OPEN ACCESS Freely available online

PLOS ONE

Comprehensive Human Transcription Factor Binding Site Map for Combinatory Binding Motifs Discovery

Arnoldo J. Müller-Molina[1], Hans R. Schöler[2,3], Marcos J. Araúzo-Bravo[1]*

1 Computational Biology and Bioinformatics Group, Max Planck Institute for Molecular Biomedicine, Münster, Germany, **2** Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, Münster, Germany, **3** Medical Faculty, University of Münster, Münster, Germany

Abstract

To know the map between transcription factors (TFs) and their binding sites is essential to reverse engineer the regulation process. Only about 10%–20% of the transcription factor binding motifs (TFBMs) have been reported. This lack of data hinders understanding gene regulation. To address this drawback, we propose a computational method that exploits never used TF properties to discover the missing TFBMs and their sites in all human gene promoters. The method starts by predicting a dictionary of regulatory ''DNA words.'' From this dictionary, it distills 4098 novel predictions. To disclose the crosstalk between motifs, an additional algorithm extracts TF combinatorial binding patterns creating a collection of TF regulatory syntactic rules. Using these rules, we narrowed down a list of 504 novel motifs that appear frequently in syntax patterns. We tested the predictions against 509 known motifs confirming that our system can reliably predict *ab initio* motifs with an accuracy of 81%—far higher than previous approaches. We found that on average, 90% of the discovered combinatorial binding patterns target at least 10 genes, suggesting that to control in an independent manner smaller gene sets, supplementary regulatory mechanisms are required. Additionally, we discovered that the new TFBMs and their combinatorial patterns convey biological meaning, targeting TFs and genes related to developmental functions. Thus, among all the possible available targets in the genome, the TFs tend to regulate other TFs and genes involved in developmental functions. We provide a comprehensive resource for regulation analysis that includes a dictionary of ''DNA words,'' newly predicted motifs and their corresponding combinatorial patterns. Combinatorial patterns are a useful filter to discover TFBMs that play a major role in orchestrating other factors and thus, are likely to lock/unlock cellular functional clusters.

Citation: Müller-Molina AJ, Schöler HR, Araúzo-Bravo MJ (2012) Comprehensive Human Transcription Factor Binding Site Map for Combinatory Binding Motifs Discovery. PLoS ONE 7(11): e49086. doi:10.1371/journal.pone.0049086

Editor: Nicholas James Provart, University of Toronto, Canada

Received July 3, 2012; **Accepted** October 8, 2012; **Published** November 28, 2012

Copyright: © 2012 Müller-Molina et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: marcos.arauzo@mpi-muenster.mpg.de

Introduction

Gene expression is regulated by the attachment of transcription factors (TFs) onto DNA binding sites located in promoter or enhancer gene regions. Each TF has a propensity to bind to a specific set of binding sites. This set can be represented by a binding motif [1]. Currently, only 10%–20% of the total human TF binding motifs (TFBMs) have been identified [2,3]. The most widely used databases of experimentally validated TFBMs are Jaspar [4] and Transfac [5]. Considering only globally traceable TFBSs (those for which the collection of all the *loci* targets is documented in the database across the whole genome), 228 and 281 TFBMs exist in Jaspar and Transfac databases, respectively. It is estimated that the total number of human TFs ranges between 1400 [2] and 2600 [3], hence 80% to 90% of the TFBMs are unknown. Furthermore, there are TFs that use more than a TFBM, originating the so called secondary motifs [6]. Such motifs add major variation and complexity to the TFBM repertoire implying a much higher number of unknown motifs. Thus, to gain a comprehensive understanding of the gene regulation process, it is necessary to discover the unknown TFBMs set. Once this set of TFBMs is predicted, the crosstalk with their corresponding TFBSs can be analyzed in a more comprehensive way.

The average length of the known TFBMs is 11.53 base pairs (see section S1.1 in Supporting Information S1). This is shorter than the required length to achieve enough binding specificity (30 information bits as shown by Wunderlich *et al.* [7]). The TFBMs of eukaryotes are shorter than those of prokaryotes [7], and therefore their binding specificity is lower. To compensate for this effect, gene expression in eukaryotes is regulated by the orchestration of multiple TFs [8]. By clustering several short motifs together, the effective spanned length of the combined pattern achieves a greater level of specificity. Given all TFBMs and their corresponding binding sites, it is worthy to extract common syntactic structures that arise in the promoter binding topologies since such structures can constitute regulatory rules that provide additional insights in the transcriptional regulation process. Such rules can also be used to select TFBMs that frequently appear with other motifs and therefore control a larger amount of functionality.

Here, we provide algorithms and databases that answer the following questions: What do unknown TFBMs look like? Given all possible TFBMs, what is the motif combinatorial binding topology of each promoter? What genes have common TF combinatorial binding patterns and are therefore likely to be switched simultaneously? Do genes regulated by the same TFs have common functions?

In order to answer such questions, we have developed new computational methods that acquire knowledge from the already known TFBSs. The novelty of these methods includes an adaptive choice of the number of aligned sequences from different species using a permutation prefix method, clustering of the selected sequences using a new distance function (conglomerative distance) to gain granularity, and an adaptive choice of the number of features to be learn from each known TFBM thanks to a new filtering technique that we term dynamic dimension selection (DDS). These methods generate a database with the following major features. Firstly, the database provides a dictionary of "DNA words" that contains all the possible binding sites. This dictionary is not associated to any particular cell type but is a key element for inferring the transcription regulations on genomic scale. Such types of dictionaries are a key tool of the techniques that break cryptographic codes [9]. Secondly, our algorithm generates a list of *ab initio* TFBM predictions that cover the unknown 80%–90% of human TFBMs. Thirdly, our method distills a list of common TFBM combinatorial "syntax" rules that arise in the gene promoters region. Fourthly, a sublist of TFBMs, based on motifs that appear in the discovered combinatorial syntax rules, can have a greater ability to regulate larger modules of cellular functionality. Finally, we predict the potential biological functionality of the newly found TFBMs and their combinatorial patterns annotating them with the gene ontology enrichment analysis of their associated gene targets.

The task of finding *ab initio* TFBMs has been tackled in the past with a large number of algorithms. A recent survey [10] states that even after a considerable effort, DNA motif finding still remains an open problem as motif finding algorithms are not able to detect motifs in mammals. Here we describe the two main categories of motif finding algorithms.

Constrained discovery algorithms

The first category is composed of algorithms [10] that work on small sequence fragments. Initially developed for predicting TFBMs from co-expressed gene clusters determined by transcriptomics experiments, and also used for searching TFBMS in the DNA fragments generated by wet-lab analyses such as DNAse footprinting assay, Electrophoretic Mobility Shift Assay (EMSA) and more recently ChIP-Chip and ChIP-Seq [11]. After sequence mapping, such experiments deliver *loci* of limited length and each of these *loci* is assume to hold a binding site with a certain probability. The algorithms operate on the sequences finding common motifs. Examples of these techniques include AlignAce [12], Gibbs Motif Sampler [13], MEME [14], PhyloGibbs [15] or Weeder [16]. We term them constrained algorithms since they are designed to find only one motif or a small set of motifs from an experiment. Their outcome is cell type specific and they do not extrapolate the knowledge from known motifs to perform *ab initio* predictions. Elemento and Tavazoie [17] argued that such techniques are not appropriate, primarily because binding site density is low. They also mentioned that those algorithms are relatively slow and can miss many regulatory elements as they are based on stochastic methods [18–21]. In practice, such approaches are reported to work only on small genomes like yeast [10]. The described problems render these algorithms ineffective as the genome becomes more complex [10]. Our objective is to discover from whole eukaryote genomes *ab initio* cell type independent motifs using computer learning techniques that take advantage of the intrinsic binding properties of the known motifs.

Unconstrained discovery algorithms

The second category comprises algorithms that systematically discover *ab initio* motifs [17,22–25]. These algorithms rely on the small number of known TFBMs compared to the number of TFs. They try to generalize the properties of the already known TFBMs to find new TFBMs. This is a challenging task since the number of binding sites targeted by TFBMs is very small compared to the number of all the possible subsequences of a genome. Table 1 shows a list of whole genome motif discovery methods and their performance characteristics. Success rates vary widely among different algorithms. In the yeast case, better results were achieved [26] due to the lower genome complexity in this organism. Since previous methods have been tested with different motif discovery conditions, the number of validations varies considerably. E.g., in the human, Elemento *et al.* [24] employed 309 validation motifs whereas Xie *et al.* [22] used only 123. Regarding the number of *ab initio* predictions, in human, FastCompare [24] produced 284 novel predictions whereas the method proposed by Xie *et al.* [22] provided 184. These numbers are still far from the final goal of 1400–2600 TFBMs. Furthermore, all methods besides the one presented in this publication, employ IUPAC (International Union of Pure and Applied Chemistry) [27] symbols to describe motifs. This is a drawback since the generated motifs are too coarse, and therefore their biological accuracy is reduced. None of the previous methods employs machine learning techniques to find properties intrinsic to known TFBMs. A common trend is to employ one-dimensional overrepresentation/conservation scores [22], but this is a very limiting approach to describe complex binding patterns. Finally, TF combinatorial co-occurrence that is a valuable filter to find TFs controlling large modules of functionality in the cell, is not performed by previous approaches.

