Journal club report

Paper: 7

Members: Čerňanský Andrej, Drobná Alica, Hamidová Barbora, Holenda Slavomír, Jakubovský Erik Róbert Ján, Kravec Marián, Martyrosian Anna Oleksandrivna

Transcriptional regulatory code of a eukaryotic genome

Introduction

Regulation of gene expression is considered to be one of the greatest scientific mysteries that remains to be solved. Differences in the subset of genes that are being turned on or off as well as the timing of it represent the key to understanding many complicated processes, including cell differentiation, resistance to stress, adaptation to changes in the environment, cell proliferation and senescence.

Compared to the prokaryotic genome, the genome of eukaryotic organisms is way larger and more complex. It consists of linear chromosomes, the number of which is species-specific, that are enveloped in the nucleus. Therefore, the mechanisms that underlie elaborate control of genetic machinery in eukaryotes are highly complicated and sometimes poorly described. One of the major control levels of gene expression is regulation of transcriptional activity.

Transcription is the process of RNA molecules synthesis, the sequence of which is encoded in the DNA. This reaction is catalysed by the enzyme RNA polymerase. However, RNA polymerase cannot accomplish the synthesis on its own, therefore the presence of other proteins is required. Transcription factors are *trans*-elements that assist RNA polymerase during transcription and are important for the assembling of the transcriptional complex. Their main function is to bind to specific sites in the sequence of DNA (*cis*-elements) and initiate the transcription by navigating RNA polymerase as well as other subunits to regulatory regions in promoters of specific genes. The combination of transcription factors that act in certain cells determines what genes should be activated under current circumstances. The function of transcription factors is depicted in Figure 1.

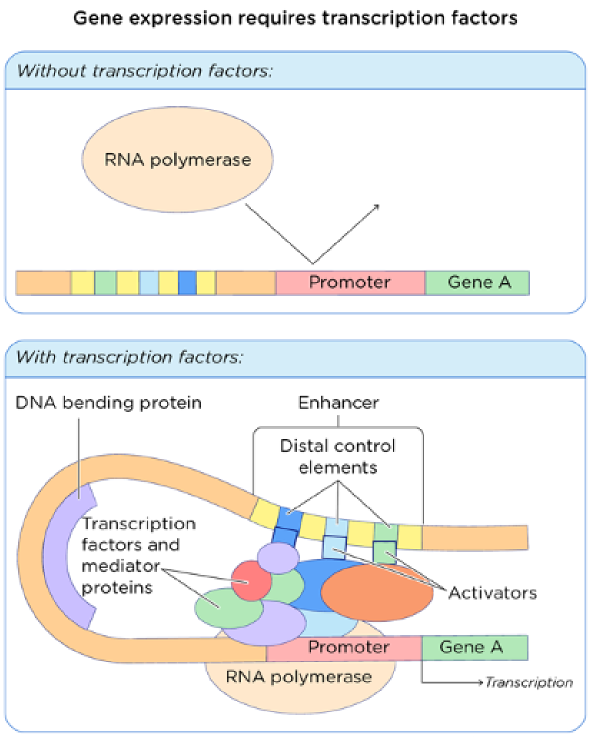

Gene expression requires transcription factors
Without transcription factors:
RNA polymerase
Promoter
Gene A
With transcription factors:
DNA bending protein
Enhancer
Distal control elements
Transcription factors and mediator proteins
Activators
Promoter
Gene A
RNA polymerase
Transcription

Fig. 1. The role of transcription factors in gene expression. Modified by Jack Westin.

Harbison *et al.* in the article "Transcriptional regulatory code of a eukaryotic genome" (2004) were trying to better understand the nature of interaction between *cis*-elements and regulators by constructing an initial map of yeast's transcriptional regulatory code. The basis of this map was to identify the elements in the DNA that are bound by transcription factors in various environments. By combining different methods, the authors managed to divide specific regulators into several groups according to the conditions under which binding occurs. They also successfully updated the current knowledge possess about transcription factors as well as the sequence specificities of their binding sites. Model organisms chosen for the research were different species of budding yeast *Saccharomyces* due to relative simplicity of its genome and ease of manipulation. Nevertheless, the authors claim that the approaches they used in yeast can be applicable to map regulatory sequences in higher eukaryotes.

Generally, the more important role the sequence plays in the genome, the more conserved it is among related species. This allowed the authors to use modern biological techniques in combination with bioinformatic methods, for example clustering, statistical tests, and phylogenetic analysis, which will be discussed further. One of the most popular biological methods, widely used to detect DNA-binding sites of transcription factors, is chromatin immunoprecipitation or ChIP. Proteins that were bound to specific regions in DNA were crosslinked with formaldehyde in vivo. The lysis of the cells was performed later, and the DNA was fragmented using ultrasound. Those fragments that were bound with transcription factors were protected against shearing. Crosslinked proteins with DNA sequences were purified using immunoprecipitation with corresponding antibody and reverse crosslinking unbound proteins from the sequence. Then samples were amplified and fluorescently labelled with ligation-mediated PCR. Lastly, the samples were hybridised to a microarray containing amplified fragments of yeast's DNA which consisted of intragenic regions and could possibly be binding sites of certain transcription factors. The results were analysed by the change of the sample's colour in case of signals overlapping.

Overview of the methods

The study employed a genome-wide location analysis to comprehensively investigate the transcriptional regulatory code within a eukaryotic genome, with a specific focus on the yeast genome. This analysis involved integrating a Myc-epitope coding sequence into the endogenous gene of 203 DNA-binding transcriptional regulators, enabling the tracking of their genomic occupancy in different growth environments. Through formaldehyde crosslinking, immunoprecipitation, and microarray-based analysis of enriched DNA regions, the researchers identified 11,000 unique interactions between regulators and promoter regions with high confidence ($P \leq 0.001$). This genome-wide location analysis provided crucial data on the binding behaviour of regulators under various conditions.

