BMC Genomics

Research article

Open Access

Coding limits on the number of transcription factors
Shalev Itzkovitz[1,2], Tsvi Tlusty[2] and Uri Alon*[1,2]

Address: [1]Dept. Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel and [2]Dept. Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

Email: Shalev Itzkovitz - shalev.itzkovitz@weizmann.ac.il; Tsvi Tlusty - tsvi.tlusty@weizmann.ac.il; Uri Alon* - uri.alon@weizmann.ac.il
* Corresponding author

This article is available from: http://www.biomedcentral.com/1471-2164/7/239

Abstract

Background: Transcription factor proteins bind specific DNA sequences to control the expression of genes. They contain DNA binding domains which belong to several super-families, each with a specific mechanism of DNA binding. The total number of transcription factors encoded in a genome increases with the number of genes in the genome. Here, we examined the number of transcription factors from each super-family in diverse organisms.

Results: We find that the number of transcription factors from most super-families appears to be bounded. For example, the number of winged helix factors does not generally exceed 300, even in very large genomes. The magnitude of the maximal number of transcription factors from each super-family seems to correlate with the number of DNA bases effectively recognized by the binding mechanism of that super-family. Coding theory predicts that such upper bounds on the number of transcription factors should exist, in order to minimize cross-binding errors between transcription factors. This theory further predicts that factors with similar binding sequences should tend to have similar biological effect, so that errors based on mis-recognition are minimal. We present evidence that transcription factors with similar binding sequences tend to regulate genes with similar biological functions, supporting this prediction.

Conclusion: The present study suggests limits on the transcription factor repertoire of cells, and suggests coding constraints that might apply more generally to the mapping between binding sites and biological function.

Background

Transcription factor proteins regulate genes by binding DNA sequences at the promoters of the target genes. Typically, each transcription factor (TF) is able to recognize a set of similar sequences, centred around a consensus sequence [1-11]. The binding probability is generally believed to be higher the more similar a sequence is to the consensus sequence.

Transcription factor proteins can be classified into super-families, each with a different DNA-binding mechanism [12-16]. For example, the winged helix super-family consists of proteins which insert an alpha helix into the major groove of the DNA, forming amino-acid-base contacts over a region spanning about 5–6 base-pairs. These proteins tend to form homo-dimers, which often contact two consecutive major grooves [13]. Thus, their binding sequences are palindromic repeats of a 5–6 base-pair sequence. Proteins from the homeodomain-like super-

family insert an alpha helix parallel to the DNA backbone, and tend to form heterodimers, thus forming more base-pair contacts than the winged helix proteins. Other super-families like the C2H2 zinc-coordinating super-family are monomer proteins with variable number of DNA recognition domains, or 'fingers', each recognizing 3 consecutive base-pairs [17,18].

The total number of transcription factors (TFs) encoded by a genome increases with the number of genes in the genome [19]. The number of TFs has been shown to scale with genome size as a power-law (the number of TFs, N, scales as the number of genes G as $N \sim G^{1.9}$ for Prokaryotes and $N \sim G^{1.3}$ for Eukaryotes [20]). This is thought to reflect the fact that the more complex the organism, the more intricate the regulation needed to respond to environmental inputs and to carry out developmental programs.

Here we ask whether there are limits on the numbers of transcription factors from different super-families. We find that the maximal numbers of transcription factors from most super-families is significantly smaller than the total number of transcription factors. The maximal number for each super-family appears to correlate with the number of possible sequences effectively recognized by the binding mechanism of that super-family. We also show that the binding sequences of different TFs from the same super-family are often highly similar, and that TFs with similar binding sequences tend to participate in similar biological processes. The results are compared to simple coding models that may provide an intuitive understanding of the origin and magnitude of the bounds on TF numbers.

Results

Maximal number of transcription factors in most super-families is bounded

The total number of TFs increases with the number of genes in the genome [20], exceeding 2700 TFs in organisms such as *Xenopus tropicalis*. However, when considering each TF super-family separately, we find that the number of TFs from most super-families reach a maximum size which is significantly lower than the total number of TFs (in other words, the size of the super-family is bounded). For example, winged helix transcription factors increase with number of genes until reaching a maximum of about 300 TFs in organisms with ~5000 genes (Table 1). Larger genomes contain winged helix TFs but at numbers which do not appear to exceed this bound.

A similar picture is observed in other super-families (Table 1), for example, the maximal number of lambda-repressor-like TFs is about 80 and the maximal number of helix-loop-helix proteins is about 185. In one super-family – the multi-domain C2H2 zinc finger super-family, the maximal number of TFs is significantly higher than in other super-families. These proteins, found mainly in eukaryotes, increase in number with genome size, following genome size as about the number of genes squared (G^a with a = 1.8+/- 0.17).

The maximal number of TFs correlates with number of degrees of freedom in the binding mechanism

In this section we compare the magnitude of the maximal TF numbers with the number of degrees of freedom in the binding mechanism of each super-family. The results are summarized in Fig 1 and Table 1.

The number of degrees of freedom of a binding mechanism is related to the number of different base-pairs that can be specifically recognized by the DNA-binding mech-

Table 1: Maximal numbers of transcription factors from each super-family in a single organism, and the organism in which the maximum is observed.

	Super-family	Maximal # TFs	Kingdom	organism	P	S	O	H	# sequences
1	lambda repressor-like DNA-binding domains	77	A,B,E	*Photorhabdus luminescens*	3	1	2	1	64
2	C-terminal effector domain	88	A,B,E	*Streptomyces avermitilis*	-	-	-	-	-
3	srf-like	122	E	*Arabidopsis thaliana*	-	-	-	-	-
4	helix-loop-helix DNA-binding domain	186	E	*Arabidopsis thaliana*	2	1	1	2	128
5	DNA-binding domain	194	B,E	*Oryza sativa*	-	-	-	-	-
6	Zn2/Cys6 DNA-binding domain	246	E	*Fusarium graminearum*	3	13	3	1	1,248
7	winged helix DNA-binding domain	299	A,B,E	*Bordetella bronchiseptica*	6	1	1	1	2,048
8	glucocorticoid receptor-like DNA-binding domain	376	A,B,E	*C.elegans*	2	9	3	2	3,456
9	homeodomain-like	417	A,B,E	*Danio rerio*	6	1	1	2	8.4×10^6
10	multi-domain C2H2 zinc fingers	1308	E	*Mus musculus*	6–30	1	1	1	-

The kingdom in which each super-family is observed is abbreviated as A – Archea, B – Bacteria, E – Eukaryotes. Estimates for the number of possible sequences are shown (see methods). P – number of variable positions in each half-site, S – number of possible spacing between half-sites, O – number of possible orientations, H – homo-dimers (1) or hetero-dimers (2). The number of sequences is $4^{P*H*O}*S/2$.
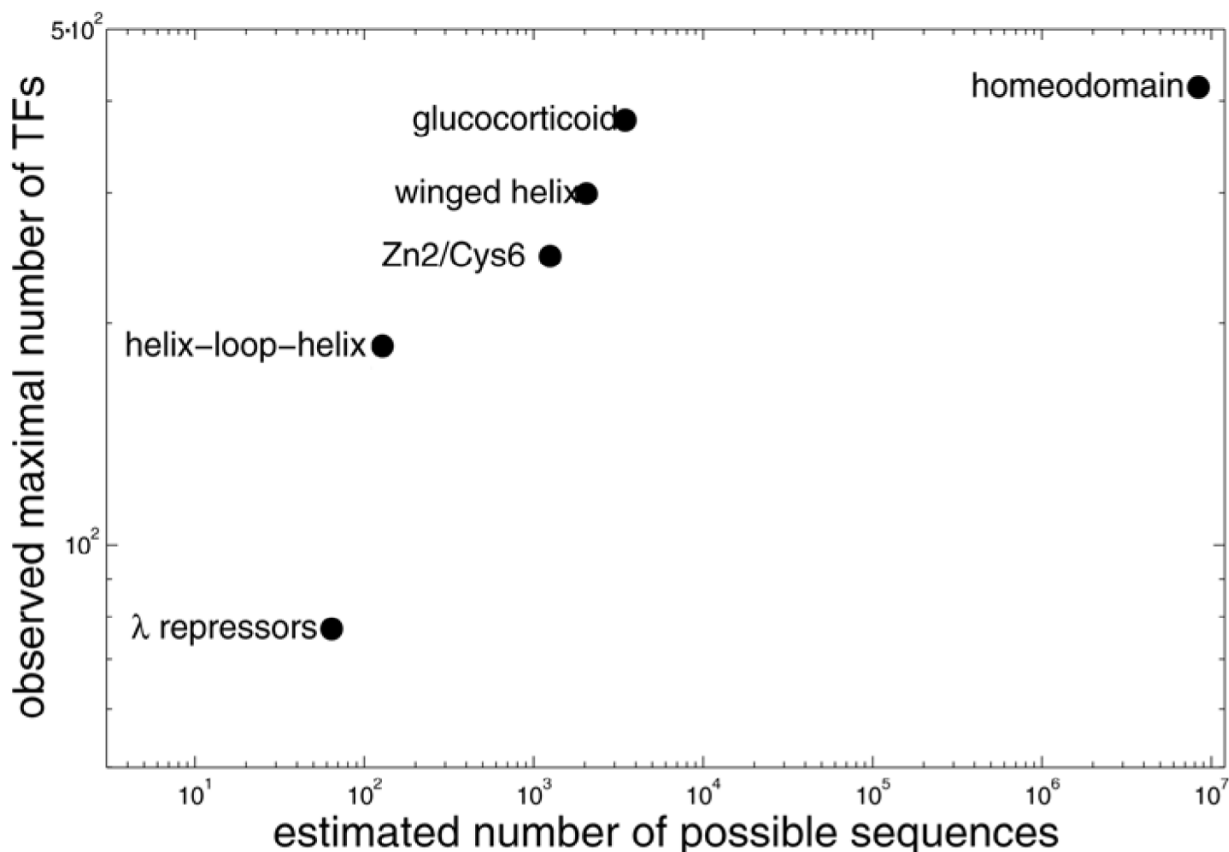
Figure 1
Correlation between the maximal number of transcription factors and number of possible sequences for six super-families, for which details of binding mechanism are known.

