Evolutionary Potential of Cis-Regulatory Mutations to Cause Rapid Changes in Transcription Factor Binding

Jasmin D. Kurafeiski, Paulo Pinto, and Erich Bornberg-Bauer*

Molecular Evolution and Bioinformatics, University of Muenster, Germany

*Corresponding author: E-mail: ebb@uni-muenster.de

Accepted: December 11, 2018

Abstract

Transcriptional regulation is crucial for all biological processes and well investigated at the molecular level for a wide range of organisms. However, it is quite unclear how innovations, such as the activity of a novel regulatory element, evolve. In the case of transcription factor (TF) binding, both a novel TF and a novel-binding site would need to evolve concertedly. Since promiscuous functions have recently been identified as important intermediate steps in creating novel specific functions in many areas such as enzyme evolution and protein–protein interactions, we ask here how promiscuous binding of TFs to TF-binding sites (TFBSs) affects the robustness and evolvability of this tightly regulated system. Specifically, we investigate the binding behavior of several hundred TFs from different species at unprecedented breadth. Our results illustrate multiple aspects of TF-binding interactions, ranging from correlations between the strength of the interaction bond and specificity, to preferences regarding TFBS nucleotide composition in relation to both domains and binding specificity. We identified a subset of high A/T binding motifs. Motifs in this subset had many functionally neutral one-error mutants, and were bound by multiple different binding domains. Our results indicate that, especially for some TF–TFBS associations, low binding specificity confers high degrees of evolvability, that is that few mutations facilitate rapid changes in transcriptional regulation, in particular for large and old TF families. In this study we identify binding motifs exhibiting behavior indicating high evolutionary potential for innovations in transcriptional regulation.

Key words: transcriptional regulation, evolution of regulatory networks, binding specificity, neutral networks, multifunctionality.

Introduction

Transcriptional Regulation

Over billions of years, natural selection has generated an amazing diversity of phenotypes performing an impressive variety of biological functions (Carroll 2005; Davidson and Erwin 2006). Genes and the traits they confer are highly entangled through (co)regulation and complex interaction networks, such as protein–protein and protein–DNA (Barabasi and Oltvai 2004). Their combinatorial complexity and nonlinear dynamic behavior helps explain the diversity of life, given that differences in gene number between species lack such explanatory power. A crucial role in increasing biodiversity is commonly ascribed to differential transcriptional regulation (Berg et al. 2004; Mustonen and Lässig 2005; Davidson and Erwin 2006; Elena et al. 2011; Payne and Wagner 2014). Consequently, transcription factors (TFs) play a major role in shaping biological processes by regulating gene expression and genetic interactions (Wray 2007) via binding to TF-binding sites (TFBSs). TFs have specific domains binding to their preferred target TFBS at the target DNA. Just like most other protein domains, the domains of TFs are functional and evolutionarily well-conserved units which can be rearranged over long evolutionary time scales (Moore et al. 2008; Schmitz et al. 2016). TFs mainly evolve by gene duplication and domain rearrangement which confer new functional potentials to TFs (Amoutzias 2004; Schmitz et al. 2016). Known TF DNA-binding domains include the homeobox (Schofield 1987; Gehring 1992, 1993), ETS (Graves and Petersen 1998; Oikawa and Yamada 2003; Gutierrez-Hartmann et al. 2007), zinc finger (Laity et al. 2001; Krishna et al. 2003; Gamsjaeger et al. 2007), fork head (Kaufmann and Knochel 1996), and basic helix-loop-helix (bHLH) (Murre et al. 1994; Chaudhary and Skinner 1999). Some TFs engage in combinatorial binding, forming dimers or multimers (Verger and Duterque-Coquillaud 2002; Amoutzias et al. 2004; Veron et al. 2007). This enables further fine-tuning of transcriptional

regulation (Stampfel et al. 2015; Reiter et al. 2017) and affects low-affinity interactions. One example for low-affinity binding through combinatorial binding is interactions between homotypic-binding site clusters in *Drosophila* (Crocker et al. 2015).

