I & & \end{bmatrix} \bullet \begin{bmatrix} M_{gj_1} \\ M_{gj_2} \\ \ldots \\ M_{gj_q} \end{bmatrix} \approx \begin{bmatrix} E^T_{g\bullet} \\ \vec{0} \end{bmatrix} \qquad \text{for Eq. 4} \qquad \text{[S8]}$$

or

$$\begin{bmatrix} \vec{A}_{i_1}^T \bullet & \alpha_2^{-1}\vec{A}_{i_1}^T \bullet & \cdots & \alpha_q^{-1}\vec{A}_{i_1}^T \bullet \\ & \sqrt{\lambda} \bullet I & & \end{bmatrix} \bullet \begin{bmatrix} M_{gi_1} \\ M_{gi_2} \\ \vdots \\ M_{gi_q} \end{bmatrix} \approx \begin{bmatrix} E_{g\bullet}^T \\ \sqrt{\lambda} \bullet \left(\vec{M}_g^*\right)^T \end{bmatrix} \qquad \text{for Eq. 5} \qquad [S9]$$

In the Eq. S9, $\vec{M}_g^*$ is a vector of user defined priors for the motif strength for gene **g**. Since the matrix in Eqs. S8-S9 is always full-rank, due to the block represented by the identity matrix I regardless of data in matrix **A**, Eqs. S8 or S9 has a unique least square solution. This completes the proof of the second part.

Note that in Eq. S1, the linear transformations of matrix M to a matrix represented by $M \bullet X$ and of matrix A to a matrix represented by $X^{-1} \bullet A$ are known as (oblique) rotation (Paatero et al., 2002). Historically, in the proof to part 1, the idea of using an *a priori* known pattern of zero elements in matrix factorization problem to reduce the rotational degrees of freedom (i.e. number of off-diagonal non-zero elements of matrix X) is a well-known one. If there are enough pre-determined known zero elements in one of factored matrices, the rotation becomes prohibited as discussed by Anderson (Anderson, 1984); Paatero and co-workers (Paatero et al., 2002); and an alternative proof to part 1 on the structure of matrix X^{-1} instead of matrix X was previously provided by Liao and co-workers (Liao et al., 2003). In such case, the matrix X and hence X^{-1} become diagonal matrices, otherwise such pre-determined pattern of zero is violated. However, the use of the pre-determined pattern of zeros alone is not enough, since the rotation cannot deal with, and therefore guarantee, the uniqueness of motif matrix M when rows of matrix A are linearly dependent as discussed in the proof to part 2 above. The main mathematical step that makes our MED formalism work is the proof in step 2, which guarantees the uniqueness of the motif matrix M regardless of the data pattern in matrix A.

Supplementary 5: Procedure for initializing non-zero matrix elements of the motif matrix M

If the regulatory motifs were defined by a consensus sequence, the motif matrix M could simply be initialized by counting how many times each specific consensus sequence occurs in each promoter (Bussemaker et al., 2001). However, since we are dealing in this case with motifs which are specified by a position specific scoring matrix (PSSM), we can weight this count by the score of each motif instance.

The initial weight M_{gm} for motif *m* and gene *g* is given by:

$$M_{gm} = \sum_i e^{S_i(g,m)} / \max_{i,g}\left(e^{S_i(g,m)}\right) \qquad [S10]$$

where $S_i(g,m)$ is the ScanAce score (Hughes et al., 2000) of the i^{th} instance of motif m in the promoter of gene g, which is approximately proportional to the log

probability of binding to that sequence (Stormo, 2000). Note that for the best-fit sequences to the PSSM for the motif, the initial value for M_{gm} reverts to a count of the number of motif instances.

Supplementary 6: Standard least square formalism used in MED

Given a matrix $M \in \Re^{mxn}, m \geq n$ and a vector of observable data $\vec{b} \in \Re^m$, the least square problem is to find a vector $\vec{x} \in \Re^n$ such that $\left\| M\vec{x} - \vec{b} \right\|$ is minimized. Geometrically, this can be represented in Fig SF4 where the vector solution $\vec{x}$ can be calculated by projecting the vector of observable data $\vec{b}$ orthogonally onto the range of the matrix *M* spanned by its columns vectors. Mathematically we can write:

$$M\vec{x} \approx \vec{b} \qquad [S11]$$

$$\Rightarrow \qquad \vec{x} = (M^T M)^{-1} M^T \vec{b} \qquad [S12]$$

Using singular value decomposition for the matrix M, we have:

$$M = U \bullet \Sigma \bullet V^T \qquad [S13]$$

Substitute S13 into S12, the solution to the least square problem can be computed using the following formula:

$$\vec{x} = V \bullet \Sigma^{-1} \bullet U^T \bullet \vec{b} \qquad [S14]$$

If the vector of observable data $\vec{b}$ in Eq. S14 is a matrix, then Eq. S14 becomes:

$$X = V \bullet \Sigma^{-1} \bullet U^T \bullet B \qquad [S15]$$

where ***X*** is now a matrix instead of a vector.

Supplementary 7: Definition of the total residual variance, convergence and selection of λ

From Eq. S11, the residual matrix ***R*** is defined as follows:

$$R = M \bullet X - B \qquad [S16]$$

therefore total residual variance **R** is then defined as:

$$\mathbf{R} = \sum_{i=1}^{m} \mathrm{var}(\vec{R}_{i\bullet}) \qquad [S17]$$

Since the second term of the Eq. 4 imposes additional constraint on the factorization for any $\lambda>0$, test error will always be larger than the corresponding with $\lambda=0$. Therefore, the value of λ must be chosen in such a way that it does not affect the test error significantly and Figs. SF5a and SF5b show both training and test errors, respectively, for $\lambda=0$ and $\lambda=10^{-4}$, which is the value we used in this work. Given this λ, we also show the trajectory of MED algorithm for both training and test error in factoring the expression matrix in Figs. SF4c and SF4d, respectively.

Supplementary 8: Extended form of Eq. 1 to account for non-linear motif-motif interactions

In the Eq. 1 in the main text, the log expression ratio of gene **g** under environmental condition **c**, $\boldsymbol{E_{gc}}$, can be approximated as follows:

$$E_{gc} \approx \sum_{j \in \Omega_g} M_{gj} A_{jc} \qquad [1]$$

Although motif-motif interactions are not explicitly included in Eq. 1, such interactions can still be captured to some degree by assigning a different weight to a motif depending on the presence of other motifs it interacts with. Nevertheless, one can also choose to include interaction terms explicitly in Eq. 1, according to the following equation:

$$E_{gc} \approx \sum_{j \in \Omega_g} M_{gj} A_{jc} + \sum_{jl \in \Omega_g x \Omega_g} M_{gjl} A_{jlc} \qquad [S18]$$

where M_{gjl} represents the influence of interaction between motifs j and l present in the promoter of gene g; and A_{jlc} represents the global proxy activity of the virtual regulator for such interaction. With this explicit form, one can investigate combinatorial regulation principles in terms of motif-motif interactions, which is an important feature in eukaryotic genomes (Kellis et al., 2003). The differences between Eqs. 1 and S18 are extra non-zero entries for each gene (extra columns) in the motif matrix M and the MED formalism can still be applied transparently.