anism. For example, lambda-repressor-like proteins recognize DNA by inserting a short alpha helix into the major groove, specifically recognizing only three base-pairs [21,22]. These three base-pairs essentially determine the binding sequence of the TF, because these proteins usually form homo-dimers in which each monomer recognizes the same sequence [13]. Thus the proteins from this super-family have a relatively constrained binding mechanism, with 64 possible binding sequences (since there are $4^3/2$ combinations of three bps, including reverse complement sequences, and there are two possible orientations of the half sites, see methods). The maximal number observed for proteins in this super-family is about 80 per genome (Table 1).

Winged helix (wH) transcription factors recognize DNA using a similar mechanism of inserting an alpha helix into the major groove. However, the longer alpha helix used by these TFs usually interacts with 6 bp positions instead of 3 bp. Just as in lambda-repressor like TFs, these 6 base-pairs determine the binding sequence, because wH pro-

teins usually form homo-dimers [13]. There does not appear to be any constraint on the possible 6 base pair sequences that can be recognized by a suitably designed wH protein. Hence, the maximal number of different sequences that can be recognized by such factors can be estimated as $4^6/2$ = 2048, more than the number of sequences available for lambda-repressor like TFs. The observed maximal number for this super-family, about 300 (Table 1, Fig 1) is higher than the maximal number for the lambda-repressor like super-family.

Three other super-families have related mechanisms: helix-loop-helix proteins, Zn2/Cys6 proteins and glucocorticoid receptor-like proteins (Table 1). These three super-families bind as dimers, in which each monomer binds a highly constrained sequence (half-site). Helix-loop-helix proteins usually recognize one of only a few conserved major-groove hexamer sequences, such as the E-box or G-box [23-25]. In these sequences, only two positions are variable. These proteins can form homo-dimers or hetero-dimers. This is the most constrained of

the three super-families (~130 possible sequences), and has the lowest observed maximal number, about 185. Zn2/Cys6 proteins bind to three bp identical half-sites. They have more possible binding sequences than helix-loop-helix proteins, because the half-sites can be at variable spacing and orientations (estimated at ~1250 possible sequences). The maximal number for this super-family is higher, about 250 (table 1, Fig 1). Glucocorticoid receptor-like proteins bind two half sites which can be at variable orientations and spacing [21] and in addition can form hetero-dimers that bind to non-identical half-sites. This super-family therefore has the most degrees of freedom of the three super-families (~3450 possible sequences), and displays the highest maximal number, about 380.

C2H2 proteins have between two to more than 30 finger domains, each recognizing three consecutive base-pairs [18,26]. These proteins have the largest number of possible binding sequences ($64^n/2$ for an n-domain protein). The maximal number of such proteins in a single organism is the highest of all super-families, consistent with the large number of degrees of freedom for the possible binding sites.

Table 1 and Fig 1 show the maximal numbers for all super-families, as well as an estimate for the number of possible sequences where data on the mechanism is available. The maximal number seems to increase with the number of possible binding sequences.

Evolutionary shift to super-families with more degrees of freedom

When examining the distribution of transcription factors among the different super-families for different organisms (Fig 2) one can observe a shift from the predominant use of super-families with less degrees of freedom and smaller maximal number of TFs to super-families with more degrees of freedom and higher maximal number of TFs. Organisms with a small number of genes predominantly use TF super-families such as Lambda-repressor like and C-terminal effectors, while as organisms with more genes shift to Zn2/Cys6 and Glucocorticoid receptor-like super-families, which have higher maximal numbers of TFs. Organisms with more genes, such as mouse and human, shift to the predominant use of C2H2 multi-domain zinc finger TFs, which have the highest maximal number.

Coding theory suggests upper bounds for transcription factor numbers

What is the origin of the bounds on the numbers of transcription factors? As one possible explanation, we consider the mapping of transcription factors to binding sequences as a coding problem, analogous to the assignment of amino acids to codons in the genetic code. We would like to emphasize that the purpose of models in this study is not to serve as descriptions of precise biochemical mechanisms, but rather as simple conceptual guides to understand the forces at play. Thus, the models neglect such details as whether each residue in the binding domain of the TF recognizes one or more DNA bases, as well as issues of DNA malleability, non-specific amino-acid base contacts, etc. The models also neglect important effects such as cooperative binding between TFs and other regulatory features.

To begin with, consider a hypothetical situation in which each sequence is assigned a different TF. In the case of winged helix TFs, for example, in which a binding sequence is effectively of length 6, there are in principle $4^6/2 = 2048$ different sequences (or 2080 if one considers separately the 64 self-complementary sequences in addition to $(4^6-64)/2$ unique non-self complementary sequences). There are therefore a maximal number of 2048 TFs that are perfectly stringent and recognize only one sequence. In reality, however, each TF recognizes a set of sequences, located around a consensus sequence [1-11]. Thus, the assignment of TFs to sequences should assign to each TF a set of adjacent sequences. This raises a difficulty, because TFs with very similar sets of binding sequences can recognize each others binding sequences.

In the theoretical case of perfectly non-overlapping sequences, in which each sequence is assigned to only one TF, Sengupta et al [6] have noted that the coding problem is similar to the problem of packing non-overlapping spheres in the space of sequences, each sphere corresponding to the sequences belonging to one TF (Fig 3a). Let us make a simple estimate of the number of TFs according to this picture. As an example, suppose that a TF can on average recognize sequences that are one letter different from the consensus (Hamming distance of one away from the consensus sequence). For winged helix proteins, there are six positions in the sequence, each of which can be changed to one of 3 other letters, resulting in $6 \times 3 = 18$ neighbours that are a Hamming distance of one away. Thus, there can be at most 2048/19~100 distinct TFs with non-overlapping sequences [27]. This is on the same order of magnitude, although lower than the observed maximal number of about 300 (Table 1).