The length of TFBS ranges from 5 to 30 nt, although most common is a binding region of 10 nt (Berg et al. 2004; Spitz and Furlong 2012; Stewart et al. 2012). Due to the relative shortness of the TFBSs it is likely that binding between a TF and its TFBS is not solely determined by sequence compatibility but also by the aforementioned combinatorial binding (Spitz and Furlong 2012), as well as the regions flanking the TFBS (Slattery et al. 2011, 2014; Song et al. 2011; Dror et al. 2015) which influence the site's accessibility to TFs (Song et al. 2011; Dror et al. 2015). G/T rich regulatory regions are associated with clusters of TFBSs for a variety of TFs, and A/T-rich regions with more specific TFBSs (Lenhard et al. 2012). Changes in TFBSs are the major cause for changes in gene regulatory networks for several reasons: first, since binding motifs are short and consequently more likely to arise through random mutations, new components of gene regulatory networks may arise (Arnone and Davidson 1997; Stone et al. 2001). Second, changes in TFBSs which regulate TFs are known to also change the functional specificity of TFs, for example, following their duplication (Schmitz et al. 2016).

Nucleotide substitutions may either weaken or strengthen binding between the TF and the TFBS. Despite components in a regulatory network depending on each other, cases of new functions arising due to substitutions in a TFBS are known (Gadau et al. 2012; Simola et al. 2013). Consequently, a balance between conservation of function and innovation is necessary for evolutionary changes to occur.

The evolutionary trajectories and possible fitness consequences of successive point mutations have been extensively researched in other systems, such as the sequence to structure maps in RNA secondary structures (Schuster et al. 1994) and in model proteins (Bornberg-Bauer 1997).

Sets of functional sequences in genotype space forming the same phenotype are referred to as a neutral set (Bornberg-Bauer 1997). If the sequences are also connected through single mutations, they are usually referred to as a neutral network (supplementary fig. 1, Supplementary Material online) and can be considered as a subset of genotype space (Wright 1932; Wagner 2008). Here, genotype space includes the set of all possible motifs consisting of 8 nt, in other words 65,536 sequences representing potential TFBSs in our data set. After removal of redundant motifs, our genotype space contains 32,896 sequences. Among the motifs forming genotype space the ones bound by a TF form the neutral set of said TF. All motifs in the neutral set that are connected through single mutations form the neutral network of the TF. A large neutral network usually indicates high mutational robustness of a function. The importance of promiscuity and mutational robustness in exploring sequence space and facilitating the emergence of new functions has been postulated and observed in several systems over the last years (Sikosek 2012; Sikosek and Chan 2014). Within the framework of neutral networks, it follows that mutants which are further away from the mutationally most stable and most uniquely binding (or folding) consensus sequences are less stable and functionally promiscuous. However, their promiscuity also sometimes links the native neutral network with another network (in which the latent function becomes the dominant function) such that these sequences can act as "evolutionary bridges" (Bornberg-Bauer 1997; Anderson et al. 2005; Wroe et al. 2007).

In the context of TFBS evolution, drift along a "native" neutral network can thus lead to transitions which are facilitated by latent traits (weak binding to another TF). However, the strengthening of such latent traits can be constrained by the need to maintain the native function (Depristo 2007). A proposed solution to this issue is the "escape from adaptive conflict" (EAC) hypothesis (Hittinger and Carroll 2007; Des Marais and Rausher 2008; Sikosek et al. 2012). In the context of protein structures, the EAC model describes multifunctional proteins as fluctuating between different structures, where the ratio with which the structures occur is determined by their thermostability. As a result, both the native function and latent function are subject to selection (Sikosek et al. 2012). Those constraints can be loosened through either the tradeoff between native and latent function or gene duplication followed by subfunctionalization (Aharoni et al. 2005; Sikosek et al. 2012).

Here, instead of fluctuations between structures, we consider promiscuity of TFBSs bound by TFs. We propose promiscuity and EAC to be important for fine tuning and expanding regulatory networks (supplementary fig. 2, Supplementary Material online).

Accordingly, we study the binding specificity and the neutral networks of different TFs to identify differences in their binding preferences. Specifically, we find that motifs, such as A/T-rich motifs, which can bind multiple TF families, exist and may act as evolutionary bridges for exploring new functions.

Methods and Methods

The UniProbe Database

Uniprobe is a database storing data on TF binding that was generated using universal protein-binding microarray technology. Binding motifs are represented by 32,896 possible sequences consisting of 8 nt.