Supplementary 9: Data

Motif data and expression data sets along with other essential data used in this work are available from our website at http://arep.med.harvard.edu/MED

Supplementary 10: p-Value calculations for motif strengths derived from the random shuffling of expression data

In the random shuffling test, we recalculated the strength of each motif (random motif strength) in the same manner as being done in Fig. 3 (in the main text) for each randomly shuffled expression matrix, for 100 times. The supplementary Figures SF6a-c show the distribution of such random motif strengths for the PAC, RRPE, MCB, and RAP1 motifs in the promoter location and with orientation, where they are most active,

along with the corresponding ones (actual motif strength) derived from the actual expression data. Since the actual motif strength is far away from the mean of the distribution of the corresponding random motif strengths by at least 28 standard deviations, the p-value for these actual motif strengths can be much smaller than 0.01, due to the Chebyshev inequality (Abramowitz and Stegun, 1972). Note that for the MCB and RAP1 motifs, we also show their distribution and actual motif strength for both motif's orientations with "←" denotes 5'← and "→" denotes 5'→ orientations.

Supplementary 11: Mathematical notations

Throughout the text, we employ a set of mathematical notations and symbols. Here are their definitions.

$\vec{X}_{j\bullet}$: denotes the **j**$^{\mathbf{th}}$ row vector of a given matrix X

$\vec{X}_{\bullet j}$: denotes the **j**$^{\mathbf{th}}$ column vector of a given matrix X

X^T : denotes the transpose of a give matrix X

$\bullet$: represents the multiplication operation

$\vec{x}$: denotes a column vector.

Reference

Abramowitz, M. and Stegun, I.A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis.* Wiley, Chichester ; New York.

Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet*, **27**, 167-171.

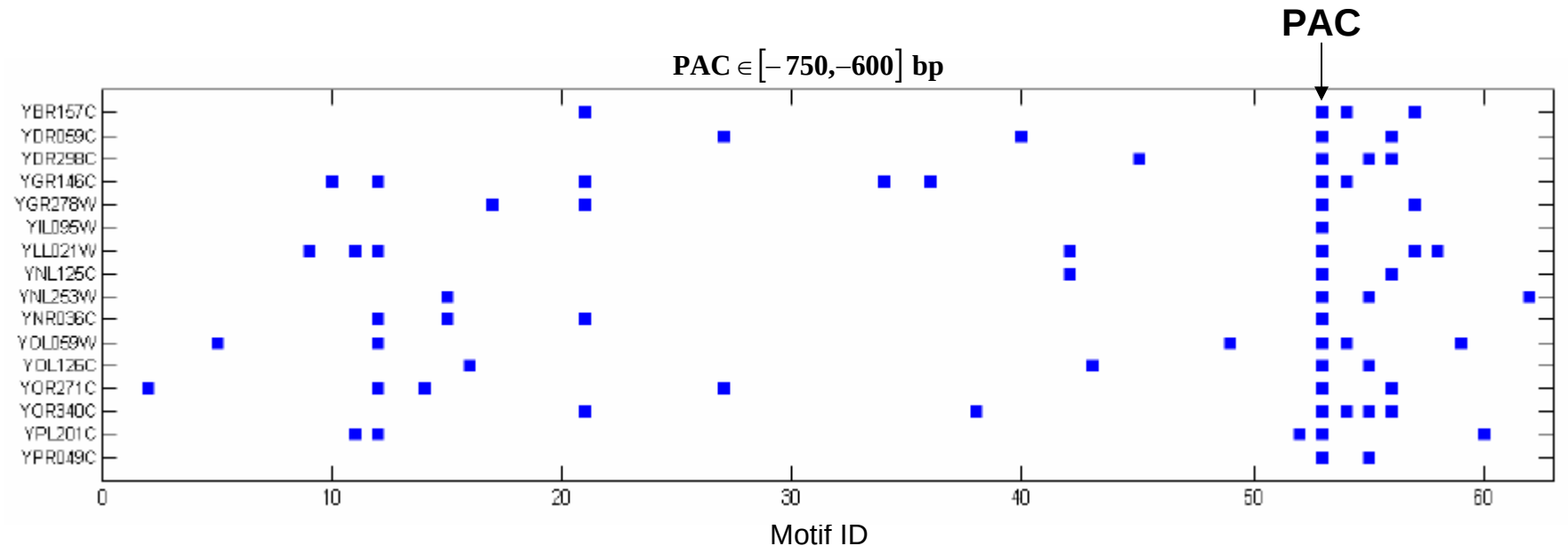
Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000) Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology*, **296**, 1205-1214.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241-254.

Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C. and Roychowdhury, V.P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*, **100**, 15522-15527.

Paatero, P., Hopke, P.K., Song, X. and Ramadan, Z. (2002) Understanding and controlling rotations in factor analytic models. *Chemometrics and Intelligent Laboratory Systems*, **60**, 253-264.

Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16-23.


PAC
PAC ∈ [−750,−600] bp
YBR157C
YDR059C
YDR298C
YGR146C
YGR278W
YIL095W
YLL021W
YNL125C
YNL253W
YNR036C
YOL059W
YOL126C
YOR271C
YOR340C
YPL201C
YPR049C
0
10
20
30
40
50
60
Motif ID