Coding theory suggests that one can increase the number of TFs by allowing sequences to overlap. This comes at a cost: TFs can mis-recognize each other's sequences leading to errors in gene expression. Optimal codes that can minimize such errors are known as "Gray codes" in information theory [28]. An optimal coding theory, which allows sequences to overlap, has been recently suggested in the context of the genetic code [29]. In the genetic code, codons differing by one base-pair correspond either to the
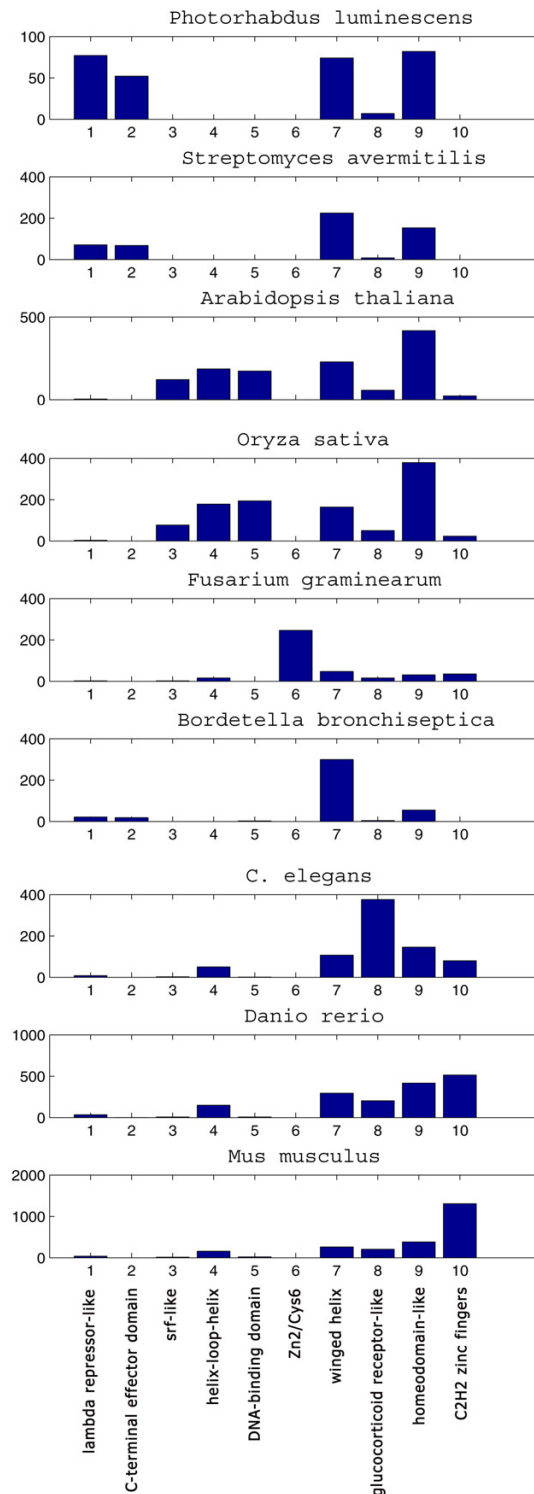
Figure 2
Distribution of transcription factors for the 9 organisms in Table 1 among the different super-families. On the x axis are the 10 super-families of table 1, on the y axis their counts in each organism. The organisms are sorted according to increasing number of genes in the genome. Note that the y-axis scale is different for different organisms.
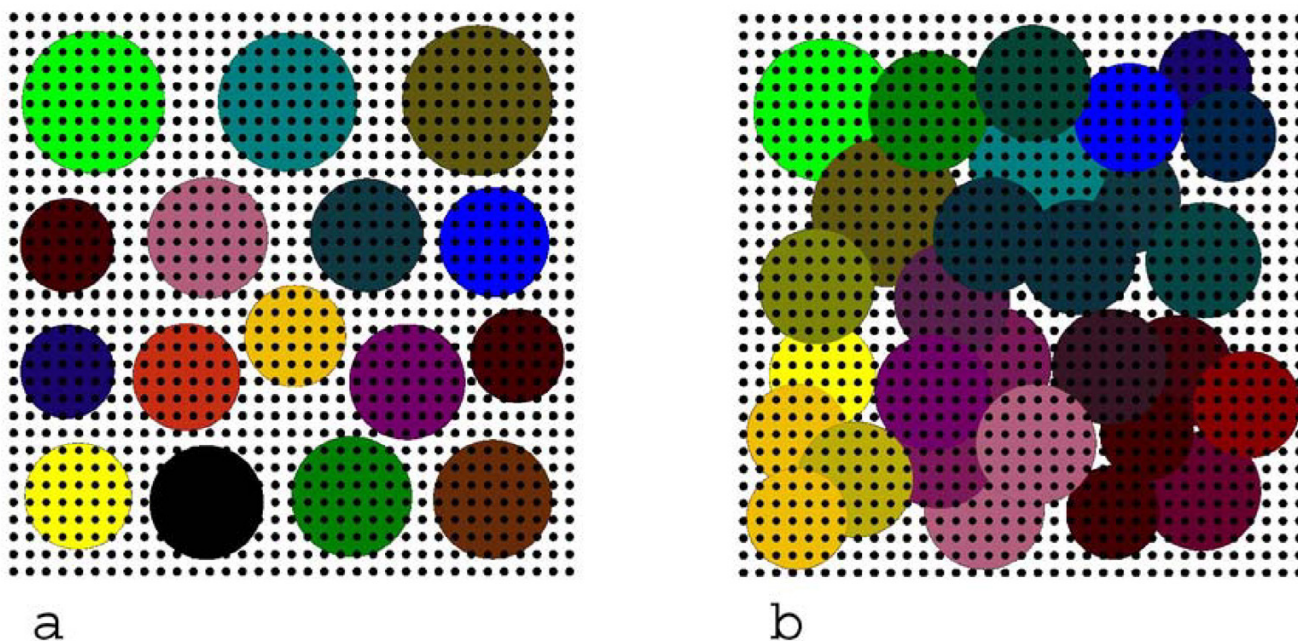
Figure 3
Conceptual coding schemes for the assignment of binding sequences to TFs. Binding sequences are displayed as points, TFs as colored spheres. Colors correspond to the biological function of each TF. a) A sphere-packing code – code-words are covered by non-overlapping spheres. The TFs do not share binding sequences. b) A smooth code – code-words are covered by overlapping spheres with similar function. TFs can share binding sequences with neighbor TFs. This type of code is predicted to be smooth, that is where TFs with shared binding sequences tend to have similar biological function, represented by spheres of similar color in the figure.

same amino acid or to chemically related amino acids [30-32]. This mapping is thought to minimize the error load caused by errors in translation [29,32].

Here we apply this theory to TFs. Since the theory takes into account the mis-binding errors, it can reach higher bounds than hard-sphere packing codes (Table 2). Importantly, the theory predicts that neighbouring "spheres", that is TFs with similar binding sequences, would tend to be close in biological function in order to minimize the error load. Thus, the TFs with overlapping sequences should regulate the same genes, or genes with similar

Table 2: Theoretical bounds for an n-length 4-letter code.

n	# code words – $4^n/2$	Coloring bound	Sphere packing bound
3	32	18	3
4	128	42	9
5	512	95	32
6	2,048	210	107
7	8,192	460	372
8	32,768	994	1310

The sphere packing bounds are from equation (10). The coloring bound is given by equation (8).

functions, so that effects of cross-recognition are minimized. Such codes are called "smooth" (Fig 3b).

Factors with similar binding sequences tend to have similar biological functions

A qualitative prediction of the smooth coding theory is that transcription factors with similar recognition sequences should tend to have similar biological effects. The reason is that factors with similar sequences can sometimes bind to each others sequences. If the biological effects of binding of TF A and B are similar, the reduction in fitness caused by such errors would be smaller. Hence, there may be a selection pressure to allocate similar sequences to biologically similar factors.