The data were generated in multiple experiments across a wide range of species, most notably *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae*. The database contains data on multiple species and more than 500 TFs. Binding strength between the TF and motifs is described by the so called enrichment score (E-score). For each TF the E-scores for all of the 32.896 contiguous motifs were checked to

determine whether the TF bound to the motif or not. The E-score is a common measure for binding strength and was calculated based on the binding affinity between a TF and the motif. Though the E-score does not directly measure binding strength it has been found to accurately correlate with binding (Badis et al. 2009; Nakagawa et al. 2013; Aguilar-Rodriguez et al. 2017). Current data are accessible in the UniProbe public database (Newburger and Bulyk 2009).

Processing and Visualization of the Binding Data

In order to investigate binding behavior we used Python scripts to process the available data. We generally set the threshold for functional binding at 0.35 (Badis et al. 2009; Nakagawa et al. 2013; Aguilar-Rodriguez et al. 2017); however for some analyses, we also investigated behavior for lower thresholds. It also has to be noted that there are known cases of low-affinity interactions of biological significance (Farley et al. 2015). Furthermore, we also investigated the role of the A/T-content of the binding site in promiscuous-binding of TFs. For the visualization of the results we used Matlab, Matplotlib (Hunter 2007), and R (R_core).

Set Distance

Several measures can be implemented for the numerical assessment of the similarity of sets. One of those measures of similarity is the Jaccard Index. We present results relative to the Jaccard index (J) due to its easy interpretability, but we should highlight that similar results are obtained when employing other measure. The Jaccard index (J) is defined as the fraction of the intersection to the union of the two sets. For sets A and B, the Jaccard index is defined as $J(A, B) = |A \cap B|/|A \cup B|$. Additionally, for two sets of motifs A and B the Jaccard distance is defined as $dJ(A, B) = 1 - J(A, B)$. The Jaccard distance is particularly useful when the sets to compare are of different sizes, as its normalization is designed to take the union of both sets.

In order to check for a correlation between the sets of motifs bound and sequence similarity the set distance of two TF was then compared with their BLAST similarity (Altschul et al. 1990). The sequences of two TF ware considered to be similar if the E-value was $<1e^{-4}$.

Creating Networks

For the creation of networks of the bound TFBSs we utilized networkx (Hagberg et al. 2008) with each motif bound by the TF being represented by a node. Edges between nodes indicate a mutational distance of 1 between them.

Results and Discussion

Binding Preferences of TFs and TFBSs

First, we characterized the binding preferences of all TFs and motifs exhibiting promiscuous binding behavior. The data set includes a variety of TFs utilizing different domain types from several species. The most prevalent species in the set are *M. musculus*, *H. sapiens*, and *S. cerevisiae* (for further details see Materials and Methods). We use the common threshold of E-score > 0.35 to identify binding interactions that are considered functionally relevant (Badis et al. 2009; Nakagawa et al. 2013; Aguilar-Rodriguez et al. 2017). To assess the general validity of the micro-array data, we compared the overall patterns of binding preferences to the existing literature.

Regulatory regions with a high G/C content are associated with multiple transcription initiation sites and a broad number of TFs; A/T-rich regions generally mark a specific transcription start site and often contain a so called TATA-box (Lenhard et al. 2012). We initially observed unspecific binding in motifs that are A/T rich (fig. 1, data labeled in black). However, the observations could be skewed due to the abundance of data on homeobox TFs and their preference to bind to A/T-rich TATA-box motifs. After separately analyzing the data according to the binding domains, the trend is no longer supported (fig. 1). Instead, binding preferences seem to mainly be determined by the preferred binding motif of the TF. The specificity of interaction has also been found to be influenced by the domain type and nucleotide composition of the motifs. Promiscuous binding has been observed between A/T-rich TFBSs and the homeobox TFs. Homeoboxes require an "ATTA" or similar motif to induce binding interaction (Gehring 1992; fig. 1; supplementary table 1, Supplementary Material online). Interestingly, the general trend of promiscuity in A/T-rich motifs is present in all analyzed binding domains at low binding strength. It decreases with increasing the threshold for binding strength for all analyzed domain types except the homeoboxes (supplementary fig. 3, Supplementary Material online). In summary, the observed low-affinity interactions are more promiscuous compared with interactions of higher affinity.