Our algorithm significantly improves the state of the art by fundamentally changing the way to tackle the problem. First, we add features intrinsic to known TFBMs (see Table 2) that are processed with a machine learning approach. This allows our algorithm to focus not only on overrepresented and overconserved elements but also on elements that match closely existing TFBMs in a hyperdimensional space. TFs are projected onto multidimensional vectors, and we attempt to find novel motifs that are in the vicinity of existing points (known TFs). Second, our system does not employ IUPAC symbolic simplification to generate motifs. Instead, it takes advantage of the richer information carried by the collection of binding site sequences targeted by each TF. Third, it uses similarity search to create more natural and biologically meaningful motifs with very high Pearson correlation with already known TFBMs. Fourth, our algorithm employs a combinatorial filter that finds motifs that are common in multiple gene promoters in a predictable topological form. This allowed us to obtain the best results with the largest validation set, the highest number of novel predictions, and the highest success rate reported so far in the literature for unconstrained whole genome motif discovery algorithms.

Results

Prediction of 4089 new transcription factor binding motifs and validation against known motifs

To disclose *ab initio* TFBMs on genomic scale, we designed a pipeline of algorithms that decrease progressively the amount of genomic data to be processed. First, we compiled "DNA word" dictionaries from all the gene promoters of 5000 base pairs upstream the transcription start site of the human genome. We designed three new filters to reduce the number of "DNA words" in the dictionaries. The reduction achieved by each filter is shown

Table 1. Comparison of unconstrained motif discovery algorithms.

Method	Year	Species	Succ.	Test	Novel	ML	Comb.	HR
Kellis *et al.*[26]	2003	*Sc*	65%	55	72	N	N	N
FastCompare[24]	2005	*Sc*; *Sb*	8%	309	398	N	N	N
FastCompare[24]	2005	*Ce*; *Cb*	4%	309	437	N	N	N
FastCompare[24]	2005	*Hs*; *Mm*	5%	309	284	N	N	N
Xie *et al.*[22]	2005	*Hs*	56%	123	174	N	N	N
Stark *et al.*[68]	2007	*Dm*	46%	87	145	N	N	N
MDOS[69]	2008	*Pf*	20%	30	26	N	N	N
Kumar *et al.* [23]	2010	*Fg*	21%	76	108	N	N	N
This publication	2012	*Hs*	81%	509	4098	Y	Y	Y

The method and the year of publication is displayed along the species. "Succ." is the success rate of the algorithm or the percentage of known motifs rediscovered. "Test" is the number of known motifs tested in the validation step. "Novel" is the number of novel motifs predicted. "ML" indicates whether a machine learning approach is used. "Comb." indicates whether a combinatorial filter is employed. "HR" indicates whether high resolution predictions are available. If IUPAC symbols are used during the enumeration phase, the prediction resolution is low. Species: *Cb* = *Caenorhabditis briggsae*; *Ce* = *Caenorhabditis elegans*; *Dm* = *Drosophila melanogaster*; *Fg* = *Fusarium graminearum*; *Hs* = *Homo sapiens*; *Mm* = *Mus musculus*; *Pf* = *Plasmodium falciparum*; *Sb* = *Saccharomyces bayanus*; *Sc* = *Saccharomyces cerevisiae*.
doi:10.1371/journal.pone.0049086.t001

in Table S1 in Supporting Information S1. The parameters employed for the dictionaries generation are shown in Figure S1 in Supporting Information S1, and examples of how such filters work are shown in Tables S2 and S3 in Supporting Information S1. The "DNA word" reduction achieved by these filters facilitates at a second stage, to complete the TFBM map through clustering. To decrease the potential TFBM candidates to an amenable number, we created a novel Dynamic Dimension Selection (DDS) filter, that using the parameters depicted in Table S4 in Supporting Information S1 dramatically condenses the number of potential TFBM as shown in Table S5 in Supporting Information S1. A final merging step identifies 4598 motifs. This number is higher than the estimated upper bound of 2600 TFs [3], but one has to consider that a TF can use several binding modes which can generate the so called secondary motifs [6]. Using as a similarity criterion a Pearson correlation of at least 0.85, our predictions matched 83% known TFs from Jaspar [4] (228 TFBMs) and 81% of Transfac [5] (281 TFBMs) datasets. To achieve these results, for each motif of length w, $w_{min} \leq w \leq w_{max}$ (where $w_{min}=4, w_{max}=16$) we trained our algorithms with a subset of known TFBMs with a length different from w. Then we predicted the existence of the trained TFBMs in the disjoint validation subset composed by the TFBMs of length w. Once known motifs were removed, there remained 4089 novel predictions. To verify that our predictions do not occur by chance, we measured the percentage of predictions that match a random control TFBM dataset. Since each TFBM is characterized by a position weight matrix (PWM) [28], to generate a dataset without nucleotide compositional biases, the elements of each of the Jaspar and Transfac PWMs were shuffled for each column while keeping the column order. We calculated how many of our predictions had a similarity higher than a threshold against at least one of the random PWMs. We selected the similarity threshold following [22], considering that a prediction matches a known TFBM when the Pearson correlation of its respective PWMs is ≥ 0.85. We found that only 1% of our predictions matched the random dataset, thus, the probability that our *ab initio* TFBMs have been generated by chance is low. Figure 1 presents the known motifs predicted by our method with the highest two matches obtained for each prediction of a specific length. The entire list of predictions is shown in the Table S6 in Supporting Information S1. Our algorithm is able to predict not only individual TFBMs but also possible combinations, such as the OCT4+SOX2 (MA0142.1) pair, which is well known in embryonic stem cells (ESCs) [8,29]. In ESCs, OCT4 activates downstream genes by binding to enhancers carrying the octamer-sox motif (OCT-SOX enhancer) for synergistic activation with SOX2 [30], playing a key role in the maintenance of pluripotency [31–33]. The pair OCT4+SOX2 is also crucial for cellular reprogramming, since both OCT4 and SOX2 are members (together with KLF4 and MYC) of the Yamanaka reprogramming cocktail [34], and (together with LIN28 and NANOG) of the Thomson reprogramming cocktail [35]. The ability to discover OCT4+SOX2 shows possible follow up applications of our algorithm for understanding the reprogramming mechanisms. When comparing with other unconstrained discovery algorithms, we found that though our method was validated with the largest amount of known TFBMs (509), it still achieved the highest success rate (81%) reported in the literature (see Table 1).

Table 2. Intrinsic properties of the TFBMs used by our algorithms.

D	Feature	Description
w	*Entropy curve*	Entropy-based fingerprint
w	*Distance distribution*	Inter binding site Hamming distance distribution of the prediction
1	*Conservation score*	Conservation score from Xie *et al.*[22]
1	*Number of conserved sites*	Total number of conserved sites for each prediction
1	*Conservation per base*	Average number of conserved species per base for each prediction site
1	*Entropy per base*	Average motif entropy per base

For each dictionary of "DNA words" of length w, the algorithm works in a $(2w+4)$ dimensional space defined by six multidimensional feature fingerprints. "D" is the number of feature dimensions.
doi:10.1371/journal.pone.0049086.t002

The TFBMs emerge forming combinatorial binding patterns

Since the TFBMs are short, they do not have enough specificity, and there are multiple evidences that TFs work together in combinatorial assemblies, such as the OCT4+SOX2 partnership [30]. We hypothesized that when a motif appears together with the same set of motifs in multiple promoters, the motif is more likely to play a relevant role in gene transcription regulation. We denote this arrangement of TFBM co-occurrence as combinatorial binding patterns (CBPs), i.e., CBPs are "motifs of motifs". To discover such patterns, we developed a computational method that identifies combinations of motifs that bind to more than one promoter. The 17831 CBPs we found cover 73% of already known TF-TF binary interactions listed in the Transcompel [5] dataset. The best CBP match against each Transcompel entry is shown in the Table S7 in Supporting Information S1.

From the initial set of TFBM predictions, we found a subset of 504 motifs that also co-occur with other motifs in multiple gene promoters. This set is denoted as STFBM (significant TFBM). Figure 2 shows the top two matches for each motif prediction set of length w that include statistically significant gene ontology (GO) enrichment over the genes targeted by the CBPs. The complete STFBM is presented in the Table S8 in Supporting Information S1. By applying the GO enrichment analysis over the genes targeted by such motifs (section S1.5 in Supporting Information S1), we annotated successfully 51% of STFBMs with GO terms with statistical significance ($P \leq 0.05$), and found that this subset is constituted exceptionally by TFs and enriched with pattern formation and morphogenesis related GO terms (section S2 .1 in Supporting Information S1).

To visualize the topology of the CBPs, we designed a technique (see section S1.5.1 in Supporting Information S1) based on performing multiple alignments of the predicted TFBS *loci* coordinates. Figure 3 gives three examples of this visualization. Figure 3A shows a very specific CBP that applies to 16 promoters. Our *ab initio* TFBM predictions "b-11-3-0" and "b-7-53-8" are mixed with the TFs Sp1, FAC1, and Zic3. Gene expression microarrays [36] have already shown that Sp1 and Zic3 are members of one gene regulation cluster in cardiac cells (cluster 7), whereas FAC1 and Zic3 belong to clusters 22, 63, and 101. Finally, Sp1 and FAC1 belong to cluster 72. Figure 3 B shows a CBP that involves AP-1/Sp1 interactions. This interaction has already been documented [37] and is related to 12-*O*-tetradecanoylphorbol-13-acetate (TPA) response in keratinocytes. This example shows the ability of our system to find co-occurrence of the same motif in different directions (see the multiple appearances of the Sp1 complement). Finally, Figure 3 C displays a comprehensive CBP that applies to 1218 genes. The TF Pax-8 is surrounded by three of our predictions. The CBP subset with the strongest connection is composed of "b-7-110-0", "b-6-116-1", and "Pax-8". As PAX-8 is an excellent marker for primary tumor sites [38], our *ab initio* TFBMs associated with Pax-8 could provide insights into tumor formation regulation. We found 9 additional CBPs that apply to more than 1000 genes (see CBP gallery in Table S9 in Supporting Information S1), indicating that a small subset of CBPs has the potential to exert regulation over large sections of the genome.