To test the prediction that TFs with similar sequences should tend to have similar function, we examined TFs in *E. coli*, yeast and human, and compared their sequence similarity by means of several distance metrics. In these organisms there exists a significant sequence similarity between the binding sites of some TF pairs (Fig 4, 5, 6). The yeast set of 94 well characterized TFs contained 18 pairs with highly similar sequences (Fig 4). The *E. coli* set of 46 well characterized TFs contained 6 pairs with highly similar sequences (Fig 5). The human set of 49 TFs con-
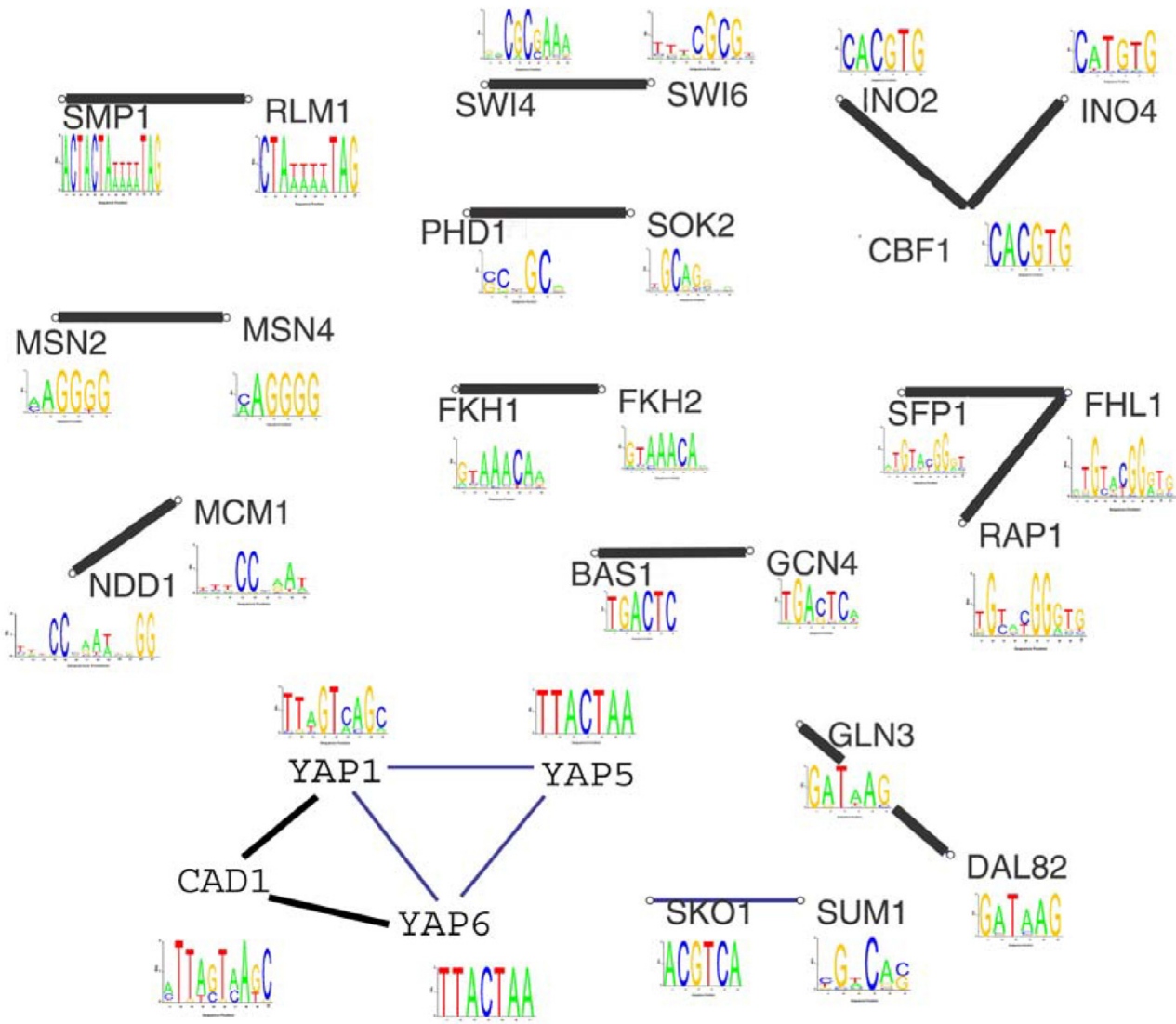
Figure 4
Transcription factors with overlapping binding sequences in *S. cerevisae*. Nodes represent TFs, edges connect pairs of TFs if their corresponding sets of binding sequences have significant overlap according to the present measure. Bold edges connect TFs which also have biological similarity according to the functional annotation and transcription network (gene co-regulation) measures. Shown are the TF logos [11]. Logo length was limited to the highly conserved base pairs for clarity.

tained 9 pairs with highly similar sequences (Fig 6). In other words, the TF "spheres" often overlap significantly (Fig 3B).

To assess the similarity in function of the TFs with similar binding sequences, we used two measures. The first similarity measure was a significant co-regulation of target genes by both factors. The second measure was the similarity in functional annotation [33,34] of the target genes of each TF. For both measures, we observed a significant enrichment of TF pairs with similar sequences and similar biological function measures.

We now provide more details on this result. To assess the functional similarity in yeast, we used an experimentally determined transcription network [35,36]. This network contained targets for 64 TFs in our data-set. About 14% (276/2016) of all TF pairs had significant target co-regulation. When considering pairs with similar binding sequences, the fraction with significant target co-regulation increases to over 50% (10/18, p-value of $5.1*10^{-5}$) (Fig 4).

As a second measure of functional similarity, we assigned to each yeast TF a profile according to the functional
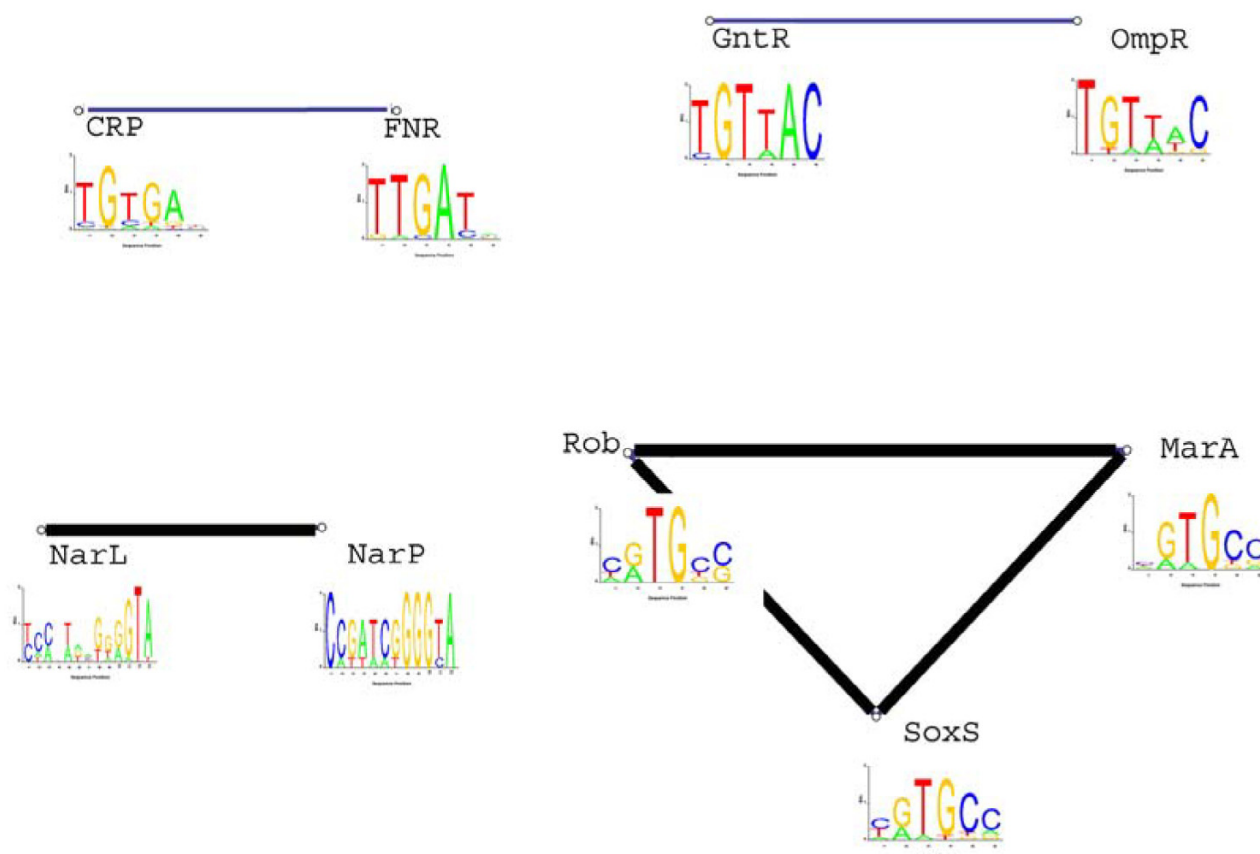
Figure 5
Transcription factors with overlapping binding sequences in *E. coli*. Nodes represent TFs, edges connect pairs of TFs if their corresponding sets of binding sequences have significant overlap according to the present measure. Bold edges connect TFs which also have biological similarity according to the functional annotation and transcription network (gene co-regulation) measure. Shown are the TF logos [11]. Logo length was limited to the highly conserved base pairs for clarity.

annotation of its target genes [33]. We then compared the average distances between the profiles of TF pairs with similar binding sequences to the average distances between the profiles of all TF pairs. We find that TF pairs with similar binding sequences have a lower average profile distance (0.17 vs 0.35, p-value of $7*10^{-5}$). Thus their targets tend to have more similar biological annotations. Examples of such pairs are the stress-response regulators MSN2 and MSN4, the stress-response YAP TFs, the cell cycle regulators FKH1 and FKH2 and the nitrogen regulators GLN3 and DAL82.