The research categorised regulators into distinct groups based on their binding behaviour. For instance, Condition-invariant regulators exhibited consistent binding to the same set of promoters in different growth environments, highlighting their stability across conditions. On the other hand, Condition-enabled regulators displayed detectable binding only in specific conditions, suggesting an environment-dependent regulatory role. The distribution of binding sites within yeast promoters was analysed, revealing non-uniform patterns. Binding sites were found to have a sharply peaked distribution, with the majority located between 100 and 500 base pairs upstream of protein-coding sequences. This information provided insights into the spatial organisation of regulatory elements within promoters.

The constructed map of the transcriptional regulatory code included 3,353 interactions within 1,296 promoter regions, outlining the regulatory potential embedded in the genome. The study leveraged genome-wide location analysis to identify *cis*-regulatory sequences, providing a foundation for understanding the regulatory mechanisms governing gene expression. The dynamic nature of regulator binding under different conditions was a key aspect revealed by this genome-wide approach, shedding light on the complexity of the transcriptional regulatory network in the yeast genome.

Raw Data Analysis:

The microarray images were subjected to scanning using an Axon200B scanner, and subsequent analysis was performed using Genepix 5.0. The columns corresponding to background-subtracted intensities and the standard deviation of the background were isolated for further examination. The intensities from the two channels, representing immunoprecipitated (test) and unenriched (control) samples, underwent normalisation. This normalisation involved utilising the median of each channel to calculate a normalisation factor, thereby aligning all datasets to a uniform median intensity. The log ratio of the test channel intensity to the control channel was then computed.

To address potential biases in the immunoprecipitation reaction, these log ratios underwent further normalisation for each spot by subtracting the average log ratio of that spot across all arrays. Subsequently, the intensities in the test channel were adjusted to achieve normalised ratio. An error model was employed to determine the significance of enrichment on each chip, and data from replicates were combined to generate a final average ratio and significance of enrichment for each intergenic region. Each intergenic region was then associated with the genes it is most likely to regulate.

Step 1: Initial Motif Discovery:

CONVERGE:

The authors designed CONVERGE to identify motifs that are simultaneously over-represented in a designated set of input sequences and conserved across multiple genomes. In the CONVERGE approach, the input sequences encompass an ungapped DNA sequence corresponding to the primary genome, accompanied by one or more optional aligned sequences that may contain gaps. The algorithm is rooted in the ZOOPS model of MEME and integrates a 5th-order Markov background model.

Diverging from MEME's strategy of searching for matches to a motif model across a set of input sequences, CONVERGE conducts searches within the multiple-sequence alignments for each sequence. Specifically, CONVERGE computes the probability of a motif occurring at a site in the alignment as the product of the probabilities of the motif occurring at the same site in each of the aligned sequences. Consequently, CONVERGE defines a site as conserved flexibly, contingent on the particular motif under investigation.

Probe Sequences:

Motif discovery programs were employed to analyse the sequences of probes that exhibited binding with a P-value of ≤ 0.001. During this analysis, it was observed that certain intergenic regions displayed extensive homology across their entire length, introducing a potential bias in motif discovery as all subsequences appeared to be overrepresented. To mitigate this bias, the researchers utilised BLAST19 to identify pairs of probes demonstrating high sequence similarity covering over 50% of their lengths. For each identified pair, the shorter intergenic region was excluded from the motif discovery computations. This corrective measure resulted in the removal of up to nine regions in some experiments, with an average removal of less than one region.

To ascertain the sequences represented on the microarrays, the researchers computed the expected products of the polymerase chain reaction (PCR) used in constructing the arrays. Primer sequences were obtained from http://www.resgen.com/products/YeIRP.php3, and the March 2002 revision of the yeast genome was sourced from SGD. Probes predicted to amplify more than two different genomic sequences were excluded from the calculations. Additionally, twenty-five probe sequences neighbouring repetitive, non-transcribed features such as telomeric repeats, X elements, and Y' elements were omitted from further considerations.

PSSM Representation:

Motifs identified by all programs underwent standardisation into a uniform position-specific scoring matrix (PSSM) for subsequent analysis. In the case of AlignACE and MDscan, which generate alignments of binding sites, these alignments were initially transformed into matrices representing the frequency of each base (A, C, G, T) at each position within the alignments. Kellis *et al.*'s method represents motifs as text strings containing ambiguity codes, and these were also converted into matrices of frequencies. For instance, if a motif featured the letter “S” at a specific position, a value of 0.5 would be assigned to both “C” and “G.”

The matrices of base frequencies were then converted to probabilities and subsequently adjusted with 0.001 pseudo-counts in proportion to the 0th-order background probabilities (3.1x10-4 pseudocounts for A and T, 1.9x10-4 pseudocounts for G and C). Log-likelihood scores were computed by dividing the estimated probabilities by the background probability for each letter and calculating the base-2 logarithm. Probability matrices directly provided by CONVERGE and MEME were used in the analysis without additional transformation.