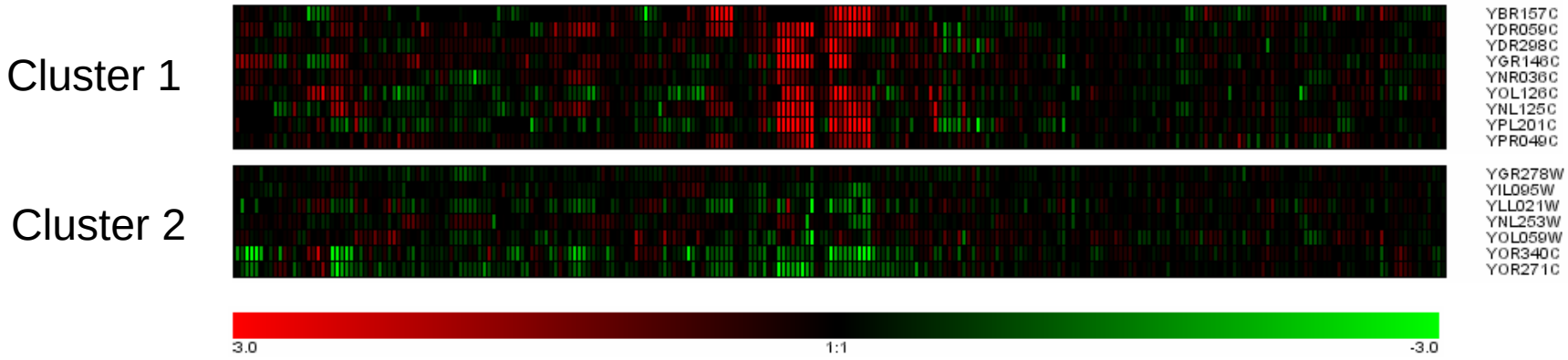

Cluster 1
Cluster 2
YBR157C
YDR059C
YDR298C
YGR146C
YNR036C
YOL126C
YNL125C
YPL201C
YPR049C
YGR278W
YIL095W
YLL021W
YNL253W
YOL059W
YOR340C
YOR271C
3.0
1:1
-3.0

Figure SF1a

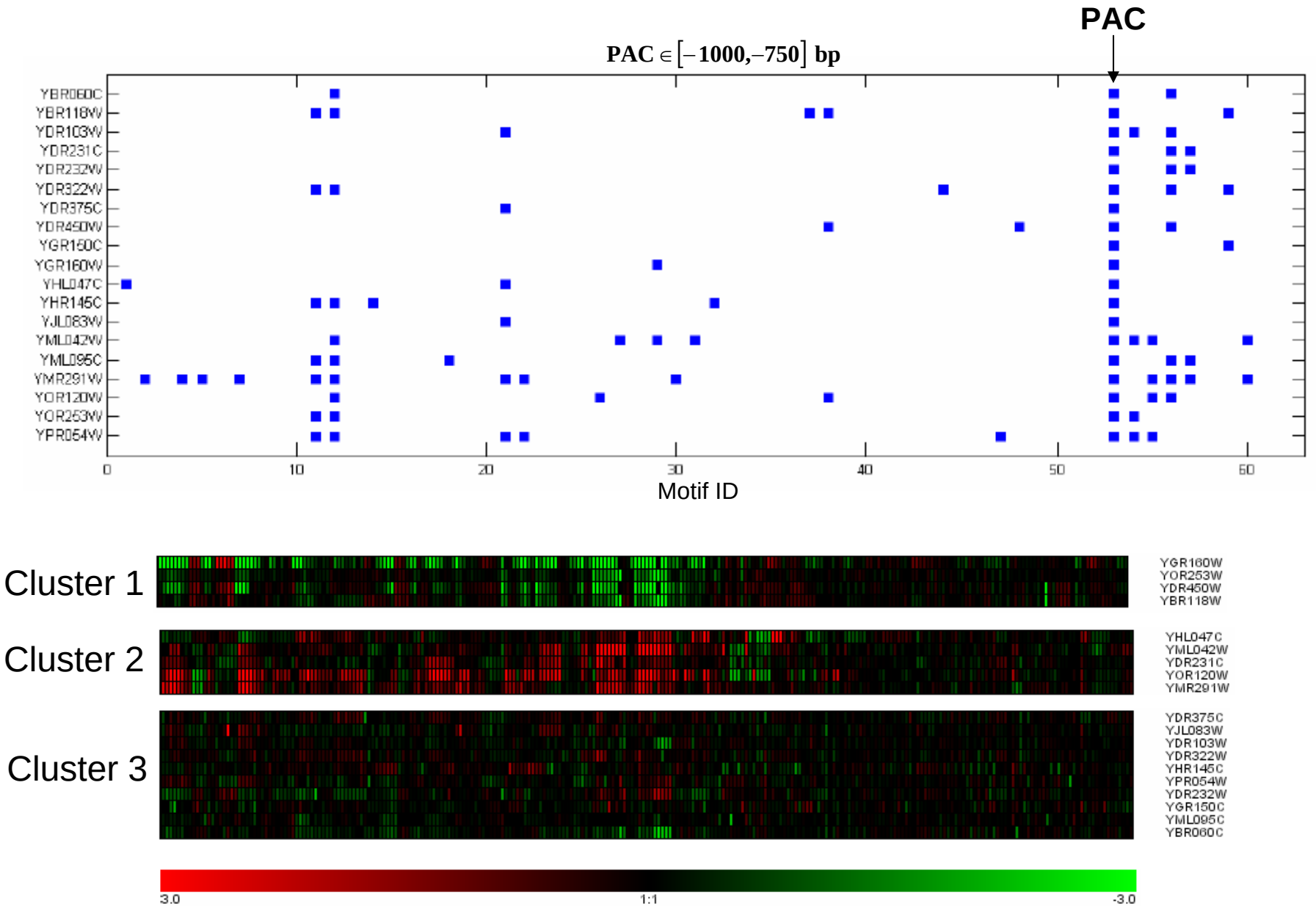

PAC
PAC ∈ [−1000,−750] bp
YBR060C
YBR118W
YDR103W
YDR231C
YDR232W
YDR322W
YDR375C
YDR450W
YGR150C
YGR160W
YHL047C
YHR145C
YJL083W
YML042W
YML095C
YMR291W
YOR120W
YOR253W
YPR054W
0
10
20
30
40
50
60
Motif ID
Cluster 1
Cluster 2
Cluster 3
YGR160W
YOR253W
YDR450W
YBR118W
YHL047C
YML042W
YDR231C
YOR120W
YMR291W
YDR375C
YJL083W
YDR103W
YDR322W
YHR145C
YPR054W
YDR232W
YGR150C
YML095C
YBR060C
3.0
1:1
-3.0