Similar results were obtained for *E. coli*. In our database of 46 TFs we found 6 pairs of TFs with significantly high binding sequence similarity (Fig 5). To assess the functional similarity between these pairs, using co-regulation criterion, we parsed the Ecocyc database [34] to obtain a network of 541 operons and 806 experimentally verified transcription interactions. Only 4% (39/1035) of all TF pairs had significant target co-regulation. When considering TFs with similar binding sequences, this fraction increases to over 65% (4/6, p-value $2.5*10^{-5}$). As a second test, we used a functional annotation for *E. coli*. We found that TF pairs with similar sequences have a lower average functional profile distance (0.26 vs 0.76, p-value of $4.2*10^{-5}$). Thus, the target genes of these TFs tend to have similar biological annotations. Examples of TF pairs with similar sites and similar functions include the drug and stress response regulators MarA, SoxS and Rob that jointly regulate at least 6 operons, and the anaerobic metabolism regulators NarL and NarP that jointly regulate 5 operons. The similarity in sequences is so large that some of these factors bind the exact same sequence in some of their co-regulated genes. However, many of their binding sequences are also distinct (the TF spheres appear not to overlap completely).
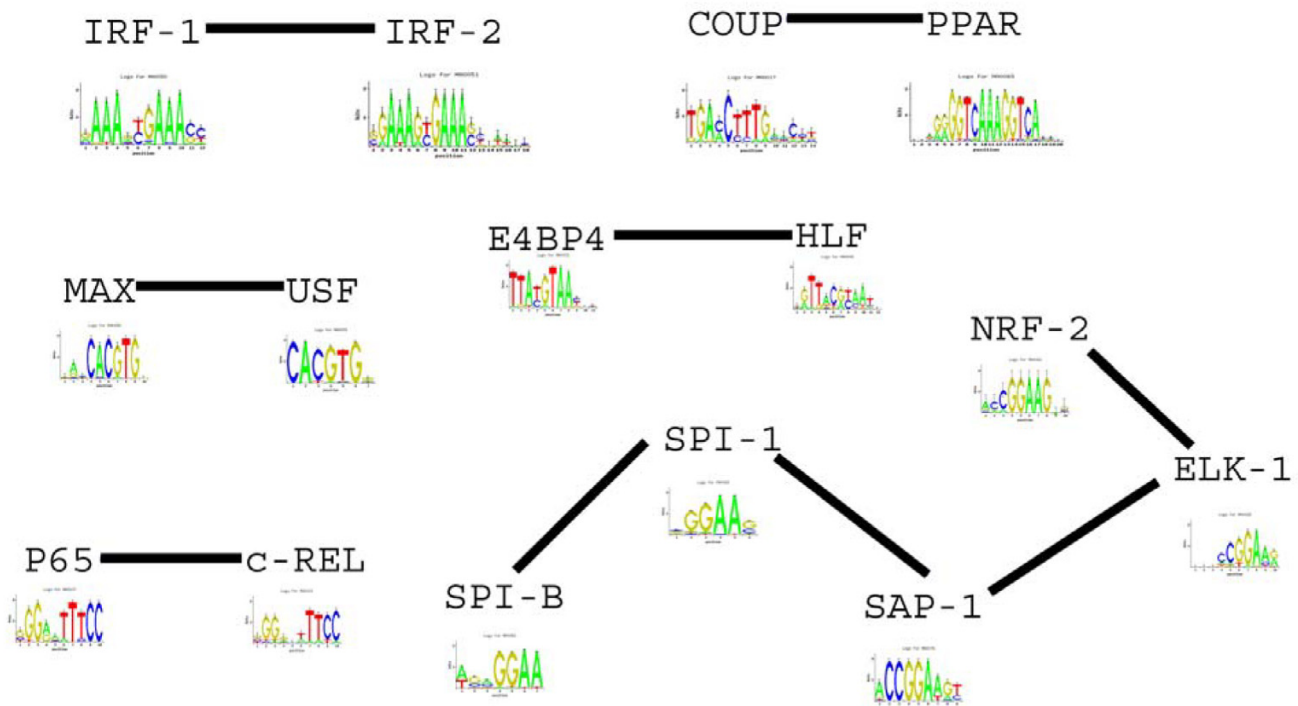
Figure 6
Transcription factors with significantly overlapping sequences in Humans. Edges connect two TFs with similar binding sequences. Sequence logos are shown for each TF.

For human, limited data currently prohibits a detailed statistical analysis. However, several examples are known where functionally related TFs recognize the same sequences. These include the interferon regulatory factors IRF-1 and IRF-2 [37], and the ETS transcription factors SPI-B and SPI-1 [38].

Many TF pairs with overlapping sequences and similar function are close paralogs (belong to the same family within the super-family) [15,39-42]. In most cases there are additional paralogous TFs with non-similar binding sequences. For example the *E. coli* regulators MarA, SoxS and Rob that recognize similar sequences, are all paralogs from the AraC/XylS family. However, they differ in their binding sequences and in their biological function from their paralog AraC. This suggests that homology of TF proteins does not fully explain the similarity in their binding sites. Gene duplication may aid in generating paralogous TFs with similar binding sites, which can then be selected according to the cost and benefit of their action on the target genes.

Discussion
The main result of this study is that the maximal numbers of TFs from most transcription factor super-families appear to be bounded. The number of these TFs in a genome does not seem to exceed a certain upper bound. These bounds range from around 80 for lambda repressor-like, to about 420 for homeodomain proteins. The bounds seem to correlate with the number of degrees of freedom of the DNA-binding mechanism in each super-family. The multi-domain C2H2 zinc fingers super-family displays a significantly higher maximal number in the present data, compared to other super-families.

To understand these bounds, we considered the coding problem faced by the cell: how to assign different sequences to each transcription factor in a way that avoids erroneous recognition in which a transcription factor binds where it shouldn't. As organisms of increasing complexity evolve there is a need for more diversity in gene regulation, through the introduction of new transcription factors. The stochastic process of DNA recognition by TFs may result in binding of a TF to binding sequences intended for another TF, if these are similar enough. This study examined the proposal that minimizing these mis-recognitions limits the maximal numbers of TFs with a given binding mechanism, that is, TFs from a given super-family.

To examine the coding problem on a qualitative level, we considered two theoretical mappings of code-words

(DNA sequences) to messages (TFs): a sphere packing code in which each TF has unique sequences not shared with other TFs, and a smooth code in which sequences of different TFs can partially overlap. The latter appears to offer a more realistic representation, because many pairs of TFs have highly similar code-words. For some super-families, the observed bounds seem to approximately agree with the theoretical bound derived for smooth codes.

A prediction of the theory in which misrecognition errors are an important constraint on the coding, is that TFs with similar binding sequences should tend to have similar biological functions. This is because misrecognition between such TFs would have a smaller impact on fitness. This prediction agrees with the present observation that the TFs with overlapping code-words are significantly closer in their biological function than expected at random.

One possible scenario for the evolution of TF super-families is as follows. Simple organisms, which require few TFs, employ certain super-families such as lambda repressor-like and winged helix. When these super-families reach their upper bounds, new super-families are needed. At these points organisms shift their TF usage to novel super-families with more degrees of freedom and higher maximal numbers (Fig 2). An example is the increased use of the C2H2 zinc finger TFs in the more advanced organisms.

It is important to note that the usage of different TF super-families is linked to the phylogenetic grouping of organisms. An example is the Zn2/Cys6 DNA-binding domain TFs which are largely restricted to fungal organisms [14,43]. This hints at an interplay between the suggested coding limits that create pressure to use different TF super-families, and the pyhlogenetic history that might dictate the specific choice of these TF super-families.

Much of the innovation of biological function has been attributed to events of gene duplications [44-47]. Several species, such as zebra-fish and Arabidopsis have undergone whole genome duplications. Such duplication events may lead to a situation where the number of TFs from a given super-family exceeds its theoretical bound, as observed for helix-loop-helix proteins in Arabidopsis. This could either create an evolutionary pressure which may lead to eventual loss of redundant TFs or result in a transient, non-equilibrium census of TFs.

To further test the present conclusions requires additional biological data. The current sets of TFs with known binding sequences (the transcriptional code) and gene-regulation networks, representing the functional annotation of the TFs, are still partial. For example, the current dataset for *E. coli* includes less than a fifth of all TFs in the organism. Once datasets are enlarged, one may get a better estimate of the bounds, the amounts of overlap in sequence space, and the functional smoothness of the transcriptional code.