MEME:

MEME stands for Multiple Expectation Maximisation for Motif Elicitation and it is an algorithm based on the Expectation-maximisation algorithm. The algorithm consists of two main steps, which are repeated until convergence. First given some model parameters we compute the expected position of the motif region. In practice, we have some probability matrix θ based on some observation which tells us the probability of base b occurring on i-th position. Given the background probabilities θ_0 we calculate the likelihood ratio of each k-mer in sequence.

$$LR(kmer_i) = \prod_{j=1}^{k} \frac{p(b_{i,j} \mid \theta)}{p(b_{i,j} \mid \theta_0)}$$

In the second step based on the newly found region, we update the model parameters, so we maximise the probability of observing this region. Since now we know the likelihood of each k-mer, to derive new probabilities, we simply create a frequency matrix of occurrences of each base at each position in k-mer, however, we scale each base with corresponding likelihood.

$$\theta_{i,b} = \frac{\sum_{j=1}^{n} LR(k - mer_j) S(i, b, k - mer_j)}{\sum_{j=1}^{n} LR(k - mer_j)}$$

where S (i, b, k-mer$_j$) is 1 when the i-th position in k-mer$_i$ is base b, 0 otherwise. When the matrices in two succeeding steps do not change the algorithm has converged and a motif was discovered. In the case of MEME the initial probability matrix is computed iteratively for each k-mer and run for two iterations. The most promising matrices are selected as initial matrices and run toward convergence.

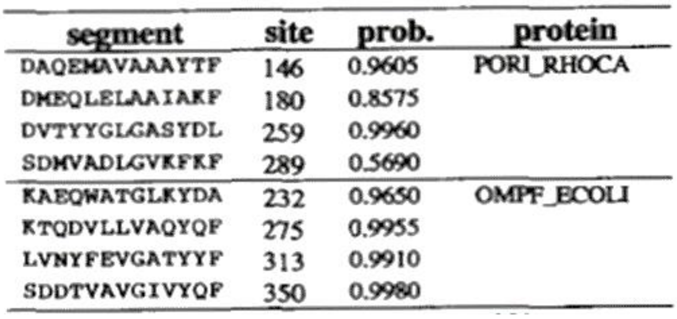
AlignACE:

AlignACE, which is abbreviation for Alignus Nucleic Acid Conserved Elements, is a program for detecting motifs in input sequences based on Gibbs sampling strategy. Gibbs sampling strategies are employed in cases when direct sampling from multivariate probability distribution is difficult and sampling from conditional distribution is easy to sample from. The overall algorithm partitions input sequences into non-overlapping regions each corresponding to a specific motif model, including the null model. For each model the algorithm stores two evolving data structures. One for alignment of related sequences and one for storing the observed target probabilities. The target probabilities are computed as follows:

$$q_{i,r} = \frac{c_{i,r} + b_r}{c + b}$$

where $c_{i,\,r}$ is the number of bases r at column i, c is number of alignments for concrete motif, b_r is number of pseudocounts for base r and b is sum of all b_r.

Tab. 1. AlignACE



segment	site	prob.	protein
DAQEMAVAAAYTF	146	0.9605	PORI_RHOCA
DMEQLELAAIAKF	180	0.8575	
DVTYYGLGASYDL	259	0.9960	
SDMVADLGVKFKF	289	0.5690	
KAEQWATGLKYDA	232	0.9650	OMPF_ECOLI
KTQDVLLVAQYQF	275	0.9955	
LVNYFEVGATYYF	313	0.9910	
SDDTVAVGIVYQF	350	0.9980	

In the beginning we select a random location of motif for each sequence. In the first step, the algorithm removes one sequence and computes a new probability matrix. In the second step, based on this matrix, we score every k-mer in sequence with likelihood ratio. Based on this ratio, we sample from k-mers and select a new motif region. These steps are performed for each sequence until convergence.

MDscan:

For ChIP-array experiments MDscan scrutinises top sequences to propose motif candidates. For each candidate, the program finds w-mers, which match with at least m base pairs with the seed (w-mer of motif). For each seed, a motif weight matrix is computed. After evaluating the matrix and computing the scores of all w-mer motifs, the top 10-50 seed motifs are selected for the next step. In the next step, retained matrices are used to find new w-mers in the remaining sequences. W-mer is only added if the motif score of the matrix increases. A segment will be removed from the matrix if the motif score increases. The process is iterative and generally converges after 10 iterations.

Step 2: Motif Scoring and Significance Testing:

In the second step of an experiment they tested the significance of motifs by comparison of their occurrence in bound and unbound probes. They used 3 approaches. Each approach used a different metric.

Enrichment:

This metric measures occurrence of motif in bound probes, compared to all possible gene targets. Downside is that it does not take into account how many motifs occur within each intergenic region. For a sequence to be considered containing a motif, there must be sites scoring at least 70% of the maximum possible score of the matrix. The P-value has been calculated accordingly:

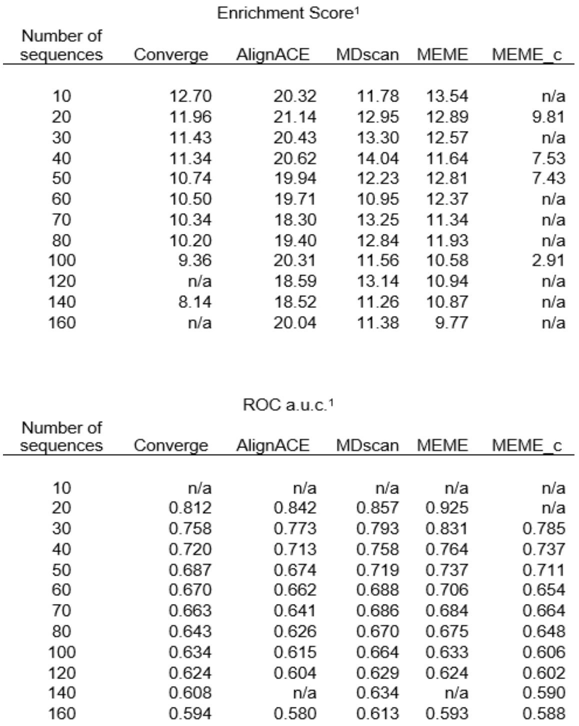
$$\boldsymbol{p} = \sum_{i=b}^{min(B,g)} \frac{\frac{BG-B}{i\ g-i}}{\frac{G}{g}}$$

where "*B* is the number of bound intergenic regions and *G* is the total number of intergenic regions represented on the microarray (or the genome). The quantities b and g represent the number of intergenic regions of *B* and *G* matching the motif." Then the enrichment score is evaluated as $-log_{10}(p)$.