Figure SF1b

PAC $\in [-150, \mathbf{ATG}]$ **bp**

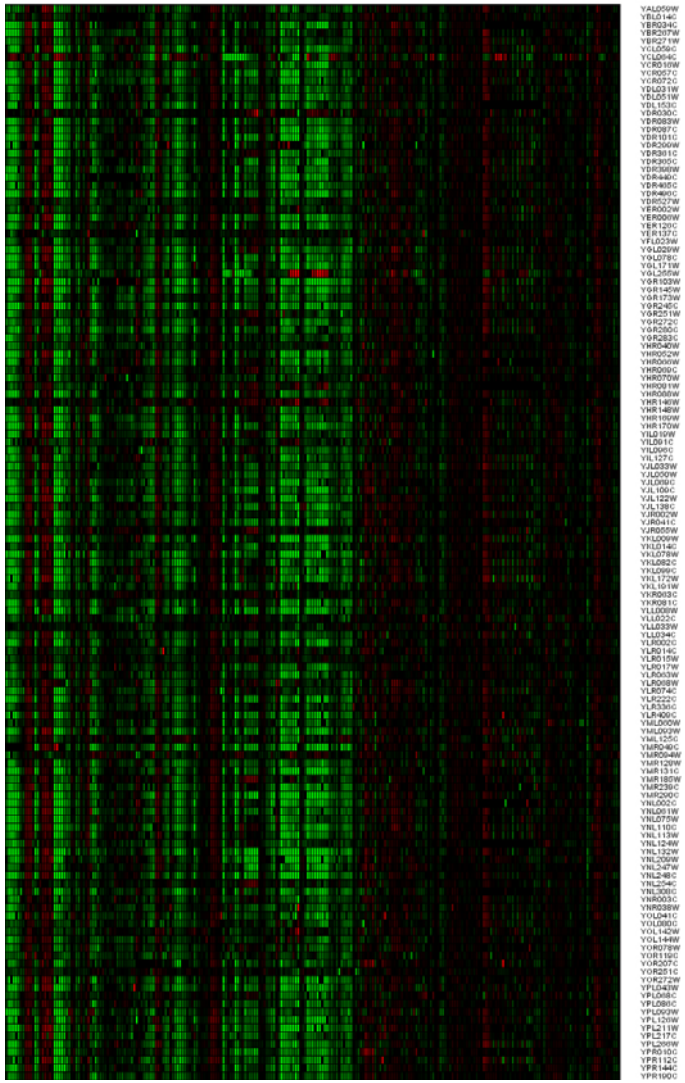
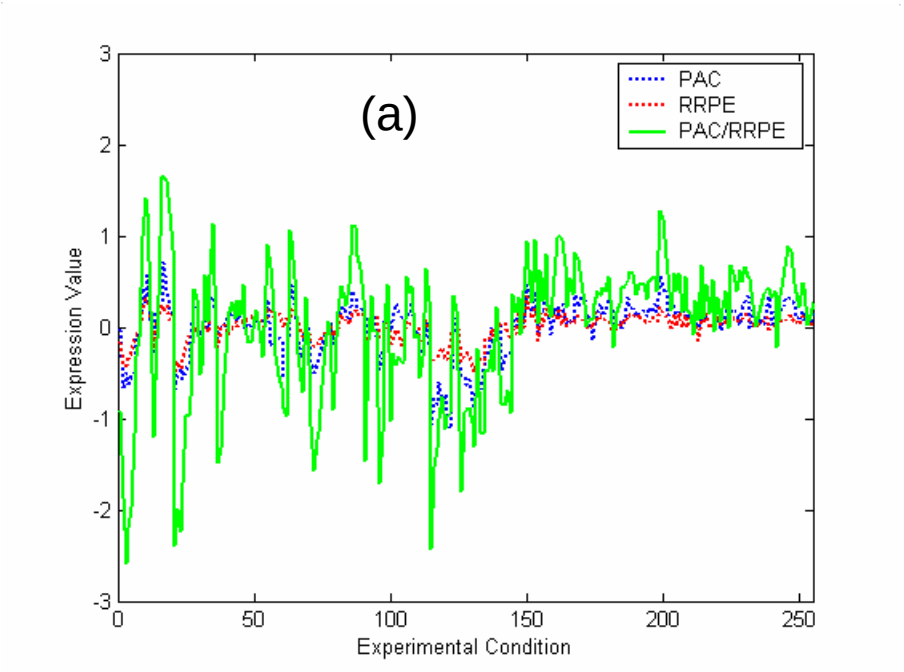


Figure SF1c


(a)
PAC
RRPE
PAC/RRPE
Expression Value
Experimental Condition
3
2
1
0
-1
-2
-3
0
50
100
150
200
250

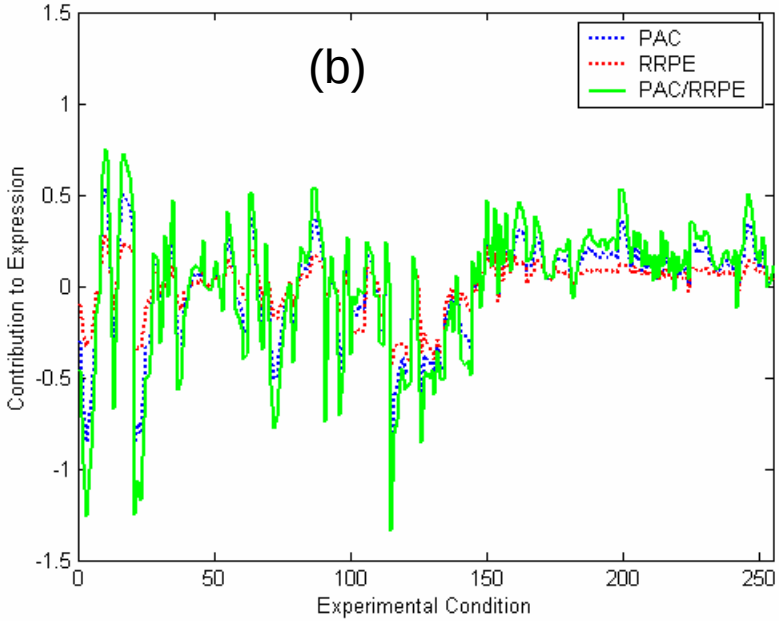

(b)
PAC
RRPE
PAC/RRPE
Contribution to Expression
Experimental Condition
1.5
1
0.5
0
-0.5
-1
-1.5
0
50
100
150
200
250

Figure SF2

(c)