At high binding affinity the preferences of homeobox TFs are in agreement with the available literature (Otting et al. 1990; Billeter et al. 1993; Gehring, 1993). This generally holds up for all the binding domains. Zinc fingers of the C_2H_2 domain type contain a conserved alpha helix, and ligands of cysteine and histidine (Laity et al. 2001; Gamsjaeger et al. 2007). The bHLH domain binds "CANNTG" as a consensus motif (Murre et al. 1994; Chaudhary and Skinner 1999) making it unlikely to bind motifs with an A/T-content of zero, one, seven, or eight (nAT[0, 1, 7, 8]). Fork heads bind to a canonical motif of "G/A T/C C/A A A C/T A" (Kaufmann and Knochel 1996), resulting in them binding motifs across the whole spectrum of motif nucleotide composition. ETS TFs bind a core motif of "5'-GGAA/T-3'" (Oikawa and Yamada 2003;

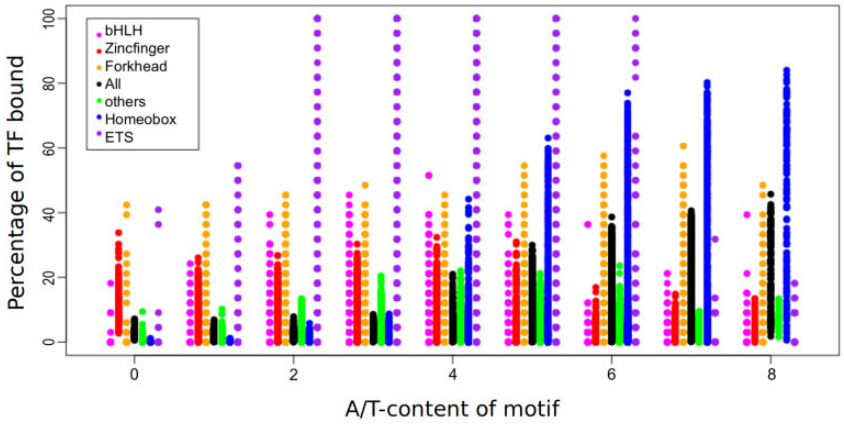

Fig. 1.—Binding preferences of TFs in relation to the A/T-content of TFBS for an E-score > 0.35. Data points represent an 8-nt long motif. The colors indicate which TF-binding domain the motif interacts with. Consequently all circles of the same color form the set of a domain. Each data point shows a motif and the percentage of TF with the same domain binding the motif.

Gutierrez-Hartmann et al. 2007; Wei et al. 2010) (supplementary table 1, Supplementary Material online).

When analyzing the mean and median numbers of bound TFs (supplementary fig. 4, Supplementary Material online) it seems to be in agreement with A/T-rich motifs binding indiscriminate as described in the literature (Lenhard et al. 2012). However, the large error bars show the amount of variance when it comes to the number of bound TFs (supplementary fig. 5, Supplementary Material online). Despite the clear trend of A/T-rich motifs binding promiscuous it is possibly exaggerated through motifs that are extremely unspecific and again the number of homeobox TFs in the data set.

Binding interactions are also more numerous at low thresholds of binding strength (supplementary fig. 4, Supplementary Material online). This indicates that weak binding events might be numerous but can also be broken up by a TF forming a stronger bond with the site. Some of the low-affinity interactions might still have a biological function when considering combinatorial binding interactions (Crocker et al. 2015). Indeed, several TFs forming homo- or heteromers occupying multiple binding motifs that are located close to each other can still lead to stable interactions despite low-affinity interactions being involved. Here, analyzed chip-seq data do, however, not allow for statements regarding combinatorial binding. When raising thresholds of binding affinity the number of motifs binding many TFs quickly decreases, indicating a tendency for the motifs being rather specific in the sets of TFs they bind to.