Along the same lines, further knowledge of TF-DNA binding mechanisms could allow one to obtain more accurate estimates for the number of degrees of freedom of each super-family in order to more accurately test the correlation of the observed bounds with this number. In the case of C2H2 zinc fingers, detailed structures of the TF-DNA complex allowed a good estimation of the mapping between residues in the TF binding domain and the DNA bases recognized by the TF [17,18]. In other super-families, however, no clear mapping has yet been devised. Therefore, the present study presented only crude estimates for the number of possible sequences of each super-family. The present study predicts that the number of degrees of freedom in super-families with a lower observed bound should be smaller than for super-families with high bounds.

The present study focused only on one level of TF-DNA interaction, the recognition mechanisms of DNA binding sequences by transcription factors. There are many additional effects that govern transcription regulation. Gene expression often depends on the combined effects of multiple TFs, integrated by a cis-regulatory input function at the promoter [48-55]. The functional role of each TF is influenced by the distance on the promoter of its binding sites from sites of other TFs and the transcription start site [51], as well as the phase of the site along the DNA helix [56,57]. In addition, there is often co-operative binding to other factors [58,59], tissue-specific TF expression [19] and differential exclusion of TFs from the nucleus that is dependent on cell type and conditions [60]. By effectively introducing more degrees of freedom into the binding mechanism, these additional effects may also alleviate the constraints anticipated by the high levels of sequence overlap observed in the present study. These effects may explain the abundance of TFs from super-families like the lambda repressor-like and helix-loop-helix TFs, for which the observed maximal number of TFs are higher than the expected bounds in the simplest coding models.

The creation of maximal diversity of TFs with minimal misrecognition error-load might not be the only factor underlying the smooth codes observed in this study. Assigning TF pairs with similar biological function to similar binding sequences may have additional functional advantages. Some of the TF pairs with overlapping binding sequences and similar biological function presented in this study, such as the yeast MSN2 and MSN4 stress

response regulators, can partially backup each others function. Other TF pairs, such as the IRF-1 and IRF-2 in humans have an antagonistic regulation mode, where one activates and another represses the same target genes on different time-scales [37]. This kind of regulation may create a transient activation profile, where target genes are activated for a short time following induction. TF redundancy [61] and antagonistic regulation may form additional 'forces' pulling TF sequence sets together, increasing the number of TFs above the strict coding bounds.

Conclusion

In conclusion, the present study suggests that there are upper bounds on the number of transcription factors from different super-families. It seems that the more constrained the binding mechanism, the lower the bound. The present bounds may be understood in terms of an optimal coding strategy, in which misrecognition errors are minimized. As predicted by such a theory, the present data suggests that TFs with similar binding sequences tend to regulate genes with similar biological functions. More generally, similar coding problems may occur in other recognition problems in biology, such as protein-RNA recognition and protein-protein interactions through defined protein recognition motifs [62,63]. Coding constraints can potentially limit the number of different protein binding motifs of a given type in the cell, in order to avoid non-specific cross recognition. It would be interesting to extend the present approach to these and other molecular recognition systems.

Methods

Transcription factor numbers

We focused on ten major super-families of transcription factors: lambda repressor-like, C-terminal effector domain of the bipartite response regulators, winged helix, srf-like domains, DNA binding domain (GCC box), helix-loop-helix, Zn2/Cys6, glucocorticoid receptor-like (hormone receptors), C2H2 and C2HC zinc fingers and homeodomains (Table 1). We used the superfamily database (version 1.69) to obtain the numbers of TFs from each super-family in different organisms. The superfamily database contains extensive annotations of structural domains of proteins in 250 sequenced organisms using Hidden Markov Model profiles [64]. The database contains 1439 super-families. We focused on the major super-families of transcription factors studied in [15], and added all super-families that contained the terms "DNA binding" or "transcription". This resulted in 32 super-families. We further filtered super-families in which the maximal number of predicted proteins in a single organism was smaller than 50. For the remaining 10 super-families, we determined the maximum number of TFs as the maximal number of proteins from each super-family after discarding organisms with less than 5 proteins and discarding the top 1%

of the remaining organisms. It is important to note that the super-family domain assignment may contain predicted transcription factors due to the appearance of the relevant structural domain, which are in fact not functional, or which have other roles in the cell [65]. Thus, the maximal numbers presently found may be an overestimate.

The coding arguments presented here apply to the number of distinct DNA binding domains. Since the vast majority of known binding sites correspond to only one TF each in a given organism, one may assume that the TF family size is approximately equal to the number of DNA binding domains used by that family.

Binding sequence databases

Position-Specific Score Matrices (PSSM) for 46 *E. coli* transcription factors, were constructed based on the RegulonDB database [66]. The set of transcription factor binding sequences for each TF were searched for aligned motifs using AlignACE [67]. We chose the top-scoring motif, and considered only TFs with four or more aligned sequences contributing to that motif. Finally we removed the non-specific DNA binding factors FIS, HNS and IHF [21].

For the yeast *S. cerevisiae*, we used 94 PSSMs based on a set of 102 PSSMs constructed by Harbison et al [9]. We filtered out proteins that either do not bind DNA directly or always bind as a complex: Gal80, DIG1, STB1, Met4, HAP2, 3, 4, 5. All PSSMs were converted to a probability representation, where the sum of each PSSM column is 1.

For humans we used the PSSMs in the JASPAR database [68]. This data set consisted of 49 PSSMs.

Measurement of sequence similarity

To measure the similarity between binding sequences of a pair of factors we assessed the distances between their PSSMs. To compare pairs of PSSMs, we use a distance measure related to the one used by Wang et al [69]. The present measure, described below, is stringent in the sense that it scores bases according to their conservation within the PSSM, and compares to randomized PSSMs that preserve these conservation profiles. It is more appropriate for the present purpose than simpler methods such as direct comparison of sequence Hamming distances, because the present interest is in the active base pairs in the site, rather than base pair differences that have little functional impact on binding.

We denote the length of the binding sequence for TF_i as n_i, and its PSSM by M_i. The PSSM is an $4*n_i$ matrix in which each column, $p_{i,k}$ is a vector of length 4 holding the probability of observing letters A,C,G,T at position k in the set

of aligned binding sequences of TF$_i$. For each PSSM we created an information profile [7] denoted by I_i. The information profile is a n_i length vector whose k'th element quantifies the conservation of position k in the PSSM:

$$I_{i,k} = 2 - H(p_{i,k}) \quad (1)$$

where

$$H(p_{i,k}) = -\sum_{m=1}^{4} p_{i,km} \log_2(p_{i,km}) \quad (2)$$

is the entropy of $p_{i,k}$. $I_{i,k}$ has a minimum of 0 when all four bases have equal probability of appearing at position k, and a maximum of 2 when all aligned binding sequences have the same base at position k (the small sample size correction of [7] was applied). As the PSSMs in our database have different lengths, and the 'important' positions for the TF-target recognition are the conserved positions, we used the information profiles I as weight vectors when comparing two PSSMs.

We define the similarity between TF$_i$ and TF$_j$ as:

$$D_{ij} = \max_{A}\left(\sum_{k \in A} I_{i,k} \cdot I_{j,k} \cdot d_{ij,k}\right) \quad (3)$$

The maximum is taken over all relative shifts A of the two PSSMs, with a minimum of 5-positions of overlap. $d_{ij,k}$ is a similarity of the k'th position of the relatively shifted PSSMs. We used two different measures for d_{ij}: One minus the Shannon-Jensen distance [70]:

$$d_{ij,k} = 1 - [H(\frac{p_{i,k} + p_{j,k}}{2}) - \frac{H(p_{i,k}) + H(p_{j,k})}{2})] \quad (4)$$

and a simple correlation between the two probability vectors $p_{i,k}$ and $p_{j,k}$. Both measures gave very similar results. For each pair we computed the similarity with the reverse complement as well and took the maximal similarity. To detect pairs with significant similarity we first chose only pairs for:

$$D_{ij} > f^* \min(D_{ii}, D_{jj}) \quad (5)$$

Where f is a numerical factor (we used $f = 0.75$). This amounts to requiring that the similarity between the sets of binding sequences of two TFs comparable to the similarity between the sequences of each TF. For the pairs that passed this criterion, we created an ensemble of 1000 random PSSM pairs and computed a distribution of similarities D^r_{ij}. The random PSSMs were created by randomly exchanging the A-T and C-G positions in each column of the original PSSMs. This operation preserves the information profile, as well as the GC content, and therefore

forms a stringent ensemble. Similar sequences were sequences which had a p-value<0.005 for D_{ij} using both similarity measures.