ROC AUC:

To get this Receiver Operating Characteristic Area Under Curve (ROC AUC), they first ranked the sets of bound and unbound probes by the number of motif matches and plotted them against each other. They then calculated the area under the curve.

Tab. 2. Enrichment and ROC AUC scores.



Enrichment Score[1]

Number of sequences	Converge	AlignACE	MDscan	MEME	MEME_c
10	12.70	20.32	11.78	13.54	n/a
20	11.96	21.14	12.95	12.89	9.81
30	11.43	20.43	13.30	12.57	n/a
40	11.34	20.62	14.04	11.64	7.53
50	10.74	19.94	12.23	12.81	7.43
60	10.50	19.71	10.95	12.37	n/a
70	10.34	18.30	13.25	11.34	n/a
80	10.20	19.40	12.84	11.93	n/a
100	9.36	20.31	11.56	10.58	2.91
120	n/a	18.59	13.14	10.94	n/a
140	8.14	18.52	11.26	10.87	n/a
160	n/a	20.04	11.38	9.77	n/a

ROC a.u.c.[1]

Number of sequences	Converge	AlignACE	MDscan	MEME	MEME_c
10	n/a	n/a	n/a	n/a	n/a
20	0.812	0.842	0.857	0.925	n/a
30	0.758	0.773	0.793	0.831	0.785
40	0.720	0.713	0.758	0.764	0.737
50	0.687	0.674	0.719	0.737	0.711
60	0.670	0.662	0.688	0.706	0.654
70	0.663	0.641	0.686	0.684	0.664
80	0.643	0.626	0.670	0.675	0.648
100	0.634	0.615	0.664	0.633	0.606
120	0.624	0.604	0.629	0.624	0.602
140	0.608	n/a	0.634	n/a	0.590
160	0.594	0.580	0.613	0.593	0.588

Conservation CC4:

This metric was used for motifs discovered by Kellis method. It compares occurrence of conserved motif to the expected ratio that was observed in the same bound probes set of 3 gap-3 motifs. The binomial probability was calculated from this ratio. Additionally this metric can be reported as equivalent to z-score.

Significance Thresholds:

While generating with discovery programs, they noticed that the procedure created motifs with high Enrichment and ROC AUC metrics. This persisted even if only random intergenic regions were generated. To counterbalance, they converted the scores from each metric into empirical probability that a motif with similar score could be found in a random sequence.

Only motifs with P-value ≤ 0.001 were accepted, in an attempt to minimise false positives. After thresholding, an improvement of identification was empirically observed. The value was estimated by running each program 50 times on randomly selected sequences on sets with 10, 20, 30, 40, 50, 60, 70, 80, 100, 120, 140, and 160 probes.

Empirical random runs were parameterised by a normal distribution. The critical values that give a P-value of 0.001 are provided in Table8. If the observed distribution was not normal, the metric was not used for evaluation for that specific setup ("n/a" in table). The CC4 metric is missing from the table because the threshold (4.95) was not dependent on the number of sequences.

Each experiment had a threshold derived from a randomisation set that was closest to the probe sequences in size. For example, AlignACE had an enrichment score of 25 after 10 runs on 23 intergenic sequences. The score distribution was calculated after 10 runs of AlignACE on each of 50 random sets of 30 intergenic sequences. The enrichment score corresponding to a P value ≤ 0.001 was 20.43. The mean enrichment score was 14.1 and the std was 2.1. The score of the candidate motif was higher, so it was considered significant.

Step 3: Motif Clustering and Averaging:

In the third step, at first, they clustered the motifs with K-medoids algorithm using custom distance.

K-medoids:

This method was used to divide all motifs into k clusters (groups), by choosing k motifs as centres, while minimising the overall distance between centre and non-centre motifs. It is NP-hard to solve exactly, so there are many ways to choose the right centres, even the cited source contains several optimisation techniques. One of them is to exchange each centre with a non-centre motif and then selecting the exchange that produces the greatest reduction.

The primary benefit is the ability to use custom distance and also the algorithm is more robust to noise and outliers compared to for example similar K-means algorithm. However, a drawback of this algorithm is its higher computational expense.

The optimisation step was repeated 500 times. They found the best number of clusters by incrementally decreasing from number 10, until the average distances between members of a cluster and medoids of other clusters were not ≥ 0.18.

Inter-Motif distance:

As previously mentioned, a custom metric was constructed to efficiently measure the distance between motifs. Distance between motifs “a” and “b” is defined as

$$D(a,b) \ = \frac{1}{w}\sum_{i=1}^{w} \ \frac{1}{\sqrt{2}}\sum_{L\in\{ACGT\}} \ (a_{i,L} - b_{i,L})^{2}$$

where “*w* is the motif width, and $a_{i,L}$ and $b_{i,L}$ are the estimated probabilities of observing base *L* at position *i* of motifs *a* and *b*, respectively”. The metric was normalised by w and $\sqrt{2}$ so it can be interpreted as a percentage of the difference. Additionally they stated that in practice the optimal alignment of motifs is not known. Therefore they “use the minimum distance between motifs among all alignments in which the motifs overlap by at least seven bases, or when the motifs are shorter, by 2 bases fewer than the shortest motif length”. Also they included the reverse complements of the motifs.

Averaging:

After clustering, average probabilities were calculated from the matrix positions for every cluster. The average motif for each cluster was computed and shortened on the ends to remove low-information positions.