Comparison of Sequence Similarity and Binding Preferences

To check if similar TFs bind similar sets of binding motifs we compared sequence similarity to the sets of bound motifs. In order to compare the sets of bound motifs we utilized the Jaccard Distance (see Materials and Methods) as a simple

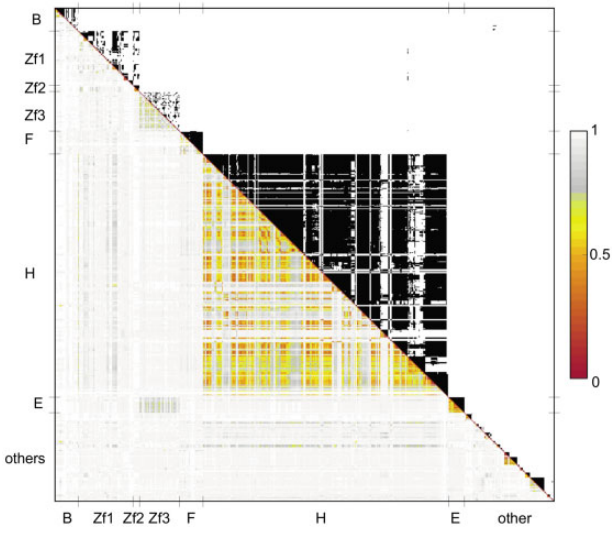

Fig. 2.—Comparison of set similarity (lower part) and sequence similarity (upper part). Set similarity is based on Jaccard distance. A distance of zero (red) equals 100% overlap of bound TFBS between two TFs. Orange indicates a distance of 0.5 (50% overlap), and yellow indicates a distance around 0.75 (25% overlap). Sequence similarity was determined using BLASTp, with all pairs of sequences with an e-value < 1 e-4 being considered as similar. Legend of domain types: B, bHLH; Zf1, C2H2 zinc finger; Zf2, zinc-cluster; Zf3, Zn2Cys6 zincfinger; F, fork head; H, Homeobox; E, ETS).

measure of set similarity, ranging from 0 (same) to 1 (no overlap). Sequence similarity is defined by a BLAST E-value $<1e^{-4}$ (Altschul et al. 1990). In general, sequences with high sequence similarity have >25% intersection between their neutral sets (fig. 2). However, there are also several exceptions, indicating that similar sequences do not always equate to similar binding preferences. A noticeable example is the zf-C_2H_2 type zinc fingers. Many pairs of this TF family share very few binding motifs despite having similar sequences. A possible cause for the differences could be that zfC_2H_2 TFs

tend to form dimers or multimers in order to bind to their targets (Yanez-Cuna et al. 2012; Schulz et al. 2015; Reiter et al. 2017), making 8 nt binding motifs insufficient to initiate binding. Furthermore, their threshold for biologically functional affinity might be lower due to combinatorial binding resulting in low-affinity functional interactions (Crocker et al. 2015). Various TF families besides zincfingers are known to form dimers and polymers in order to bind their targets. Examples are TFs utilizing the ETS domain (Verger and Duterque-Coquillaud 2002), and fork head TFs (De Val et al. 2008). Unlike the pairs of zfC_2H_2 pairs of ETS TFs were able to bind similar sets of motifs. Also, despite only short binding motifs being available and therefore combinatorial binding being impossible, we did not observe failure to bind. In other cases some pairs of TFs bind similar sets of motifs despite their sequences being dissimilar. For example, the ETS TFs share around 25% of their binding motifs with Zn_2Cys_6 TFs. However, the overlap is higher than commonly observed for dissimilar TFs, but still low compared with the overlap between TFs with similar sequences.

Networks of Bound TFBS

Each TF binds an individual set of motifs and these interactions can be characterized as a network, with nodes representing the sites bound by the TF and edges connecting two nodes indicating a nucleotide substitution. All motifs connected through a series of single mutations can be considered part of the neutral network of a TF. We then analyzed properties which are widely used to characterize the topological properties of a network in general (Seidman 1983; Albert and Barabasi 2002; Bettstetter 2002; Newman 2003). Specifically, we analyzed: 1) network density, which is defined as the fraction of possible connections between nodes that are observed in the network and corresponds to how densely the nodes are connected to their neighboring nodes; 2) the node degree, which describes the number of edges for each node in the network. Nodes with a high network degree have many neighbors with a mutational distance of 1; 3) network size, defined by the number of nodes in a network, the higher the number of nodes in the network the higher the promiscuity of the TF. Accordingly, the networks of promiscuous TFs are large and possess higher node connectivity but lower network density (fig. 3*A* and *B*). Nodes with high connectivity can be considered as "hubs" in their networks. Here, we consider nodes with a connectivity of 12 or higher in at least one network as hubs, because they have $\geq$50% of their possible neighbors in the network. The majority of the "hub" motifs is very similar and has an A/T-content of seven or eight out of 8 nt. Due to this high similarity the majority of the "hub" motifs form a densely connected neutral network (supplementary fig. 6, Supplementary Material online). The density of the network further illustrates how close the motifs are clustered in sequence space.