The intrinsic specificity of the CBPs is limited to subsets of ten or more genes

To study to which degree the CBP are specific to genome *loci*, we analyzed the distribution of the gene targets covered by each CBP. Figure 4 A depicts the cumulative percentage of CBPs that apply to a certain number of genes and shows an elbow at the level of 10 genes. Figure 4 B displays the number of genes versus the CBPs of a specific motif count showing that the average number of genes switched by CBPs is always equal to or greater than 10. These results indicate that, on average, 90% of those CBPs are shared among 10 different genes. The analysis of the average number of genes covered by CBPs with a specific amount of motifs shows that as the number of motif count increases (Figure 4 C), the gene count reduces and stabilizes at about 30 motifs. Few CBPs hold more than 30 motifs. Such observation is highlighted by the heat map in Figure 4 D, showing that the percentage of CBPs that holds x motifs and y genes has a high density region in the area of 10 genes targeted by less than 30 motifs. These results reveal that the newly discovered arrangements of TFBMs into CBPs increase the binding specificity only to an extent that a CBP can switch 10

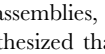
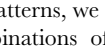
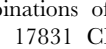
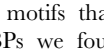
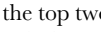
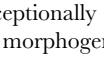
ID	Predicted	Known	ρ	TF name
b-6-0-5			0.998	P$EMBP1_02
b-14-99-0			0.996	MA0082.1
b-8-166-0			0.995	MA0018.2
b-8-221-0			0.990	MA0121.1
b-6-40-0			0.989	V$AML1_01
b-7-204-0			0.983	I$BRK_Q6
b-4-0-0			0.982	P$STF1_01
b-9-0-0			0.982	V$CKROX_Q2
b-7-67-1			0.982	V$ETF_Q6
b-5-33-0			0.978	MA0064.1
b-5-1-1			0.974	MA0075.1
b-9-27-0			0.968	MA0242.1
b-11-71-0			0.962	P$MYB80_01
b-10-6-0			0.960	V$ELK1_04
b-10-60-1			0.960	V$LHX3_01
b-13-144-0			0.957	V$HNF4_01_B
b-13-137-0			0.952	V$FOX_Q2
b-4-7-0			0.950	P$MYBAS1_01
b-12-1-1			0.946	V$DR1_Q3
b-11-248-0			0.932	F$ATF1PCR1_01
b-12-85-0			0.913	V$HNF6_Q6
b-14-57-0			0.910	V$SOX9_B1
b-16-1-0			0.867	MA0060.1
b-15-13-0			0.867	MA0137.2
b-15-3-0			0.866	MA0142.1

Figure 1. Comparison of predicted motifs against known regulatory motifs. "ID" is the TFBM identification string and ρ is the Pearson correlation of the similarity between the predicted and the known TFBMs. The TF names starting with MA correspond to Jaspar TFBM identifiers, the names with a $ symbol correspond to Transfac identifiers.
doi:10.1371/journal.pone.0049086.g001

ID	Predicted	z	CBP	Function	Process	Component
b-5-14-0		Inf.	c-1367	-	-	GO:0044459
b-5-54-0		182.7	c-1736	-	GO:0032501	-
b-6-62-1		94.95	c-1912	-	GO:0043170	-
b-6-20-5		88.12	c-1925	GO:0001071	-	GO:0005634
b-7-4-27		Inf.	c-1546	-	GO:0007155	GO:0016021
b-7-76-2		Inf.	c-1546	-	GO:0007155	GO:0016021
b-8-307-0		Inf.	c-1526	GO:0003676	GO:0034641	-
b-8-328-0		Inf.	c-1542	-	GO:0010646	-
b-9-36-0		Inf.	c-1526	GO:0003676	GO:0034641	-
b-9-206-0		32	c-2627	-	GO:0032502	-
b-10-185-0		Inf.	c-1541	GO:0005509	-	-
b-10-129-0		59.78	c-2063	GO:0030528	GO:0000122	GO:0005634
b-11-80-0		Inf.	c-1551	-	GO:0007275	-
b-11-217-0		Inf.	c-1550	-	GO:0031326	-
b-12-72-0		Inf.	c-1539	GO:0004871	GO:0032501	GO:0005886
b-12-503-0		Inf.	c-1551	-	GO:0007275	-
b-15-22-0		Inf.	c-1541	GO:0005509	-	-
b-15-210-0		9.42	c-10938	GO:0005198	GO:0048584	GO:0016023

Figure 2. Novel predicted motifs that appear frequently in CBPs (STFBM). The position within this subset is included along the motif ID, the motif logo, the significance z score, the corresponding CBP ID. Additionally, the most significant molecular function, biological process and cellular component ontologies are included for each motif.
doi:10.1371/journal.pone.0049086.g002

genes on average. Thus, the CBPs exert clusters with a granularity of more than 9 genes. This suggests that the simultaneous switching of fewer than 10 genes requires other mechanisms besides TF combinatorial binding. These computational results are in agreement with the findings of Lieberman-Aiden *et al.* [39], who coupled proximity-based ligation with massively parallel sequencing (Hi-C) and with the gene expression microarray results of Vogel *et al.* [36]. The former showed that chromatin segregates into two genome-wide compartments, where the open one is consistent with a knot-free fractal globule that preserves the ability to unfold any genomic locus. The latter showed that a large proportion of the human transcriptome is organized into gene clusters that are partially regulated by the same TFs.

The *ab initio* predicted TFBS and CBPs are related with developmental functions and transcription regulation processes

To reveal the possible biological meaning of the newly discovered TFBMs and CBPs, we generated a list of significantly enriched GO terms associated with the genes targeted by the motifs and CBPs. We were able to annotate with a statistical significant enrichment the GO terms of 93% of the predicted CBPs and 51% of the TFBMs, which shows that the newly discovered TFBMs and CBPs are likely to be biologically meaningful. We developed the concept of "ontology maps" to visualize the relationship between GO terms and TFBMs or CBPs (section S1.6.2 in Supporting Information S1) with "graphs as maps" (GMAP) [40]. Figure 5 presents the map of CBP molecular function ontologies. All elements enclosed in the same "country" of the map have links to similar sets of ontologies and therefore cluster together. The visualization shows that the surrounding CBPs also bind to genes exhibiting similar ontologies. The text size reflects the most common, significant GO terms. Interestingly, we found that such terms are related to molecular function ontologies of TFs revealing a cascade of TF binding events, i.e. the CBPs we discovered have a trend to bind to the regulatory region of other TFs. As an example of TF activity, the white highlight in the upper left region of Figure 5 (near the nucleic acid binding TF activity "country") marks the position of the c-10688 CBP, whose CBP arrangement is depicted in Figure 3 A. Additionally, the upper right corner of Figure 5 depicts a cluster of CBPs connected strongly with the "RNA binding" ontology. The corresponding "RNA splicing" term is found to be a biological process significantly regulated by CBPs, as shown in Figure 6. This correspondence suggests that the modular nature of the protein blocks codified in sets of exons to generate different isoforms could require the concurrence of a large number of TFs. Figure S6 in Supporting Information S1 depicts the ontology map of CBPs cellular components. We also found a significant enrichment in the TFBMs GO analysis of transcription regulation related terms such as "transcription factor regulator activity" and "transcription factor binding", and of developmental related terms such as "developmental process", "anatomical structure development", "tissue development" and "organ morphogenesis" (see Figures S7, S8 and S9 in Supporting Information S1).