Step 4: Conservation Testing for Averaged Motifs:

They looked at the conservation of averaged motifs and then decided to focus on motives that satisfied two types of criteria. The first requirement was that the frequency of conserved instances of a motif must be at least as high within intergenic regions that are bound to the transcriptional factor as among all intergenic regions. They decided the second criterion to be that discovered motifs must have at least three conserved instances that are bound. If the aligned sequence of at least two other species also matched the motif, it could be called a “conserved instance”.

Step 5: Assignment of a Single Motif to Each Regulator:

On average, there were three significant motives assigned to the regulators. The motifs could explain the binding specificity of the protein or the binding could be influenced by other factors. They compared the results with databases such as YSM and Transfac but for 21 regulators there was no data available. In such cases, choosing the motif with the best enrichment score was the best option.

The motif from the database was used regarding the rest of the regulators, for which either:

- Less than ten intergenic regions were bound for effective motif discovery.
- Discovered motifs similar to the literature were eliminated by the conservation.
- None of the discovered motifs matched the literature.

These motifs were only included if they had at least one conserved instance that was bound. In such a way 102 found motifs were used in all subsequent analyses. Through further analysis, the authors were able to produce a map of active binding sites for 102 regulators in intergenic regions, they could create it when finding all conserved occurrences of each motif bound by the corresponding factor. A severe binding P-value threshold of $P \leq 0.001$ was used, and the definition of conservation as described earlier.

Promoter Classification:

One of the analyses conducted involved classification, with two distinct classifications performed. In the first classification they were classifying promoters based on arrangements of DNA binding sites within them. They found four types of arrangement; we can see their visualisation in Fig. 2.

The first arrangement is the most straightforward, featuring a single binding site. For instance, the motif for the Gcn4 regulator was identified independently within the binding sites of various genes.

The second arrangement consists of multiple binding sequences for a single regulator. They observed that certain regulators, such as Dig1 and Mbp1, exhibit statistically significant preferences for this type of arrangement. In their statistical analysis, the researchers tested the null hypothesis, assuming "The distribution for each regulator is the same as the average distribution for all regulators." Here, the term "distribution" refers to the ratio between the number of repetitive and non-repetitive arrangements. The obtained P-value for these two regulators was approximately 10^{-8}, leading to the rejection of the null hypothesis.

The third class consists of most "chaotic" arrangements, this class contains promoters with binding sites for different regulators. This suggests combinatorial regulation of genes. The authors expect that, in many cases, various regulators can be employed to generate different responses under varying growth conditions.

The fourth category is characterised by 'co-occurring' motifs. Promoters in this category harbour binding sites for pairs of specific regulators that appear together more frequently than would be expected by chance. They found 94 distinct pairs of regulators for which statistical testing with null hypothesis "Binding for those two regulators is independent" resulted in P-value smaller than 0.005 and thus the rejection of null hypothesis.

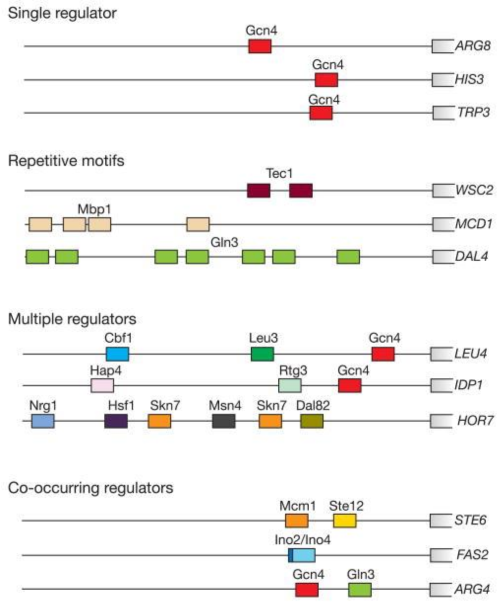


Fig. 2. The first classification of promoters based on arrangements of DNA binding sites.

The second categorisation they performed involves classifying regulators based on how their behaviour changes in various environments. They split all regulators into four categories shown in Fig. 3.

The first category is called 'Condition-invariant'. Regulators in this category bind the same set of promoters in two different environments. However, this does not necessarily imply that regulated genes are activated regardless of the environment. For instance, one of the extensively studied regulators in this category, Leu3, is essential for the activation of the genes it regulates. However, its sufficiency is not assured, as it requires the association of leucine metabolic precursor to transform Leu3 from a negative to a positive regulator. Several additional regulators in this category are recognised to exhibit similar characteristics, leading the authors to reasonably propose that activation or repression functions of some less-studied regulators in this category have requirements in addition to DNA binding.

The next category is called 'Condition-enabled'. These regulators do not exhibit binding in one environment but display considerable binding activity in another. In this class, they found, for example regulator Msn2, which is a well-studied regulator known to be excluded from the nucleus in a “calm” environment but accumulates when the cell is subjected to stress. They suggest that a lot of regulators in this category are similarly regulated by nuclear exclusion.

The 'Condition-expanded' category consists of regulators that in one environment bind to additional promoters compared to the other environment. For instance, in a nutrient-limiting environment, the levels of the Gcn4 regulator increase sixfold, resulting in its enhanced binding to a greater number of sites compared to nutrient-rich environments. According to the authors many other regulators in this category may exhibit similar behaviour, with their set of binding sites being regulated by their concentration in the nucleus.

The last category is 'Condition-altered' which means that the set of promoters changes between two environments. Regulator Ste12 is a great example of this behaviour, the set of promoters to which this regulator binds is dependent on interactions with other regulators whose concentrations vary across environments. That’s why authors suggest that this inter-regulator dependency might be the reason for altered promoters set for many of regulators in this category.