Measurement of similarity of biological function of TFs

We defined two pairs of transcription factors as functionally similar if they jointly co-regulate a significant number of target genes. This information was obtained from transcription regulation networks: For yeast we used the network of [35,36]. For E. coli we used the network based on the data in the Ecocyc database [34]. It is important to note that these networks are based on direct experimental measurements, and not on putative interactions based on binding site predictions. For each pair of TFs we used a hyper-geometric test to assess whether the number of genes regulated by both TFs is significantly larger than expected from the fraction of target genes of each TF alone. This measure of functional similarity normalizes for the variable number of target genes for different TFs. We used a hyper-geometric test to detect enrichment of TF pairs with significant target co-regulation in the group of TF pairs with similar sequences.

A second measure of biological similarity was based on the similarity of the functional annotation of the gene targets of each factor. We used functional annotations[33,34] for the top tier of the annotation tree (except sub-cellular localization and general annotations such as "protein with binding function"). For yeast we used the following functional categories from the FunCat database [33]: metabolism, energy, cell cycle and DNA processing, cell rescue, defence and virulence, interaction with the cellular environment, interaction with the environment, transposable elements, viral and plasmid proteins, cell fate, development, biogenesis of cellular components, cell type differentiation.

For *E. coli* we used the following functional categories from the Ecocyc physiological roles annotation [34]: carbon utilization, degradation of macro-molecules, energy metabolism – carbon, energy production/transport, biosynthesis of building blocks, biosynthesis of macromolecules (cellular constituents), central intermediary metabolism, metabolism of other compounds, cell division, cell cycle physiology, motility, chemotaxis, energytaxis (aerotaxis, redoxtaxis etc), genetic exchange/recombination, adaptations, protection, defense/survival, DNA uptake. Each TF was assigned a profile vector in which each position holds the fraction of its target genes with the respective functional annotation. We used a Student t-test to compare the average of the Euclidean distances between profile vectors of TF pairs with significant sequence similarity to the average of the Euclidean distances between the profile vectors of all pairs. Lack of data

in humans currently prohibits a systematic measure of functional similarity of TFs.

Assessment of the number of possible sequences

We considered three features in the binding mechanism of each super-family which contribute to the number of possible sequences: the number of variable positions in each half-site, the relative spacing and orientation between them, and whether the sites are identical (for homo-dimeric TFs) or not (hetero-dimeric TFs). The number of possible sequences for a super-family with P variable positions in each half-site, O possible half-site orientations and S possible half-site spacings is $4^{P*H}*O*S/2$, where H = 1(2) if the super-family binds as homo-dimers (hetero-dimers) respectively. In our calculations we divide by 2 to account for reverse complementary sequences. The present study presents these estimates for six of the ten TF super-families (Table 1) for which data on these three features were available.

For the Lambda repressor-like super-family, previous work suggested 3 variable positions in each half site of the homo-dimer [21,22] and 2 relative orientations of the two monomers [71], resulting in $4^3*2/2 = 64$ possible sequnces. Helix-loop-helix proteins can bind as either homo-dimers or hetro-dimers, and have two variable positions at each half-site, resulting in an estimated number of possible sequences of $4^4/2 = 128$. Zn2/Cys6 proteins, such as the yeast GAL4 protein, bind two sequences of length three, with variable spacing (0–12 nucleotides) and three possible orientations (direct, inverted or everted repeats) [43], resulting in $4^3*3*13/2 = 1248$ possible sequences. The glucocorticoid receptor-like proteins have two variable positions at each half site [21], and can bind as hetero-dimers in three different orientations (direct, inverted or everted repeats) and variable spacing ranging from 0–8 [72], resulting in $4^4*3*9/2 = 3456$ possible sequences. Winged helix and homeo-domain recognize 6 positions as either homo-dimers or hetro-dimers respectively, with a constant relative orientation and half-site spacing, resulting in 4096/2 and $4096^2/2$ possible sequences, respectively. An estimation of the number of possible sequences for multi-domain C2H2 zinc fingers is not given, as these proteins can contain a variable number of finger domains (between 2 and more than 30) per protein.

Coding theory bounds – coloring number bound

We treated the mapping between transcription factors and binding sequences as a coding problem, where the code-words are short DNA sequences of a given length and the messages are TFs. The code-words are abstracted as nodes of a graph, where edges connect any two code-words which differ by one base-pair (Hamming distance 1). This scenario has some similarities with the genetic code,

where the code-words or "codons" are 3-base-pair strings, and the messages are the amino acids encoded by these codons. The error-load is the reduction in organism fitness due to erroneous binding of factor A to the code-word assigned to factor B. At one extreme, minimal error-load can be achieved by mapping all code-words to a single transcription factor. At the other extreme, maximal diversity is achieved by mapping each code-word to a different transcription factor, resulting in the same number of TFs as possible code-words.

If one assigns a biological function to the TFs that bind each code-word, it is clear that minimization of error-load would tend to smoothen this mapping. That is, TFs that are likely to bind similar sequences should have a similar biological function. For every mapping of binding sequences to factors, one can assign an error load, which measures the average impact of erroneous binding. It has recently been proposed [29] that in the limit of large errors, the maximal number of coded messages is bounded by the coloring number of the minimal surface which can embed the code-word graph [29]. Heawood's formula states that the coloring number is:

$$chr(\gamma) = \text{int}[\frac{1}{2}(7 + \sqrt{1 + 48\gamma})] \qquad (6)$$

where int [x] denotes the largest integer not greater than x. The coloring number depends on γ, the genus of the surface embedding the graph:

$$\gamma = 1 - \frac{1}{2}(V - E + F) = 1 - \frac{1}{2}V(1 - \frac{d}{4}) \qquad (7)$$

Here $V = q^n/2$ is the number of possible code words encoded by a q-letter string of length n, assuming a code-word which is the reverse complement of another is not available for independent assignment. $E = V*(d/2)$ is the number of graph edges, and $F = V*(d/4)$ is an estimate of the number of faces of the surface embedding the graph. $d = (q-1)n$ is the number of neighbors of each code-word (by identifying sequences and their reverse complements each code-word has twice as many neighboring code-words as a code without the reverse-complement constraint but about half of these code-words are not available for independent assignment). Using this we get the following estimate for the coloring number:

$$B(n,q) = chr(\gamma) \cong 3.5 + \sqrt{0.75q^n * [n(q-1) - 4)]} \qquad (8)$$

The bound for codes with different n are shown in Table 2.

<div style="display: flex;">
<div>

Coding theory bounds – sphere packing bound

An alternative possibility for the code mapping is one in which every sequence is assigned to only one transcription factor [6], and the probability of a misread error is thus negligible. The target DNA binding sequences of each transcription factor can be represented as a sphere in the code-word space (Fig 3). The center of each sphere is the consensus sequence, and all sequences differing from the consensus sequence by *e* positions (Hamming distance *e* from the consensus sequence) are assumed to be bound with a non-negligible probability by the TF [2,8,10]. Here we assumed *e = 1*.

Unlike the smooth code, the spheres here are non-overlapping. The volume of a "sphere" of radius *e*, which contains all code-words with Hamming distance of *e* or less from a given code-word, is [73]:

$$U(n,q,e) = \sum_{i=0}^{e} \binom{n}{i}(q-1)^i \qquad (9)$$

The number *N* of non-overlapping spheres of radius 1 is bounded by:

$$N \le \frac{1}{2} A_q(n,1) \qquad (10)$$

where:

$$A_q(n,1) = \left\lfloor \frac{q^n}{U(n,q,1)} \right\rfloor \qquad (11)$$

The upper bound in (11) is called the sphere-packing bound [27]. Codes which achieve this bound are called "perfect codes". In such codes the code-word space is fully covered by non-overlapping spheres of Hamming radius 1. Generally, the number of non-overlapping spheres is smaller, as some code-words remain uncovered by any sphere. The factor of one half in Eq. (10) stems from the fact that each sequence effectively represents also its reverse complementary sequence [27].

Authors' contributions

All authors participated in the design of the study and writing of the manuscript. SI performed all the computational aspects of the work. All authors read and approved the final manuscript.