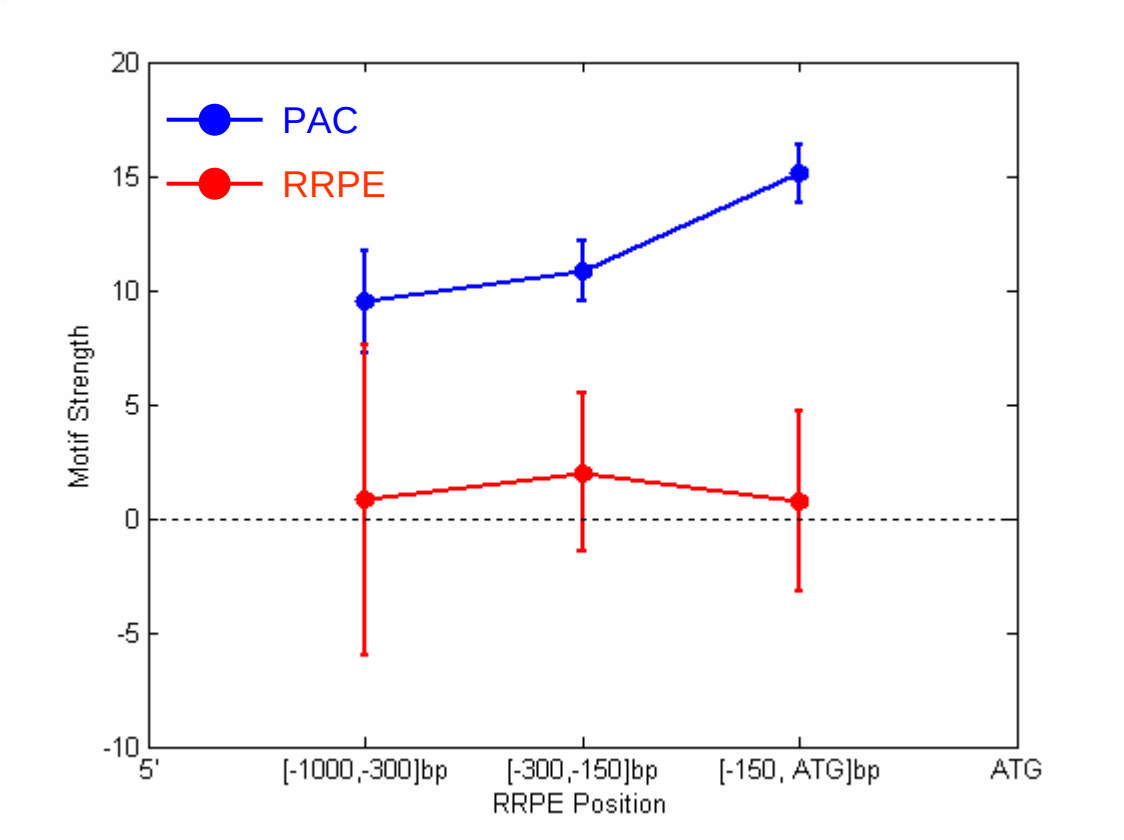


Figure SF2C

A

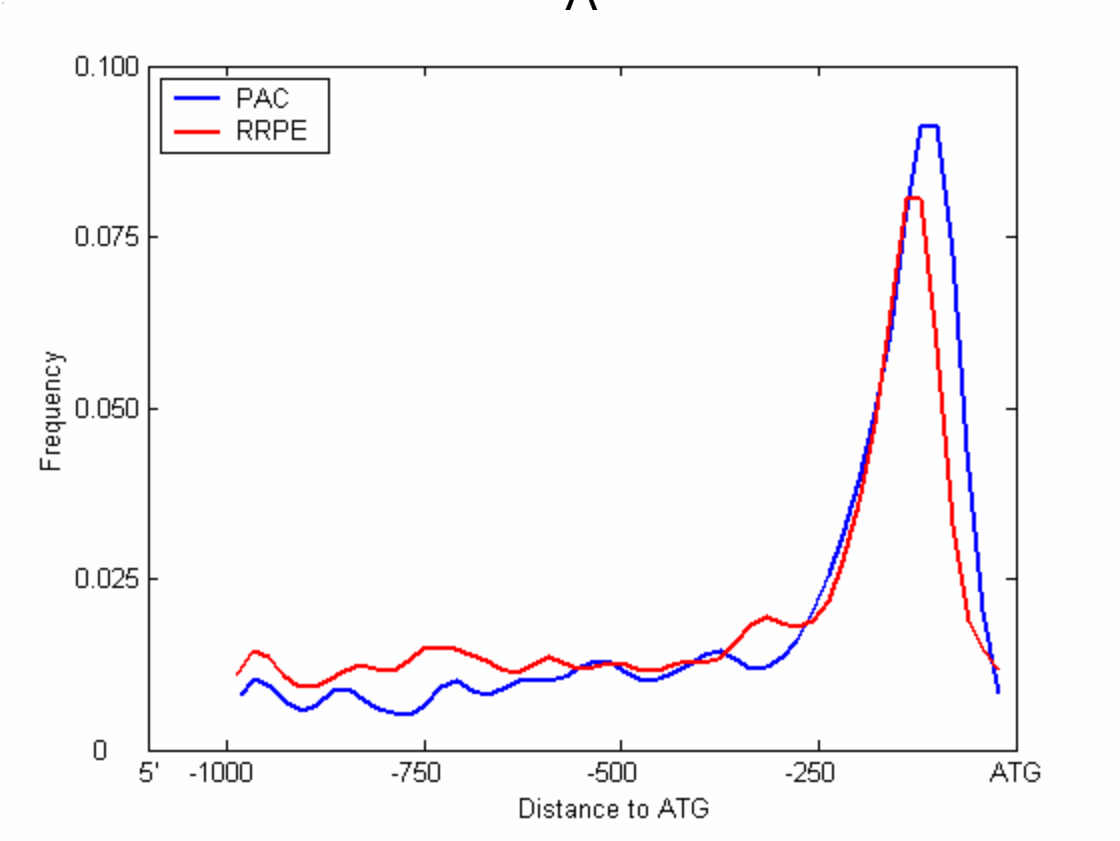


Figure SF3A

B

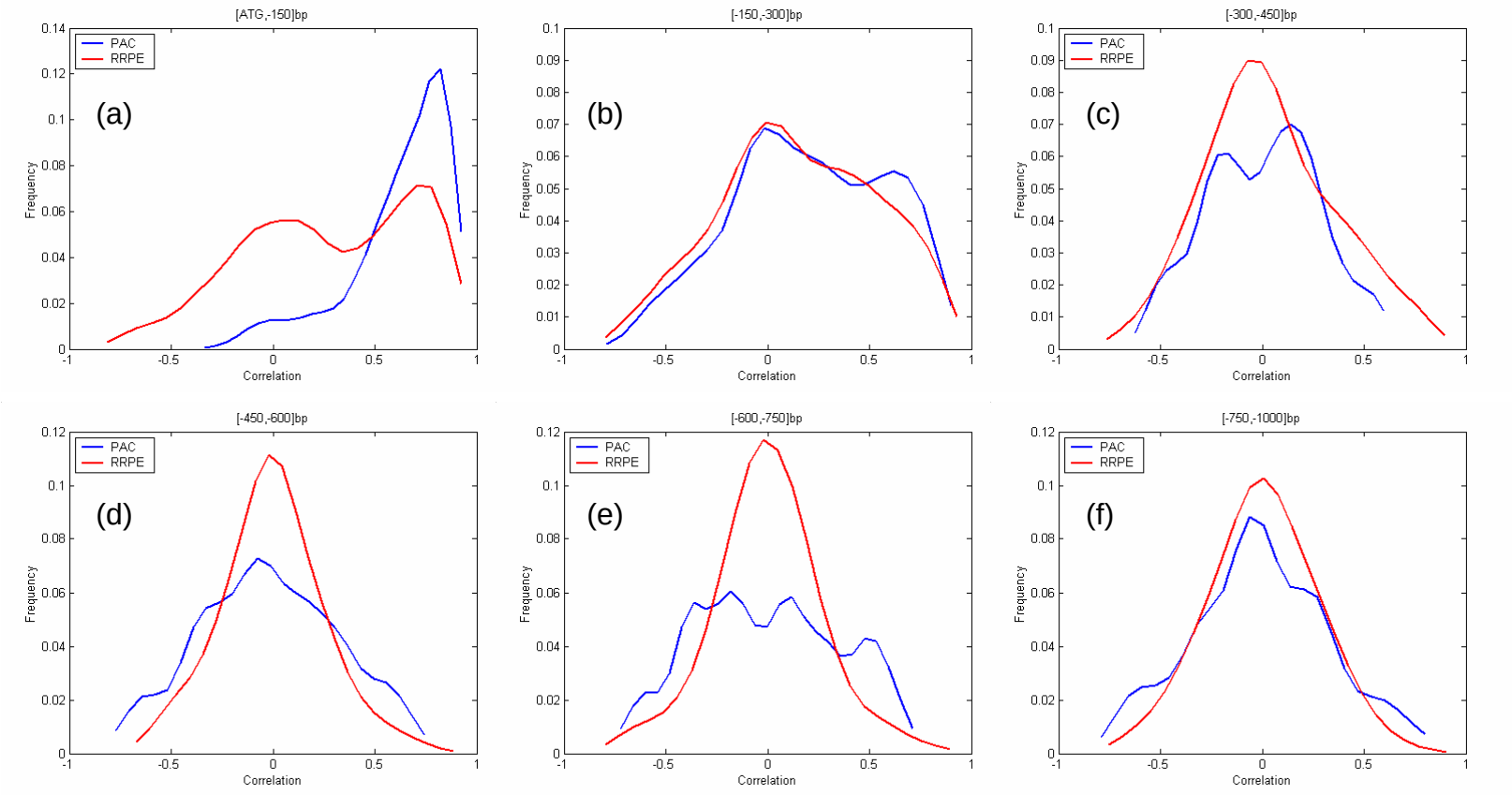


Figure SF3B

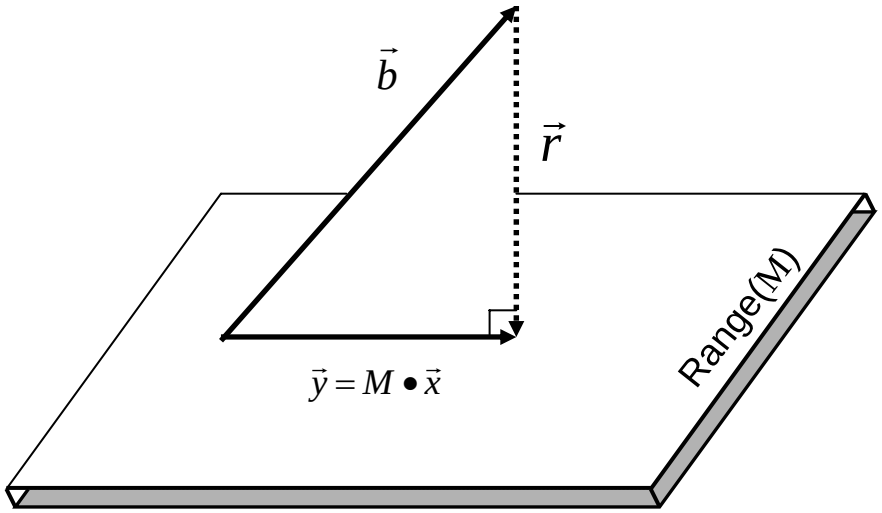


Figure SF4

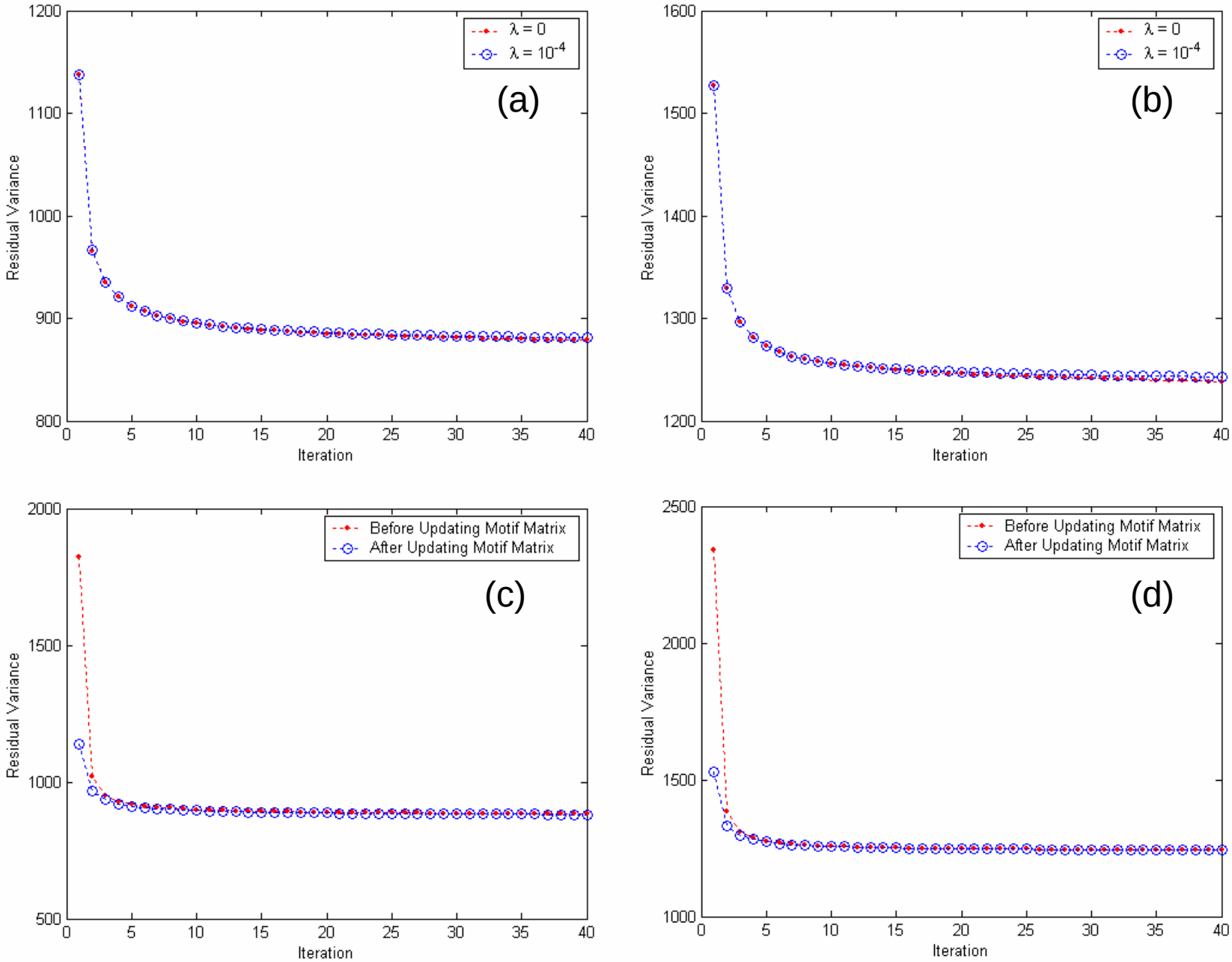