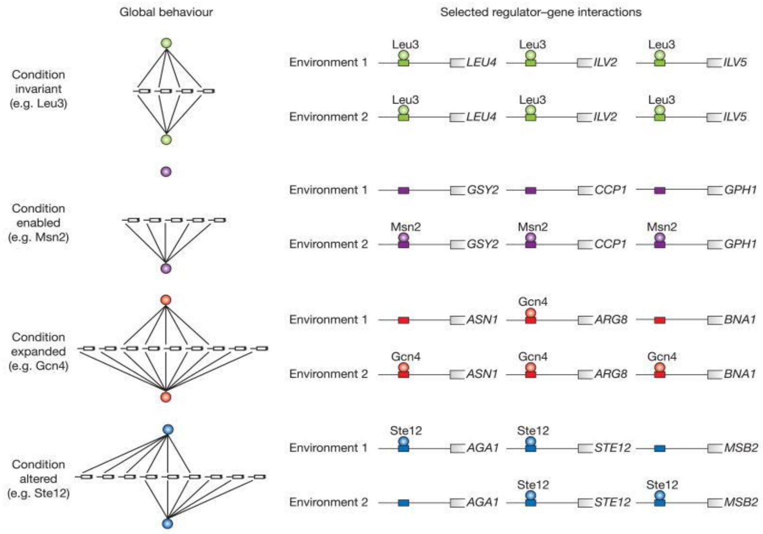

Global behaviour
Selected regulator–gene interactions
Condition invariant (e.g. Leu3)
Condition enabled (e.g. Msn2)
Condition expanded (e.g. Gcn4)
Condition altered (e.g. Ste12)
Environment 1
Environment 2
Leu3
LEU4
ILV2
ILV5
Msn2
GSY2
CCP1
GPH1
Gcn4
ASN1
ARG8
BNA1
Ste12
AGA1
STE12
MSB2

Fig. 3. The second classification of promoters based on their behaviour in various environments.

Experimental Confirmation of Predicted Specificity:

They compared the discovered motifs to those in the literature and selected the regulators for which the biggest difference was observed. For example, the newly discovered motif for Cin5 holds for P ≤ 10-38.4, while its previous motif showed the enrichment of only P ≤ 0.02 in the probes. This suggests a much stronger association between the newly discovered motif and Cin5.

Summary:

By using various biological and informatic tools, Harbison *et* al. were able to provide significantly more information about promoter cis-regulatory motifs in the yeast genome than was previously available.

Other approaches:

In a similar study, Yella and Vanaja investigated the prevalence of non-B DNA structural motifs in promoter regions in 1180 genomes belonging to 28 taxonomic groups in all domains of life. Non-B DNA involves every DNA structure other than canonical and the most common B-DNA form, such as curved DNA, cruciform DNA, G-quadruplex, triple-helical DNA, slipped DNA structures, and Z-DNA, all shown in Fig. 4. For the first time, this study searches for these specific DNA forms that contain promoter motifs.

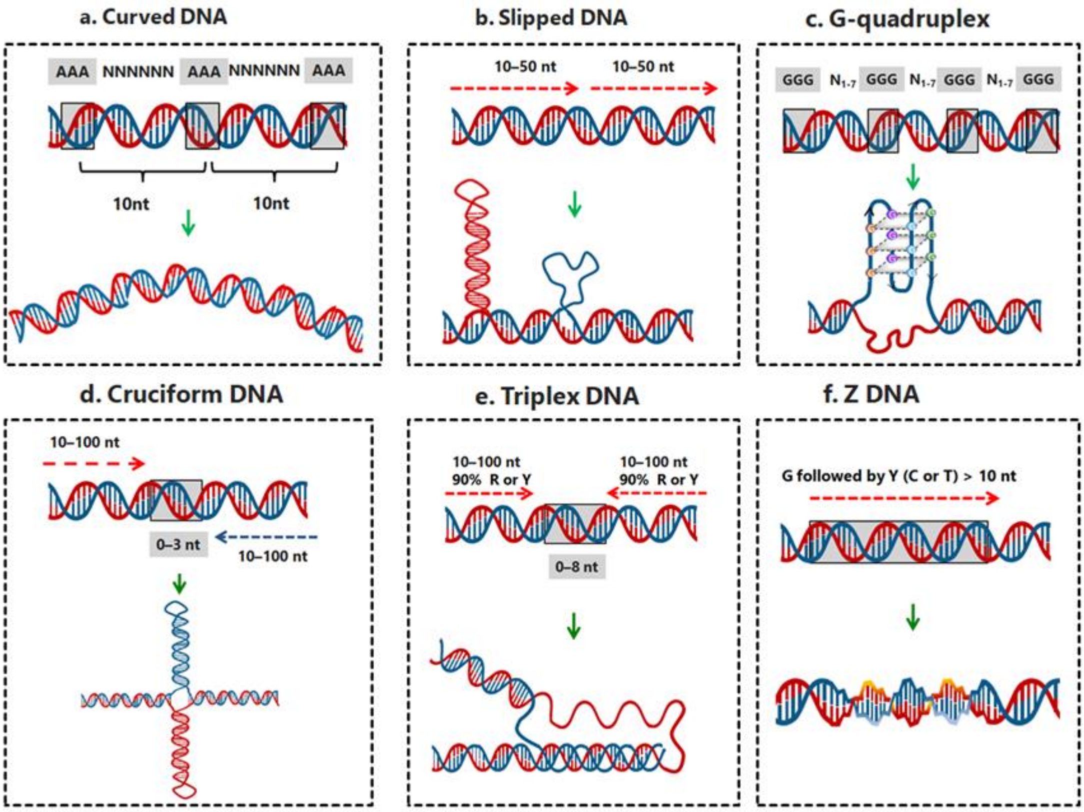
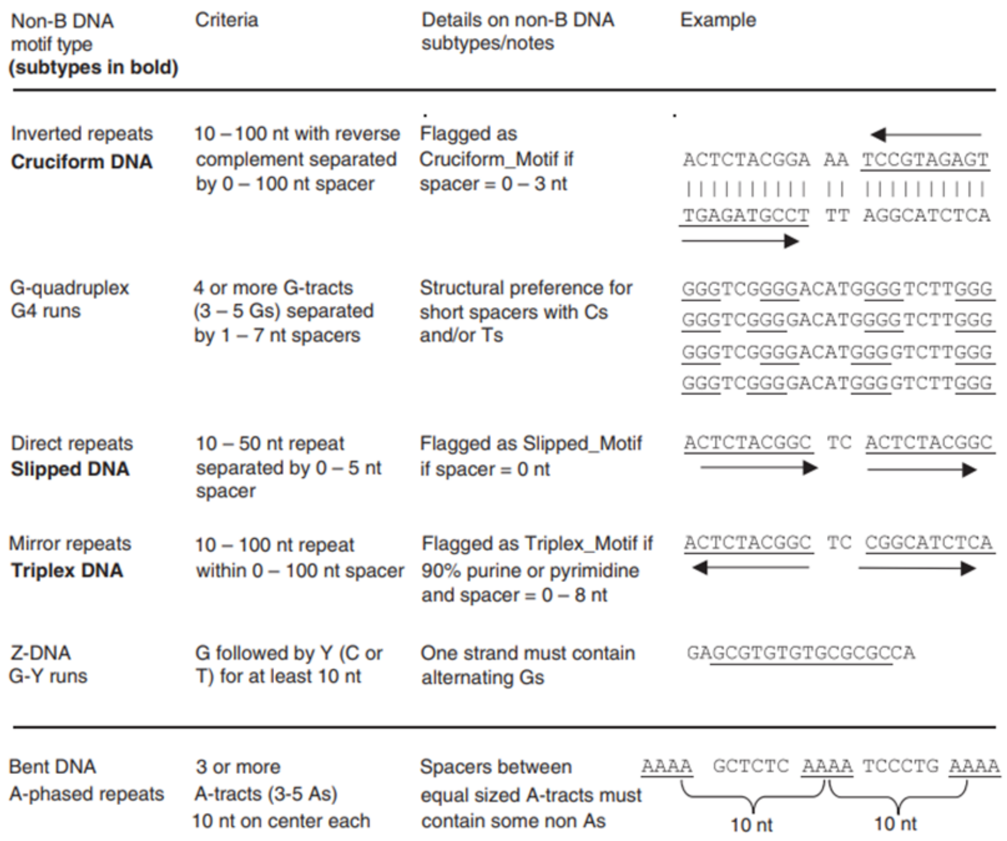