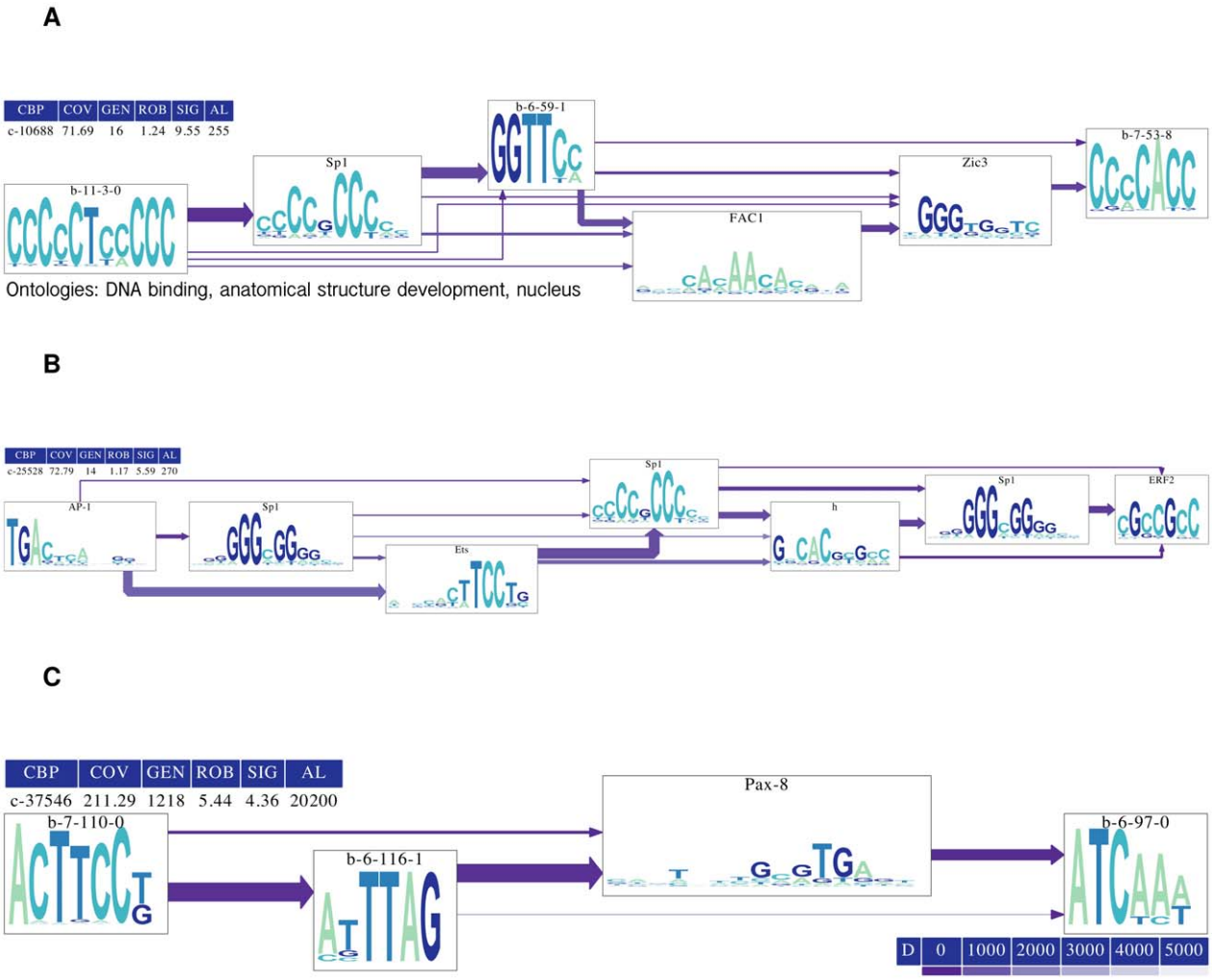


Figure 3. Visualization of discovered CBPs. Visualization of CBPs with (A) , (B) low and (C) high number of target genes. Each arrow connecting two motifs *a* and *b* implies that the motif *b* is to the "right" of the motif *a*. The arrow thickness reflects the percentage of promoters where the motif relationship occurs. The gradient "D" is the number of base pairs separating each binding site. "CBP" is the unique identifier for each discovered pattern. "COV" (coverage) is the average number of base pairs available for binding. "GEN" (generality) is the number of genes targeted by the pattern. "ROB" (robustness) is the average number of times a motif is repeated within the pattern. "SIG" (significance) is the *z*-score of the actual GEN in the distribution of GEN that would be generated if the promoters were randomly shuffled. "AL" is the motif multiple alignment score. Higher values of "AL" represent better alignment.
doi:10.1371/journal.pone.0049086.g003

Discussion

We have created novel proximity algorithms that infer a catalog of human regulatory motifs and their combinatorial binding patterns from previously known TFs. Our methods predicted 81%–83% of the known TF binding motifs and discovered 4089 novel TFBMs. This is the highest success rate achieved so far in mammalian whole genome TFBM discovery and it is the most widely validated method with rediscovering 509 known motifs of the Jaspar and Transfac databases. This success rate shows that our machine learning approach is useful to improve the predictions quality. Since our algorithm is not based on IUPAC strings, it creates smoother motifs that closely match experimentally discovered TFBMs than the methods based on IUPAC representations.

A combinatorial pattern filter selected a subset of 504 motifs that co-occur frequently with other motifs in 17831 CBPs. These motifs hold co-occurrence patterns that span hundreds of genes. A high percentage of our CBP and TFBM predictions can be annotated with statistical significance with GO terms. Such annotation reveals that our CBP and TFBM predictions are strongly related to transcription activity and development. Thus, the TFs show a trend to target other TFs. This additional level of regulation has been discovered in several specific cases such as the Microphthalmia-associated transcription factor (MITF) regulation by SOX-10 and PAX-3 in the Waardenburg syndrome [41], the modulation of chondrogenesis onset by Runx [42], the specification of lymphoid cell fates [43], or the ESC regulatory circuitry [44,45]. We found that such property seemed to be an intrinsic genomic sequence hallmark with expanding influence on genomic scale. In our study, genes involved in pattern formation, morphogenesis and development were also regulated by our *ab initio* TFBMs and CBPs with high significance. This identifies development as one of the most demanding regulatory processes, in agreement with the pivotal role of transcription regulation during development [46].

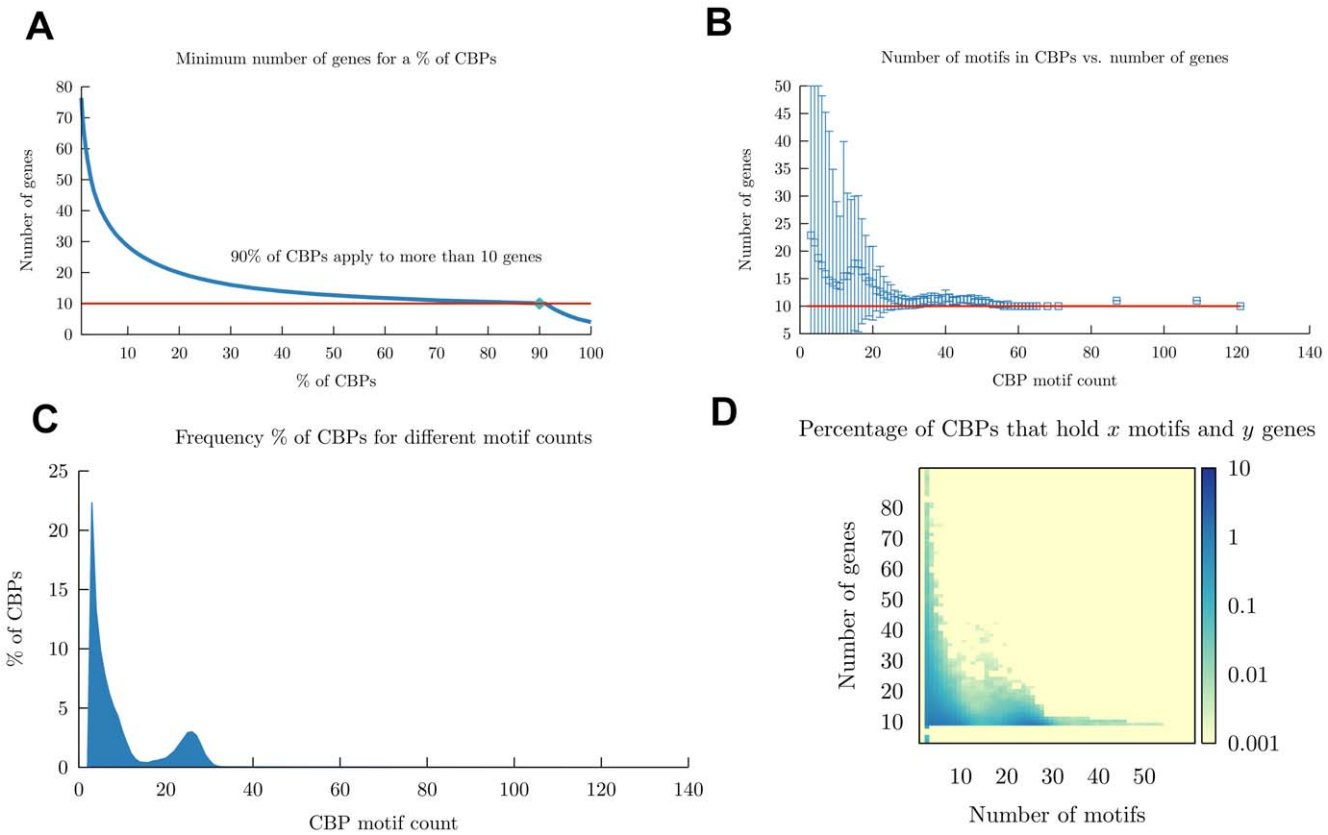


Figure 4. CBPs apply to 10 or more genes. (A) Cumulative percentage of CBPs (x axis) that apply to a given number of genes (y axis). (B) Relationship between CBP motif count and gene count. The y axis represents the average number of genes contained in the CBPs. One standard deviation surrounds each point. The x axis represents all the CBPs with a certain number of motifs. The horizontal red line marks the 10 gene boundary. (C) Frequency of CBPs for different motif counts. The x axis represents the motif count for all the CBPs. The y axis represents the percentage of CBPs that hold the given size. (D) Percentage of CBPs that apply to a given number of genes and motifs.
doi:10.1371/journal.pone.0049086.g004

The discovered TFBMs and CBPs hold properties that appear to be intrinsic to the genome, as they have been revealed using a method that applies the same algorithm over all promoter regions. Potential features encoded in the genome confer additional regulatory capabilities onto the cell, the degree of which may depend on the particular cell type. Thus, gene regulation at the TF level appears to be focused on the regulation of other TFs influencing the development of the organism. Accordingly, the development of an organism could entail cellular functions whose regulation requires higher levels of complexity.