(a)
λ = 0
λ = 10^-4
Residual Variance
Iteration
(b)
λ = 0
λ = 10^-4
Residual Variance
Iteration
(c)
Before Updating Motif Matrix
After Updating Motif Matrix
Residual Variance
Iteration
(d)
Before Updating Motif Matrix
After Updating Motif Matrix
Residual Variance
Iteration

Figure SF5

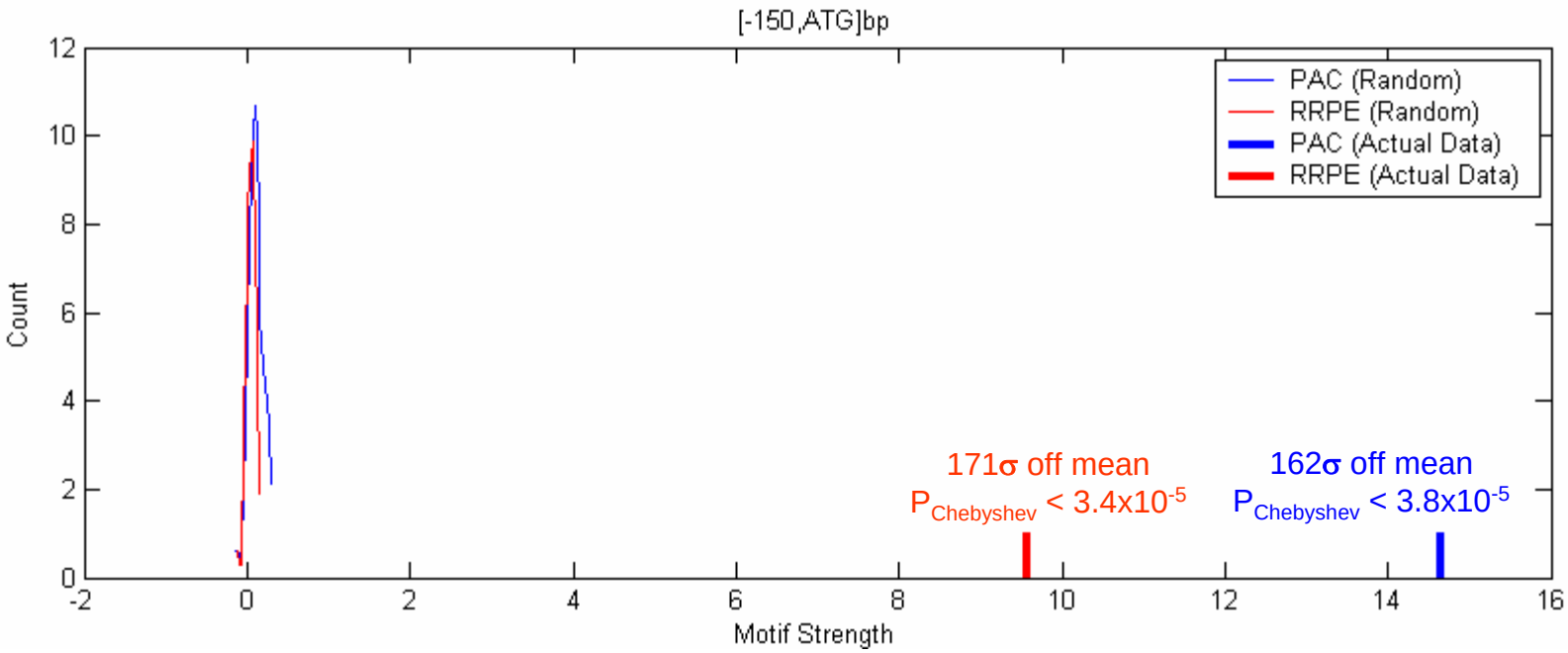

[-150,ATG]bp
PAC (Random)
RRPE (Random)
PAC (Actual Data)
RRPE (Actual Data)
171σ off mean
P Chebyshev < 3.4x10-5
162σ off mean
P Chebyshev < 3.8x10-5
Count
Motif Strength
0
2
4
6
8
10
12
-2
0
2
4
6
8
10
12
14
16