Acknowledgements

</div>
<div>

References

1. Robison K, McGuire AM, Church GM: **A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome.** *J Mol Biol* 1998, **284**:241-254.
2. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16**:16-23.
3. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
4. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.** *Proc Natl Acad Sci U S A* 2000, **97**:10096-10100.
5. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss I, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Res* 2000, **28**:316-319.
6. Sengupta AM, Djordjevic M, Shraiman BI: **Specificity and robustness in transcription control networks.** *Proc Natl Acad Sci U S A* 2002, **99**:2072-2077.
7. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences.** *J Mol Biol* 1986, **188**:415-431.
8. Gerland U, Moroz JD, Hwa T: **Physical constraints and functional characteristics of transcription factor-DNA interaction.** *Proc Natl Acad Sci U S A* 2002, **99**:12015-12020.
9. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**:99-104.
10. Berg J, Willmann S, Lassig M: **Adaptive evolution of transcription factor binding sites.** *BMC Evol Biol* 2004, **4**:42.
11. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18**:6097-6100.
12. Garvie CW, Wolberger C: **Recognition of specific DNA sequences.** *Mol Cell* 2001, **8**:937-946.
13. Luscombe NM, Austin SE, Berman HM, Thornton JM: **An overview of the structures of protein-DNA complexes.** *Genome Biol* 2000, **1**:REVIEWS001.
14. Pabo CO, Sauer RT: **Transcription factors: structural families and principles of DNA recognition.** *Annu Rev Biochem* 1992, **61**:1053-1095.
15. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks.** *Curr Opin Struct Biol* 2004, **14**:283-291.
16. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM: **The many faces of the helix-turn-helix domain: transcription regulation and beyond.** *FEMS Microbiol Rev* 2005, **29**:231-262.
17. Mandel-Gutfreund Y, Margalit H: **Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites.** *Nucleic Acids Res* 1998, **26**:2306-2312.
18. Benos PV, Lapedes AS, Stormo GD: **Probabilistic code for DNA recognition by proteins of the EGR family.** *J Mol Biol* 2002, **323**:701-727.
19. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424**:147-151.
20. van Nimwegen E: **Scaling laws in the functional content of genomes.** *Trends Genet* 2003, **19**:479-484.
21. Luscombe NM, Thornton JM: **Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity.** *J Mol Biol* 2002, **320**:991-1009.
22. Ptashne M: **A genetic switch.** , cell press&blackwell science; 1992.
23. Massari ME, Murre C: **Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms.** *Mol Cell Biol* 2000, **20**:429-440.
24. Beltran AC, Dawson PE, Gottesfeld JM: **Role of DNA sequence in the binding specificity of synthetic basic-helix-loop-helix domains.** *Chembiochem* 2005, **6**:104-113.
25. Jones S: **An overview of the basic helix-loop-helix proteins.** *Genome Biol* 2004, **5**:226.
26. Kaplan T, Friedman N, Margalit H: **Ab Initio Prediction of Transcription Factor Targets Using Structural Knowledge.** *PLoS Computational Biology* 2005, **1**:e1.
27. Marathe A, Condon AE, Corn RM: **On combinatorial DNA word design.** *J Comput Biol* 2001, **8**:201-219.

</div>
</div>

28. Hamming RW: **Coding and information theory.** New Jersey, Prentice-Hall; 1980.
29. Tlusty T: **Emergence of a genetic code as a phase transition induced by error-load topology.** . submitted.
30. Woese CR: **Order in the genetic code.** *Proc Natl Acad Sci U S A* 1965, **54:**71-75.
31. Swanson R: **A unifying concept for the amino acid code.** *Bull Math Biol* 1984, **46:**187-203.
32. Haig D, Hurst LD: **A quantitative measure of error minimization in the genetic code.** *J Mol Evol* 1991, **33:**412-417.
33. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW: **The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes.** *Nucleic Acids Res* 2004, **32:**5539-5545.
34. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli.** *Nucleic Acids Res* 2005, **33:**D334-7.
35. Garten Y, Kaplan S, Pilpel Y: **Extraction of transcription regulatory signals from genome-wide DNA-protein interaction data.** *Nucleic Acids Res* 2005, **33:**605-615.
36. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
37. Tanaka N, Kawakami T, Taniguchi T: **Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system.** *Mol Cell Biol* 1993, **13:**4531-4538.
38. Ray-Gallet D, Mao C, Tavitian A, Moreau-Gachelin F: **DNA binding specificities of Spi-1/PU.1 and Spi-B transcription factors and identification of a Spi-1/Spi-B binding site in the c-fes/c-fps promoter.** *Oncogene* 1995, **11:**303-313.
39. Babu MM, Teichmann SA: **Evolution of Transcription Factors and the Gene Regulatory Network in E. coli.** *Nucleic Acids Research* 2003, **31:**1234-1244.
40. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication.** *Nat Genet* 2004, **36:**492-496.
41. Papp B, Pal C, Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast.** *Trends Genet* 2003, **19:**417-422.
42. Tanay A, Gat-Viks I, Shamir R: **A global view of the selection forces in the evolution of yeast cis-regulation.** *Genome Res* 2004, **14:**829-834.
43. Todd RB, Andrianopoulos A: **Evolution of a fungal regulatory gene family: the Zn(II)2Cys6 binuclear cluster DNA binding motif.** *Fungal Genet Biol* 1997, **21:**388-405.
44. Volff JN: **Genome evolution and biodiversity in teleost fish.** *Heredity* 2005, **94:**280-294.
45. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH: **Zebrafish hox clusters and vertebrate genome evolution.** *Science* 1998, **282:**1711-1714.
46. Meyer A, Schartl M: **Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions.** *Curr Opin Cell Biol* 1999, **11:**699-704.
47. Seoighe C: **Turning the clock back on ancient genome duplication.** *Curr Opin Genet Dev* 2003, **13:**636-643.
48. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20:**1377-1419.
49. Yuh CH, Bolouri H, Davidson EH: **Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.** *Science* 1998, **279:**1896-1902.
50. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153-159.
51. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:**185-198.
52. Buchler NE, Gerland U, Hwa T: **On schemes of combinatorial transcription logic.** *Proc Natl Acad Sci U S A* 2003, **100:**5136-5141.
53. Setty Y, Mayo AE, Surette MG, Alon U: **Detailed map of a cis-regulatory input function.** *Proc Natl Acad Sci U S A* 2003, **100:**7702-7707.
54. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34:**166-176.
55. Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L: **Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast.** *J Mol Biol* 2006, **360:**213-227.
56. Muller J, Oehler S, Muller-Hill B: **Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator.** *J Mol Biol* 1996, **257:**21-29.
57. Kolb A, Busby S, Buc H, Garges S, Adhya S: **Transcriptional regulation by cAMP and its receptor protein.** *Annu Rev Biochem* 1993, **62:**749-795.
58. Ptashne M, Gann A: **Transcriptional activation by recruitment.** *Nature* 1997, **386:**569-577.
59. Zhang Z, Gu J, Gu X: **How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution?** *Trends Genet* 2004, **20:**403-407.
60. Struhl K: **Yeast transcriptional regulatory mechanisms.** *Annu Rev Genet* 1995, **29:**651-674.
61. Kafri R, Bar-Even A, Pilpel Y: **Transcription control reprogramming in genetic backup circuits.** *Nat Genet* 2005, **37:**295-299.
62. Dueber JE, Yeh BJ, Bhattacharyya RP, Lim WA: **Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry.** *Curr Opin Struct Biol* 2004, **14:**690-699.
63. Zarrinpar A, Bhattacharyya RP, Lim WA: **The structure and function of proline recognition domains.** *Sci STKE* 2003, **2003:**RE8.
64. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:**903-919.
65. Kummerfeld SK, Teichmann SA: **DBD: a transcription factor prediction database.** *Nucleic Acids Res* 2006, **34:**D74-81.
66. Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12.** *Nucleic Acids Res* 2004, **32:**D303-6.
67. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.** *J Mol Biol* 2000, **296:**1205-1214.
68. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-4.
69. Wang T, Stormo GD: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19:**2369-2380.
70. Lin J: **Divergence measures based on Shannon entropy.** *IEEE Trans Inform Theory* 1991, **37:**145-151.
71. Kolkhof P, Teichmann D, Kisters-Woike B, von Wilcken-Bergmann B, Muller-Hill B: **Lac repressor with the helix-turn-helix motif of lambda cro binds to lac operator.** *Embo J* 1992, **11:**3031-3038.
72. Sandelin A, Wasserman WW: **Prediction of nuclear hormone receptor response elements.** *Mol Endocrinol* 2005, **19:**595-606.
73. Peterson WW, Weldon EJ: **Error-Correcting Codes.** , The MIT Press; 1972.