The predicted CBPs are generated greedily in an attempt to extract potential syntactic patterns that are as general as possible. We have created a computational approach to stretch general genomic syntactic rules as much as possible, with as many motifs as possible. Even though we accepted a correct CBP as a syntactic rule that applies to at least two genes, we discovered a consistent and surprising trend to generate sets of rules that apply to 10 or more genes.

Previous unconstrained motif discovery approaches [22], generate less specific motifs as more species are added, and their sensitivity analysis showed that their method predicted with less accuracy (69 hits out of 123 tested motifs). Instead, our method takes advantage of additional information provided by the growing number of aligned species in databases such as the UCSC genome browser [47]. It has the capability to improve its performance as more genomic data becomes available, because we employed a permutation technique whose key feature is to select adaptively the number of closer species to focus on relevant matches only (see Materials and methods section).

Our method is designed to learn from some additional TF intrinsic properties (Table 2) that profit from the extended sequence alignments. These properties do not require sequence overrepresentation and focus on the fingerprints of specific sequences. Thus, our algorithm searches for TFBMs that do not necessarily occur frequently in the genome. Our usage of TF intrinsic properties allows considerable reduction of the input noise (see section S1.4 in Supporting Information S1). Our "DNA word" dictionary compilation algorithm adapts locally to evolution changes. Besides evolution, the multiple alignment heuristic [48] employed may also introduce artifacts into the data. Since our method learns from the feature space, better results have been achieved.

We hypothesized that evolution occurs at different rates along the genome. Therefore, local genomic regions may have different conservation levels among species [49]. Such difference may not be considered in a general phylogenetic tree based on whole genome sequences. Other methods take such a generalistic approach [23], but the adaptive selection of aligned sequences provided by our permutation prefix approach for filtering "DNA words" makes our algorithm more flexible by adapting locally, depending on learned features. Permutation prefixes have been used recently as fingerprints in the field of similarity search [50–53].

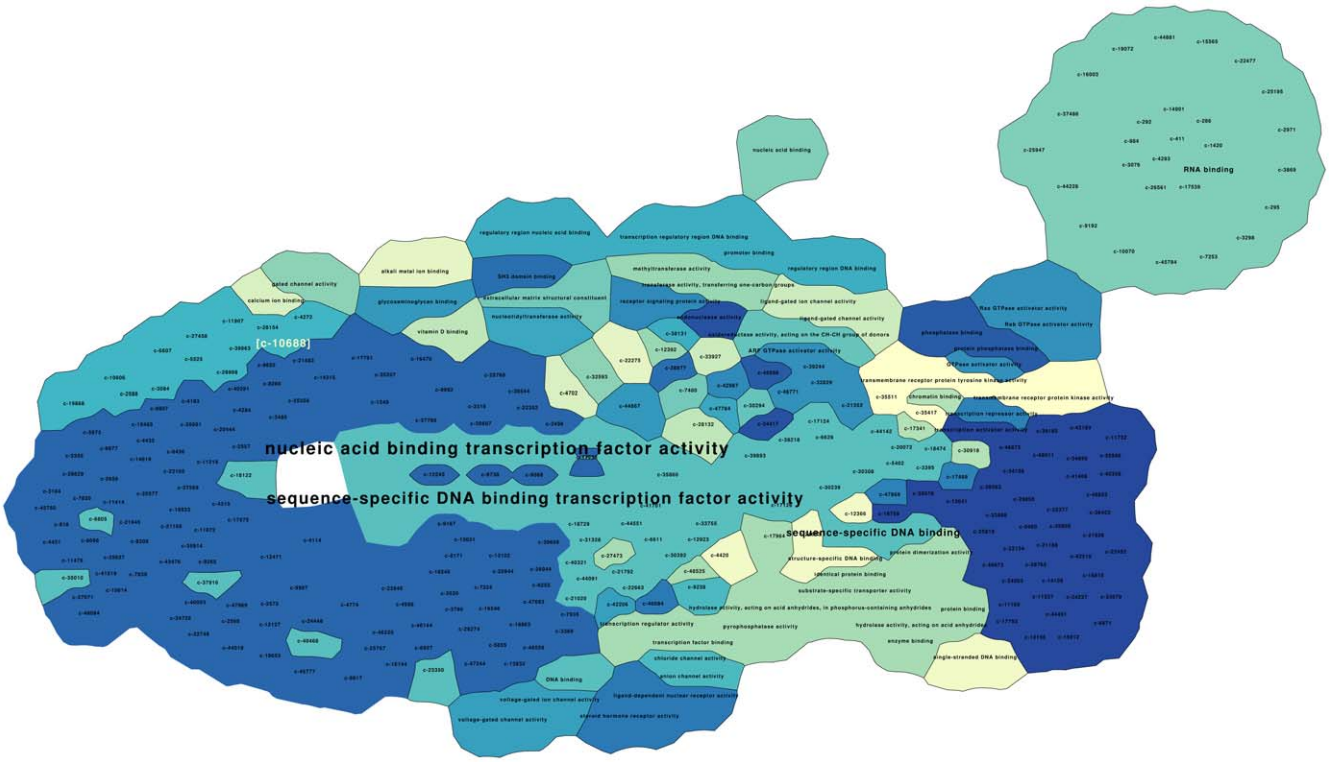


Figure 5. Map of significant CBP molecular function ontologies. Regions with the same color represent the same cluster of objects. Borders of the "countries" underline boundaries of CBP clusters. The ontology term font size is adjusted to reflect the frequency with which it is associated to the CBPs. The CBP c-10688 has been enhanced for readability in the upper left corner of the visualization map.
doi:10.1371/journal.pone.0049086.g005

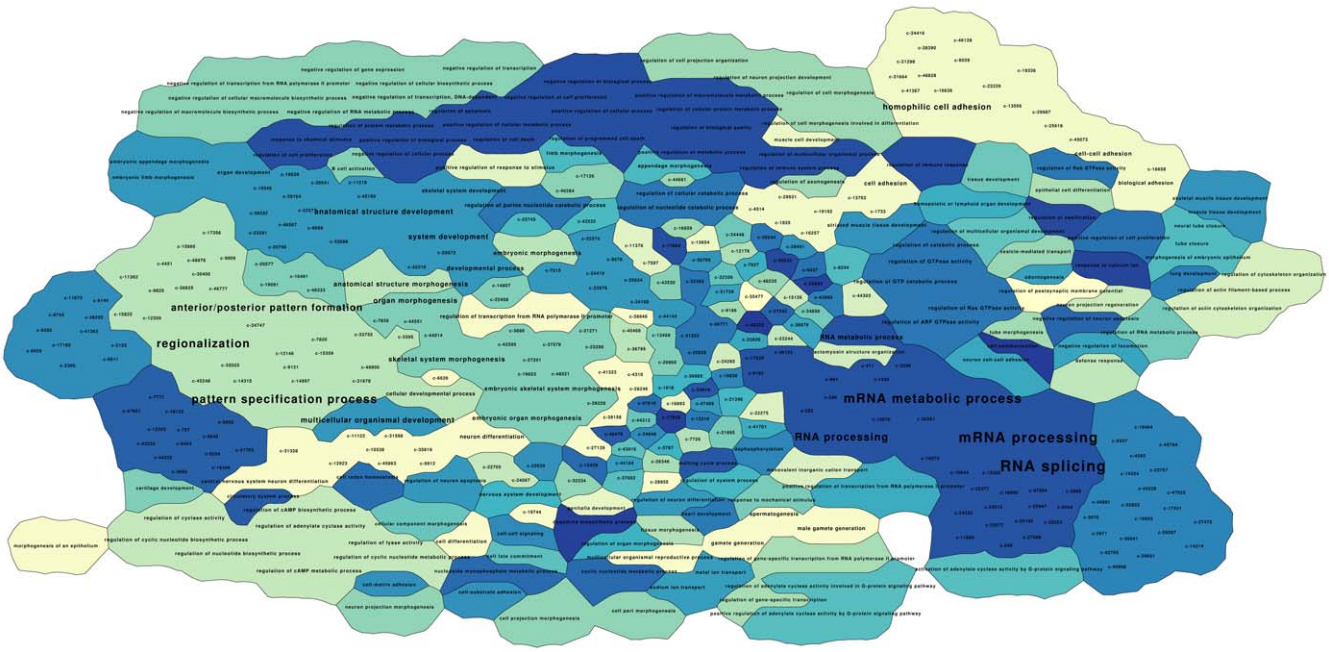


Figure 6. Map of the significant CBP biological process ontologies. Regions with the same color represent the same cluster of objects. Borders of the "countries" underline boundaries of CBP clusters. The ontology term font size is adjusted to reflect the frequency with which it is associated to the CBPs.
doi:10.1371/journal.pone.0049086.g006

The DDS algorithm introduced a novel way to handle high dimensional data by focusing on certain features of the space. This allowed us to project each prediction into a high dimensional space without the need to evaluate the relevance of different features. The important features are found at classification time and depend on the target object. By concentrating on a limited subset of components, the DDS algorithm removes noise and focuses on the important features. The dimension selection is dynamic and depends on each object, and unlike the Euclidean metric, other l_p metrics, and some recent approaches [54,55], the differences of the components are not mixed when the similarity is calculated. Since the DDS treats each single component independently, there is no need to normalize data with different magnitudes. The combination of these key properties allowed the DDS to remove vast amounts of spurious noise (see Table S5 in Supporting Information S1), thus, reducing the number of false positives.