Fig. 4. Set of non-B DNA structures predicted in the study: curved, slipped, G-quadruplex, cruciform, triplex, Z-DNA structures with typical parameters.

The authors used a combination of high-throughput sequencing data and computational tools to identify and analyse these cis-regulatory motifs. For every organism, they extracted genomic regions adjacent to the translation start sites. These datasets were retrieved from the NCBI database, the Fungi Genome Databases, the UCSC genome browser, and PlantGDB. Information about GC contents and genomic sizes was found on NCBI's FTP database. They used a common and minimal definition of promoter regions for genes as the upstream region [-200 to -1], other regions that were accepted were [-500 to -301] and [+101 to +300]. As a negative control, they randomised every promoter sequence for each genome 10 times. Then they searched these promoter regions to find non-B DNA structures stated in Fig. 5. using the non-B DNA Motif Search Tool (nBMST). They had other possibilities of choosing the analysis tool such as QuadFinder, QGRS Mapper, TTS Mapping, TFO, Z-Hunt II and IRF but none of these provides the searching for every non-B DNA form.



Non-B DNA motif type **(subtypes in bold)**	Criteria	Details on non-B DNA subtypes/notes	Example
Inverted repeats **Cruciform DNA**	10 – 100 nt with reverse complement separated by 0 – 100 nt spacer	Flagged as Cruciform_Motif if spacer = 0 – 3 nt	ACTCTACGGA AA TCCGTAGAGT ||||||||| || ||||||||| TGAGATGCCT TT AGGCATCTCA
G-quadruplex G4 runs	4 or more G-tracts (3 – 5 Gs) separated by 1 – 7 nt spacers	Structural preference for short spacers with Cs and/or Ts	GGGTCGGGGACATGGGGTCTTGGG GGGTCGGGGACATGGGGTCTTGGG GGGTCGGGGACATGGGGTCTTGGG GGGTCGGGGACATGGGGTCTTGGG
Direct repeats **Slipped DNA**	10 – 50 nt repeat separated by 0 – 5 nt spacer	Flagged as Slipped_Motif if spacer = 0 nt	ACTCTACGGC TC ACTCTACGGC
Mirror repeats **Triplex DNA**	10 – 100 nt repeat within 0 – 100 nt spacer	Flagged as Triplex_Motif if 90% purine or pyrimidine and spacer = 0 – 8 nt	ACTCTACGGC TC CGGCATCTCA
Z-DNA G-Y runs	G followed by Y (C or T) for at least 10 nt	One strand must contain alternating Gs	GAGCGTGTGTGCGCGCCA
Bent DNA A-phased repeats	3 or more A-tracts (3-5 As) 10 nt on center each	Spacers between equal sized A-tracts must contain some non As	AAAA GCTCTC AAAA TCCCTG AAAA 10 nt 10 nt

Fig. 5. Non-B DNA-forming motif search criteria.