Figure SF6a

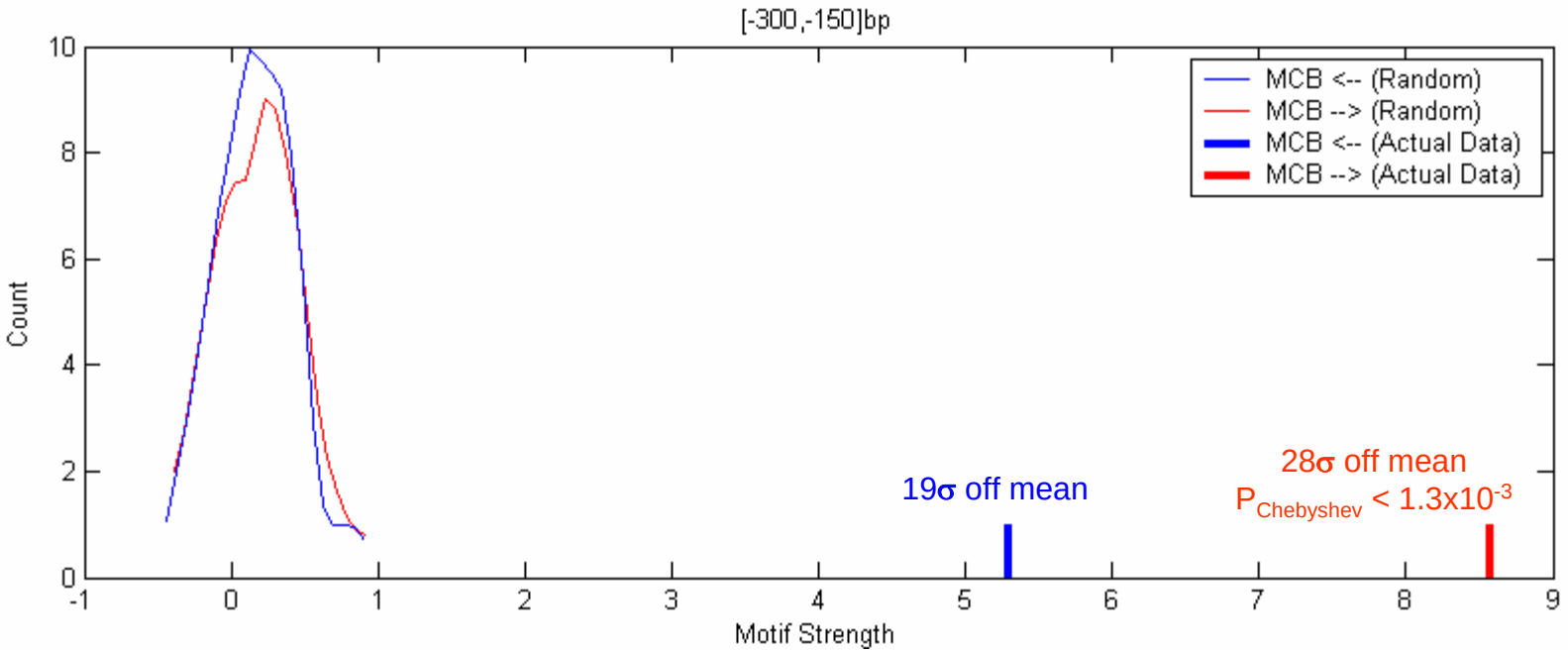

[-300,-150]bp
MCB <-- (Random)
MCB --> (Random)
MCB <-- (Actual Data)
MCB --> (Actual Data)
19σ off mean
28σ off mean
P Chebyshev < 1.3x10-3
Count
Motif Strength
0
2
4
6
8
10
-1
0
1
2
3
4
5
6
7
8
9

Figure SF6b

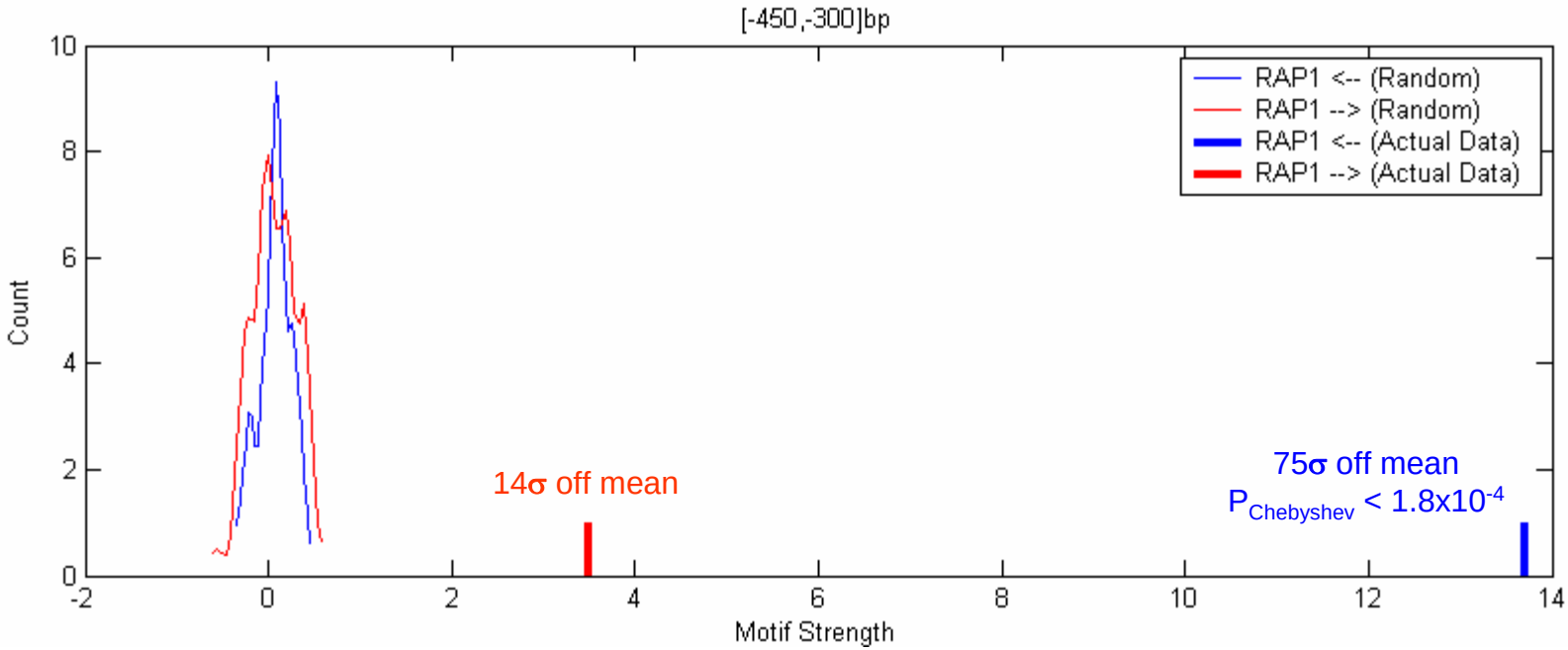

[-450,-300]bp
RAP1 <-- (Random)
RAP1 --> (Random)
RAP1 <-- (Actual Data)
RAP1 --> (Actual Data)
14σ off mean
75σ off mean
P Chebyshev < 1.8x10-4
Count
Motif Strength
0
2
4
6
8
10
-2
0
2
4
6
8
10
12
14

Figure SF6c