With our algorithms we generated a catalog of predictions that contains the "DNA words" dictionary, the motif predictions, the binding site location for each known and novel motifs, the CBPs predictions, and the visualizations for motifs and CBPs which can be downloaded from http://computational-biology.mpi-muenster.mpg.de/publications/TFBM/.

Materials and Methods

Our method for TFBM discovery first creates a dictionary of "DNA words" with high probability to be bound by TFs is employed to predict binding motifs using a new clustering method. Once binding motifs are predicted, the algorithm finds TF combinatorial binding patterns (CBPs). Figure 7 is an overview of our computational pipeline.

Compilation of "DNA word" dictionaries

The first step is to create a dictionary of DNA sequences of length w, with $w_{min} \le w \le w_{max}$ (where $w_{min}=4, w_{max}=16$), that includes potential binding sites for all the TFs. To compile this dictionary, our algorithm learns intrinsic properties (see Table 2) of known TFBSs and then generates a universal list of "DNA words" that are likely to be a TFBS. The input of the method is the multiple alignment format (MAF) dataset of 45 species provided by the UCSC genome browser [47], and known TFBMs from Jaspar [4] and Transfac [5] that are used as learning/validation data. To reduce the number of "DNA words", and thus, to decrease the computational burden we designed three filtering techniques, the conservation curve, the permutation prefix, and a merging method.

The first filtering method is based on the conservation curve (Figure 7 C), is employed in order to reduce the number of "DNA word" candidates. Note how the entropy inversely matches the conservation curve in this example. This entropy/conservation relationship does not necessarily have always to match, but it does reveal that the shape of the conservation profile carries an important intrinsic property of the TFBS.

The second filtering method is based on similarity permutations. For each "DNA word" of length w (Figure 7 A), we employ the order of similarity between the aligned sequences of 45 different species and the "DNA word" that is to be analyzed. This order creates a permutation. The similarity criteria used to create the permutations is based on the Hamming distance [56] (defined as the number of positions d that differ between two sequences). This permutation of species sequences is a valuable fingerprint to prune the search space as shown in Figure 7 B. The compilation dictionary algorithm keeps track of the closest species of the known

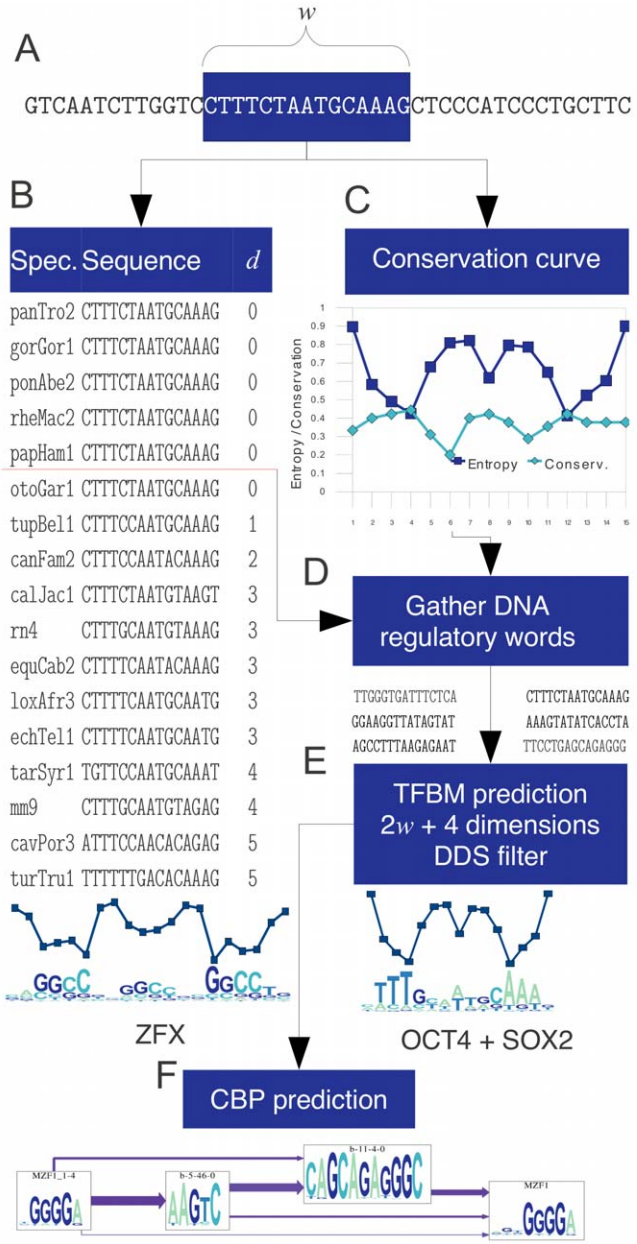


Figure 7. Method overview. (A) Extract "DNA words" of length w. (B) Filter the "DNA words" with the permutation prefix method, by ordering the aligned sequences of different specie genomes (Spec.) by the closeness d to the target sequence. The horizontal red line marks the end of the species permutation prefix. (C) Reduce the number of "DNA words" by the conservation curve. (D) Gather the intersection of "DNA words" generated in (B) and (C) with the merging method. (E) Predict the TFBM candidates using the DDS filter over a projection of $(2w+4)$ features (see Figure 8). (F) Take the predicted TFBMs to generate the catalog of combinatorial binding patterns of all the discovered motifs.
doi:10.1371/journal.pone.0049086.g007

binding sites, and uses them as learning data. Since the species order changes depending on the similarity of the aligned sequences, such alignment creates a permutation. The permutations that belong to known TFBSs become the learning set. Then, it is possible to filter each site only when the permutation is present in the learned set. Since a permutation of 45 elements will likely be unique for each site in the genome, we take as a prefix the first p

closest species, that is smaller than the total number of species. The prefix value p was set to 5 for DNA dictionaries greater than $w=10$ bases (for $w<10$ see Figure S1 B) in Supporting Information S1. In Figure 7 B, the red line represents the permutation prefix in the example.

The "DNA word" extraction process shown in Figures 7 B and 7 C creates two different sets of "DNA words". We take the intersection of them to feed the TFBM prediction algorithm using the merging method (Figure 7 D). This set intersection of "DNA words" is analogous to a universal dictionary of potential TF binding targets. Table S1 in Supporting Information S1 shows the resulting sizes for each dictionary generated by the conservation curve method, the permutation prefix method and the merging method. This reduction is crucial for increasing the quality of the results and for making it computationally feasible to generate the motif clustering. Once this dictionary is created, we proceed to predict TFBMs using clustering techniques (Figure 7 E).

Completion of the TFBM map through clustering

Once obtained a set B_w of "DNA words" of potential regulatory meaning, the next step is to extract biologically meaningful motifs. We expanded previous work on the subject of motif prediction [22–25] to integrate with a machine learning approach, not only comparative genomics, but also intrinsic TF properties. We assume that each "DNA word" in B_w is the center of a cluster of size k. This cluster is generated by obtaining the k closest (measured with a conglomerate distance) elements in B_w. The algorithm find the subset of elements that create an entropy curve that matches better the entropy curves generated from the training set T. Each cluster is ranked according to a quality criteria based on phylogenetic conservation and on average entropy. Then the predictions that do not follow the general patterns of motifs are eliminated using the new developed DDS filter. Finally, the clusters that share similar motifs are removed and the similar clusters are grouped.

Binding site cluster creation. Our TFBM discovery method is based on the search for signature "DNA words" using a similarity search strategy [57] using a novel sequence distance approach that we term "conglomerate distance". The idea behind this metric is to consider in the distance computation, the groups of identical sub-sequences shared between two sequences, providing extra weight to clusters of fragmented sub-sequences. Given two sequences s_1 and s_2 of length w we define its "conglomerate distance" as

$$D(s_1,s_2)=w^2-\sum_{i=1}^{c}L(s_i)^2, \qquad (1)$$

where c is the number of identical sub-sequences s_i between s_1 and s_2, and $L(s_i)$ is the length of each of such sub-sequences. The conglomerate distance is smaller when a larger number of contiguous bases or "chunks" is equal. It provides higher granularity than the Hamming distance and is specially useful for large sequence lengths. Thus, we have obtained better results with the "conglomerate distance" than with the Hamming distance. For a detailed explanation of the "conglomerate distance" see section S1.4.2 in Supporting Information S1 and pseudo-code in Figure S2 in Supporting Information S1. For each "DNA word" we take the top k closest "DNA words" based on the conglomerate distance (Eq. 1) and create an initial TFBM prediction. We refine the "DNA words" with a greedy filtering algorithm (see section S1.4.2 and pseudo-code in Figure S3 in Supporting Information S1), so that the entropy profiles match better known entropy profiles of other motifs. Figure 8 A is an example of this clustering. Each "DNA word" cluster becomes a TFBM prediction. We employed recent similarity search techniques [58,59] that allowed us to perform large scale searches at very high speed to obtain "DNA word" clustering. The algorithm projects each "DNA word" into a low dimensional space of 16 bits. This allows high speed similarity searches and the efficient construction of the nearest neighbors graph [60] of the "DNA word" dictionary.