To quantify the prevalence of non-B DNA motifs in promoter regions in individual genomes, they implemented a percentage of association of non-B DNA motifs, which is defined as the percentage of promoter regions in a species with at least one occurrence of the non-B DNA motif. Statistical analysis using one-way ANOVA and Tukey's test confirmed the significance of higher non-B DNA values in promoter regions compared to upstream, downstream and shuffled sequences ($p < 0.01$). For the correlation analysis to assess the relationships between non-B DNA in promoter regions, GC percentage, and genome size across all organisms surveyed, they considered promoter regions using the Spearman rank-order correlation coefficient (ρ; $p < 0.001$). Their findings suggest that non-B DNA motifs are strikingly prevalent in promoter regions and exhibit variability linked to genomic, GC content, and taxonomic groups, spanning all domains of life.

Their results are summarised in Fig. 6.

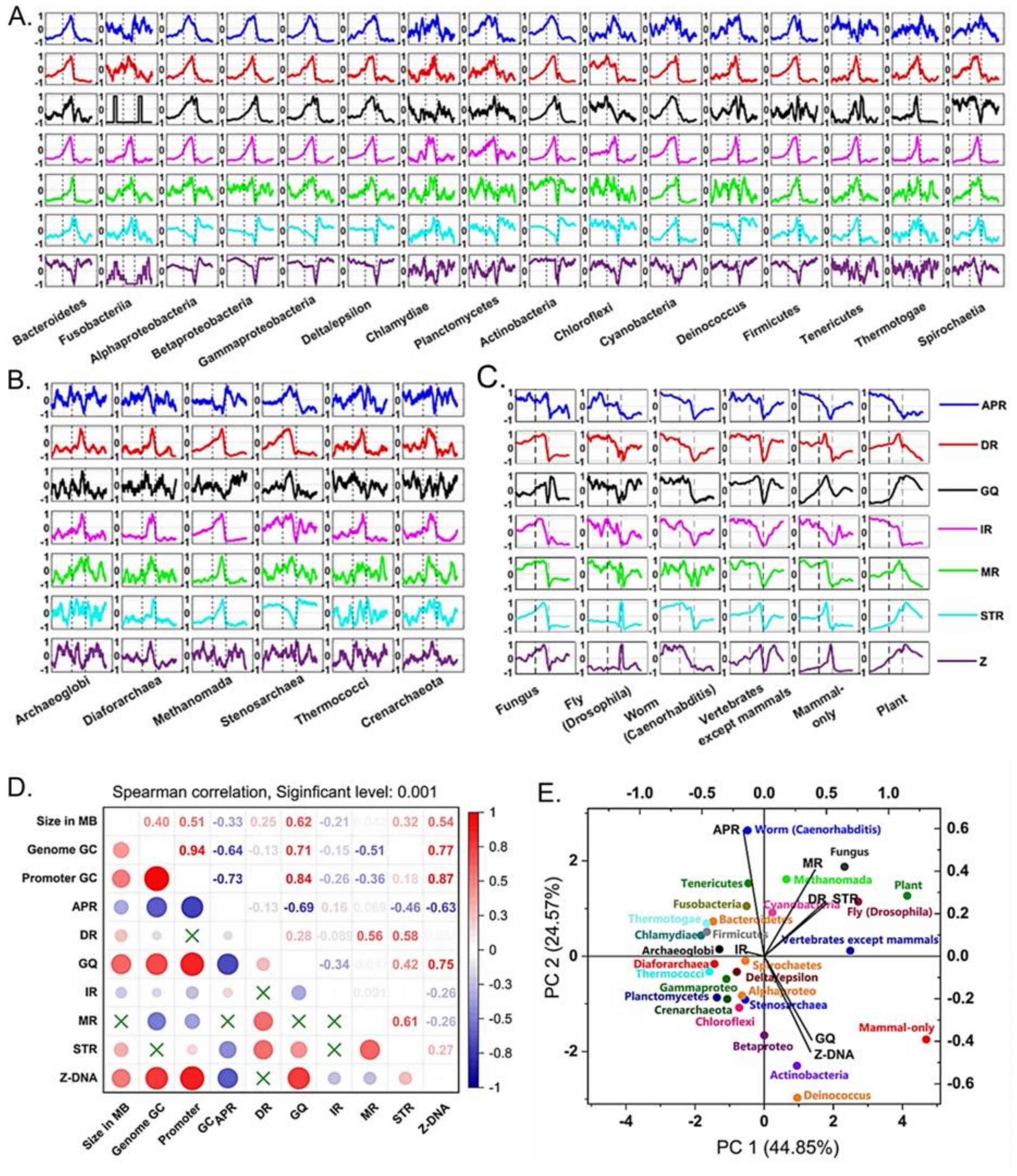


Fig. 6. Positional frequencies of seven non-B DNA motifs in: (A) bacteria, **(B)** archaea, and **(C)** eukaryotes. The position relative to the translation start site in all plots shows is marked on the x-axis. The y-axis shows the frequency of positions of non-B-DNA, scaled from -1 to +1. **(D)** Correlation between genome size, promoter GC, genome GC and seven motifs. Colour red, blue and green correspond to positive, negative and no correlation, respectively. **(E)** Principal component analysis (PCA) of mean non-B DNA association of promoter regions in 28 taxonomic phyla. The biplot of PCA analysis shows that PC1 (44.85%) is strongly influenced by Z-motifs and GQ, while APR contributed more to PC2 (24.57%).

By a use of the evolutionary timetree of life (TTOL) analysis they showed that the frequency of the G-quadruplex motifs in eukaryotic genomes increases as a function of the time of divergence, and they are under positive selection guided by evolutionary forces. They also determined the unique enrichment of non-B DNA forming motifs in individual taxa. Notably, cruciform DNA predominates in prokaryotes, while slipped DNA and triplex DNA motifs are dispersed in eukaryotes, and G-quadruplexes accumulate highly in mammalian promoters.

The authors suggest that these findings have implications for our understanding of the regulatory roles of non-B DNA configurations in evolutionary contexts.