Cluster prediction ranking. Once the clusters are fitted, they are ranked according tro two quality criteria. During the dictionary generation, we keep for each "DNA word" of length w, the number of times the word is conserved, c, and the number of its appearances, t. With the recorded c_i and t_i of each site i of the cluster, we calculate the phylogenetic conservation $score=\sum c_i/\sum t_i$ dividing the total number of conserved instances $\sum c_i$ by the total number of instances $\sum t_i$. Additionally, to deal with the case of large w in which sequences appear only once, we define a *cscore*, normalizing the total number of appearances divided by the maximum number of appearances $max(t)$. We add both scores to create a ranking value.

$$R=score+cscore=\frac{\sum c_i}{\sum t_i}+\frac{\sum t_i}{max(t)}. \qquad (2)$$

Higher values of R represent a better score. Finally, in the case of equal score, the order is decided by the average entropy:

$$AH(PWM)=\frac{-\sum_{j=1}^{w}\sum_{i=1}^{4}p_j(x_{ij})\log_4 p_j(x_{ij})}{w}, \qquad (3)$$

where x_{ij} are the elements of the PWM with w columns, and p_j is the probability of occurrence of the base represented by x_{ij} in the column j. Figure 8 B shows an example of this ranking criterion.

Binding site cluster filtering by dynamic dimension selection (DDS). Once the predictions are sorted by its relevance provided by the ranking criterion (Eqs. 2 and 3), the motifs that emerge naturally are filtered with a novel algorithm called DDS filter (Figure 7 E). This filter is necessary since the intrinsic properties describing the potential TFBM are of a different nature (see Table 2). Depending on each TFBM, the weight of their contribution to the cluster generation is different. The DDS adaptively searches which components of the feature space are most appropriate to cluster each potential motif. To determine which motif predictions are more likely to belong to TFs, we look at feature projections of the TFBMs. The filter is a machine learning algorithm that processes as learning data, known TFBMs. Then, it discriminates between predictions that follow intrinsic properties of TFBMs. For each motif of length of w, the filter works on the projection of the motif into a feature set of $(2w+4)$ fingerprints. Thelist of feature fingerprints is given in Table 2 .

Since the projected space is of high dimension, the curse of dimensionality [61] makes it difficult to discriminate between motifs, e.g., for $w=15$, it generates 34-dimensional fingerprint vectors and standard methods like one-class support vector machines [62] do not prune the search space effectively. To solve this filtering problem we developed the DDS filter. Given a d-dimensional object query $Q=(q_1,\ldots,q_d)$, for each component q_i the DDS orders the objects in the database according to their distance to q_i. These sorted lists of results $L=(l_1,\ldots,l_m)$ (where m is the number of object in the database) are iterated one list l_j at a time. The DDS records incrementally the object IDs found in each

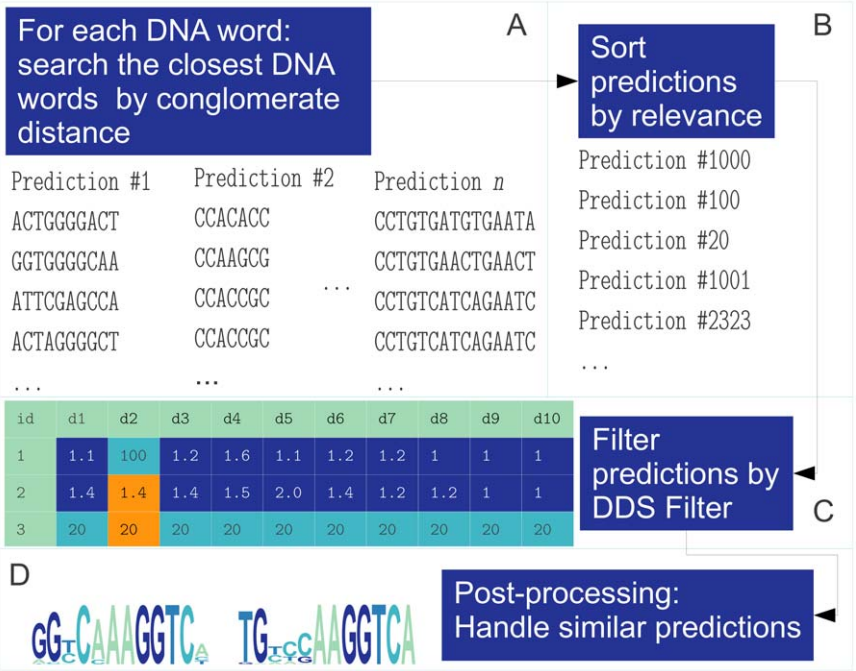


Figure 8. TFBM prediction method. (A) Cluster "DNA words" of length w by the conglomerate distance. A greedy algorithm removes sites that do not keep the motif within a certain distance of the learning set. (B) Sort predictions by the phylogenetic score given by Eqs. 2 and 3. (C) Filter those predictions that are far away from the learning set in the projected space of $(2w+4)$ features. An example of how the DDS filter works is represented in the table with three vectors of 10 components each. (D) Post-process similar predictions by grouping them if they are close ($\rho \geq 0.85$), and an example of two similar grouped predictions.
doi:10.1371/journal.pone.0049086.g008

component until it finds an object that appears a minimal number of times or until the search reaches a maximum number of iterations. An additional stage stores as a binary vector the components used to validate the training data. The algorithm also precomputes a binary vector B for all the objects in the database (that remains when the object in question is removed) and finds the closest mask in B for each query Q. The query is accepted to belong to a cluster if the Hamming distance between the closest vector of the cluster and the query is less than a threshold. The DDS filter and its pseudo-code implementation (Figure S4 in Supporting Information S1) is described in the section S1.4.2 in Supporting Information S1. The key property that DDS filter exploits to remove spurious TFBM predictions is that only a subset of the feature set is relevant to perform the filtering and this subset is different depending on the target object. Figure 8 C shows an example of how the DDS filter works for three vectors of 10 components each. Vector 2 is compared against vectors 1 and 3. The components of vector 2 that are closer to vector 1, are shown in dark blue. and those closer to vector 3 in orange. The Euclidean distance between vectors 1 and 2 is 98.60, and between vectors 2 and 3, 58.98. d_2 is the only component in which vectors 2 and 3 are closer. If we remove this component, the Euclidean distance between vectors 1 and 2 decreases to 1.01. The final clustering step purges similar predictions (those with more than 80% of shared sites) and clusters with low ranking score. Additionally, those clusters with Pearson correlation $\rho \geq 0.85$ are merged. Figure 8 D depicts two predictions that are combined into a cluster since $\rho \geq 0.85$. Once the prediction of TFBMs is complete, the output is a list of groups of "DNA word" clusters or TFBMs.

CBP prediction algorithm

After discovering TFBMs, and their respective binding sites, we uncover the syntax of the crosstalk among them on the genomic scale, estimating the combinatorial binding patterns (CBPs) of the TFBMs. To achieve this, we developed a computational method that finds common "motifs of motifs" based on the similarity of their topological features. The algorithm extracts the promoters that closely resemble an initial "query" promoter. Then, an iterative process links the motifs so that their multiple alignments are as large as possible and target as many genes as possible. We define a "group" as a set of motifs, each separated by not more than 1000 base pairs. The first step of the CBP generation algorithm is to create these groups. For each extracted group, the algorithm starts by searching for groups that share at least 3 motifs. Once this candidate list is made, it is ordered according to the degree of similarity shared with the group. The similarity is the number of shared motifs, and higher similarity is preferred. The algorithm greedily adds candidates. Next, we perform a multiple alignment using the center star algorithm [63]. First, this algorithm finds a center group and uses it to compute pairwise alignment with the other sequences adding spaces as needed. The pairwise alignment is calculated with the Smith Waterman aligner [64]. We set the difference cost to 5, and the open and extended gap penalties to 0. Motifs must align in at least 3 groups to be considered. Thus, we generate CBPs with motifs that are topologically aligned. We explain the pseudo-code (Figure S5 in Supporting Information S1) of the CBP prediction algorithm in detail in the section S1.5 in Supporting Information S1. Those TFBM predictions that occur frequently with others are extracted into CBPs which help us to decipher the common elements required in gene regulation (Figure 7 F). The motifs that appear in the generated CBPs set are extracted as the subset STFBM.

Ontology analysis and visualizations

We calculated the statistical significance of the ontologies of the list of gene targets associated with all the STFBM and the CBPs generated. To predict the gene targets of each STFBM, we used as in [65] the Berg-von Hippel method [66]. The GO terms were obtained from the AmiGO web server [67]. The statistical significance of the GO terms of each list of genes was analyzed using an enrichment approach based on the hyper-geometric distribution. The GO terms were backpropagated from the final term appearing in the gene annotation to the root term of each ontology. As a background set, we used the list of all the genes in the human genome with annotation on AmiGO. The multitest effect influence was corrected by controlling the false discovery rate, using the Benjamini-Hochberg correction. We developed the ''ontology maps'' concept to visualize the relationship between TFBMs or CBPs and ontology terms. For two sets of ontologies A and B, associated to TFBMs or CBPs, we calculated the ontology similarity with the following distance:

$$ontd(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}, \quad (4)$$

where $|A|$ is the cardinality of set A. We take the top k TFBMs or CBPs based on Eq. 4, to establish a topological graph linking those top closest objects. Finally, we create a link between each TFBM or CBP and their correspondent GO terms whose enrichment P-values satisfy a significance level α. We chose $k = 10$ and $\alpha = 0.00001$ (for cellular component GOs, $\alpha = 0.0001$). We use the GMAP algorithm [40] to visualize the topological graph.

Supporting Information

Supporting Information S1 Section S1 of this document contains a more detailed description of the materials and methods; section S2, additional ontology maps, validation results of TFBMs and CBPs, and lists of newly found STFBMs and CBPs; and section S3, a detailed description of the generated database files and their formats. The predicted catalog of regulatory elements can be downloaded from: `http://computational-biology.mpi-muenster.mpg.de/publications/TFBM/`
(PDF)

Acknowledgments

The authors would like to thank Daniela Gerovska, Karin Hübner, Rashel Grindberg, Holm Zaehres, David Obridge and all members of the Department of Cell and Developmental Biology of the Max Planck Institute for Molecular Biomedicine for fruitful discussions and comments during the preparation of this manuscript.

Author Contributions

Conceived and designed the experiments: MJAB. Performed the experiments: AJMM MJAB. Analyzed the data: AJMM MJAB. Contributed reagents/materials/analysis tools: AJMM MJAB. Wrote the manuscript: MJAB AJMM. Read and approved the final manuscript: AJMM HRS MJAB.

References

1. D'haeseleer P (2006) What are DNA sequence motifs? Nat Biotech 24: 423–425.
2. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252–263.
3. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology 14: 283–291.
4. Bryne JC, Valen E, Tang MHE, Marstrand T, Winther O, et al. (2008) JASPAR, the open access database of transcription factor binding profiles: new content and tools in the 2008 update. Nucleic Acids Research 36: D102–D106.
5. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research 34: D108–D110.
6. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720–1723.
7. Wunderlich Z, Mirny LA (2009) Different gene regulation strategies revealed by analysis of binding motifs. Trends in Genetics 25: 434–440.
8. Remenyi A, Schëoler HR, Wilmanns M (2004) Combinatorial control of gene expression. Nat Struct Mol Biol 11: 812–815.
9. Al-Kadi IA (1998) Origins of cryptology: the Arab contribution, Norwood, MA, USA: Artech House, Inc. pp. 93–122.
10. Das MK, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics 8 Suppl 7: S21.
11. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. Genome Research 16: 1455–1464.
12. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16: 939–945.
13. Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, et al. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. Bioinformatics 23: 1718–1727.
14. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369–373.
15. Siddharthan R, Siggia ED, van Nimwegen E (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1: e67.
16. Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 32: 199–203.
17. Elemento O, Tavazoie S (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol 6: R18.
18. Syed Z, Stultz C, Kellis M, Indyk P, Guttag J (2010) Motif discovery in physiological datasets: a methodology for inferring predictive elements. ACM Trans Knowl Discov Data 4: 2.
19. Bailey TL, Boden M, Whitington T, Machanick P (2010) The value of position-specific priors in motif discovery using MEME. BMC Bioinformatics 11: 179.
20. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, et al. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9: 447–464.
21. Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci USA 86: 1183–1187.
22. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nat 434: 338–345.
23. Kumar L, Breakspear A, Kistler C, Ma LJ, Xie X (2010) Systematic discovery of regulatory motifs in *Fusarium graminearum* by comparing four *Fusarium* genomes. BMC Genomics 11: 208.
24. Elemento O, Tavazoie S (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biology 6: R18.
25. Ettwiller L, Paten B, Souren M, Loosli F, Wittbrodt J, et al. (2005) The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. Genome Biology 6: R104.
26. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.
27. Nomenclature Committee of the International Union of Biochemistry (NC-IUB) (1986) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Mol Biol Evol 3: 99–108.
28. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16–23.
29. Ferraris L, Stewart AP, Kang J, Desimone AM, Gemberling M, et al. (2011) Combinatorial binding of transcription factors in the pluripotency control regions of the genome. Genome Res 21: 1055–1064.
30. Masui S (2010) Pluripotency maintenance mechanism of embryonic stem cells and reprogramming. Int J Hematol 91: 360–372.
31. Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, et al. (2007) Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9: 625–635.
32. Okumura-Nakanishi S, Saito M, Niwa H, Ishikawa F (2005) Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells. J Biol Chem 280: 5307–5317.
33. Yuan H, Corbi N, Basilico C, Dailey L (1995) Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes Dev 9: 2635–2645.

34. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131: 861–872.
35. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318: 1917–1920.
36. Vogel JH, von Heydebreck A, Purmann A, Sperling S (2005) Chromosomal clustering of a human transcriptome reveals regulatory background. BMC Bioinformatics 6: 230.
37. Banks EB, Crish JF, Welter JF, Eckert RL (1998) Characterization of human involucrin promoter distal regulatory region transcriptional activator elements-a role for Sp1 and AP1 binding sites. Biochem J 331 (Pt 1): 61–68.
38. Laury AR, Perets R, Piao H, Krane JF, Barletta JA, et al. (2011) A comprehensive analysis of PAX8 expression in human epithelial tumors. Am J Surg Pathol 35: 816–826.
39. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.
40. Gansner E, Hu Y, Kobourov S (2010) Gmap: drawing graphs as maps. In: Eppstein D, Gansner E, editors, Graph Drawing, Springer Berlin/Heidelberg, volume 5849 of Lecture Notes in Computer Science. pp. 405–407.
41. Potterf SB, Furumura M, Dunn KJ, Arnheiter H, Pavan WJ (2000) Transcription factor hierarchy in Waardenburg syndrome: regulation of MITF expression by SOX10 and PAX3. Hum Genet 107: 1–6.
42. Flores MV, Lam EY, Crosier P, Crosier K (2006) A hierarchy of Runx transcription factors modulate the onset of chondrogenesis in craniofacial endochondral bones in zebrafish. Dev Dyn 235: 3166–3176.
43. Singh H (1996) Gene targeting reveals a hierarchy of transcription factors regulating specification of lymphoid cell fates. Curr Opin Immunol 8: 160–165.
44. Jaenisch R, Young R (2008) Stem Cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132: 567–582.
45. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947–956.
46. Lobe C (1992) Transcription factors and mammalian development. Current Topics in Dev Biol 27: 351–383.
47. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2010) The UCSC genome browser database: update 2011. Nucleic Acids Research 31: 1–7.
48. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664.
49. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, et al. (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471: 216–219.
50. Chávez E, Figueroa K, Navarro G (2008) Effective proximity retrieval by ordering permutations. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 30: 1647–1658.
51. Skala M (2009) Counting distance permutations. Journal of Discrete Algorithms 7: 49–61.
52. Tellez E, Chavez E, Graff M (2011) Scalable pattern search analysis. In: Pattern Recognition, Springer Berlin/Heidelberg, volume 6718 of Lecture Notes in Computer Science. pp. 75–84.
53. Tellez ES, Chavez E (2010) On locality sensitive hashing in metric spaces. In: Proceedings of the Third International Conference on Similiarity Search and Applications. New York, NY, , USA: ACM, SISAP '10, pp. 67–74.
54. Tung AKH, Zhang R, Koudas N, Ooi BC (2006) Similarity search: a matching based approach. In: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, VLDB '06, pp. 631–642.
55. Aggarwal CC, Yu PS (2000) The igrid index: reversing the dimensionality curse for similarity indexing in high dimensional space. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, , USA: ACM, KDD '00, pp. 119–129.
56. Deza MM, Deza E (2009) Encyclopedia of Distances. Berlin Heidelberg: Springer.
57. Zezula P, Amato G, Dohnal V, Batko M (2005) Similarity Search: The Metric Space Approach. Secaucus, NJ, , USA: Springer-Verlag.
58. Müller-Molina AJ, Shinohara T (2009) Efficient similarity search by reducing i/o with compressed sketches. In: SISAP. IEEE, pp. 30–38.
59. Müller-Molina AJ (2009) Obsearch: a high performance similarity search engine for java. In: Proceedings of the 2009 Second International Workshop on Similarity Search and Applications. Washington, DC, , USA: IEEE Computer Society, SISAP '09, pp. 143–145.
60. Samet H (2005) Foundations of Multidimensional and Metric Data Structures. San Francisco: Morgan Kaufmann Publishers Inc.
61. Chavez E, Navarro G, Baeza-Yates R, Marroquin JL (2001) Searching in metric spaces. ACM Comput Surv 33: 273–321.
62. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New Support Vector Algorithms. Neural Comput 12: 1207–1245.
63. Gusfield D (1997) Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. New York: Cambridge University Press.
64. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147: 195–197.
65. Sarkar D, Siddiquee KAZ, Araúzo-Bravo MJ, Oba T, Shimizu K (2008) Effect of *cra* gene knockout together with *edd* and *iclr* genes knockout on the metabolism in *Escherichia coli*. Archives of Microbiology 190: 559–571.
66. Berg OG, von Hippel PH (1987) Selection of dna binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of Molecular Biology 193: 723–750.
67. Ashburner M, Ball CA, Blake JA, et al (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 25–29.
68. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219–232.
69. Wu J, Sieglaff DH, Gervin J, Xie XS (2008) Discovering regulatory motifs in the Plasmodium genome using omparative genomics. Bioinformatics 24: 1843–1849.