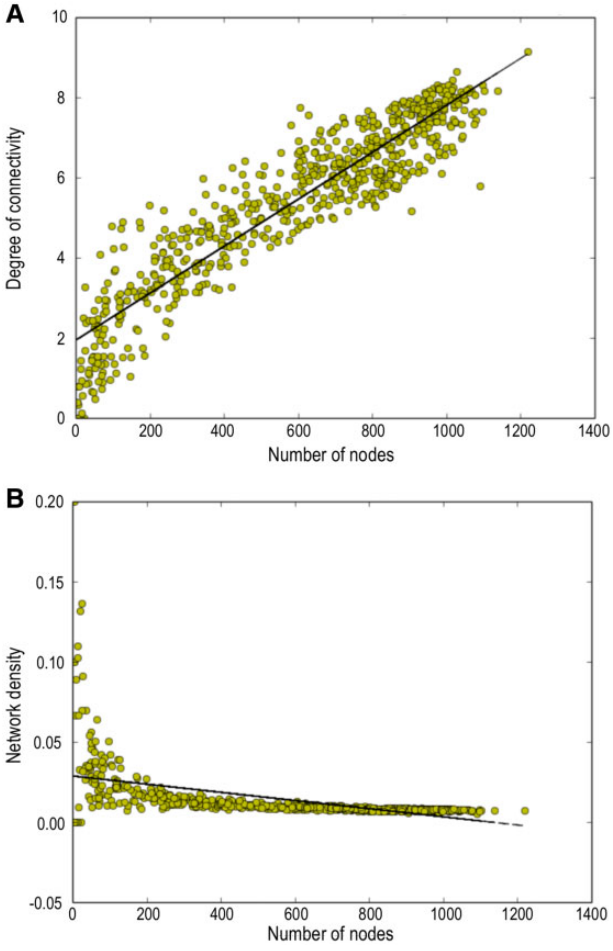

Fig. 3.—(*A*) Scatterplot illustrating the correlation of node degree in a network with network size. Node degree is the number of edges from a node to other nodes. Here, a high node degree indicates the TFBS has many mutational neighbors in the set of bound sites. The figure shows the average node degree for all 701 networks in connection to the number of nodes in the network. We find high node numbers to be positively correlating with high node degrees. (*B*) Scatterplot illustrating the correlation of network density in a network with network size. We observe a tendency for bigger networks to be less densely connected.

Binding Motifs Shared by TFs Using Different Binding Domains

Binding competition and complex binding interactions are a major factor in determining the development of an organism. We compared the sets of motifs bound by multiple binding domains to identify possible combinations of TFs involved in combinatorial or competitive regulation. We considered all possible combinations of three TFs with different binding domains and picked groups sharing at least 5% of motifs. All TFs fulfilling this condition were considered as candidates for further analysis. The candidate TFs and their connections with each other can be seen in supplementary figure 7, Supplementary Material online. We found at least one pair of TFs sharing TFBSs that are known to activate and repress

the translation of the same gene. The TF Yox1p and Fkh2p activate and repress Mcm1p, respectively (Darieva et al. 2010). Yox1p and Fkh2p bind to completely different regions of Mcm1p; therefore, they do not seem to be directly competing for the same motifs despite the intersection of motifs both TFs can interact with. Further known protein–protein interactions are shown in supplementary figure 8, Supplementary Material online; however, since our data rely on suboptimal binding interactions, possible interactions might not have been observed in vivo.

The motifs from the intersections between the sets of the candidate TFs were also further analyzed as they are also candidates for being involved in complex regulatory interactions. Again, comparable to the "hubs", we investigate the nucleotide composition of the motifs. Contrasting the Gaussian distribution of available motifs that peaks at $n(A/T) = 4$, we observe two peaks in numbers of binding sites for motifs with $n(A/T) = 3$ and $n(A/T) = 7$. Despite only a small fraction of available motifs being part of the set binding multiple TFs binding domains the vast majority of motifs belong to a single connected network (supplementary fig. 9, Supplementary Material online). In the network, we can see two subgroups of nodes that correspond to the nucleotide compositions of the motifs. Generally, the motifs in this set are mutationally dissimilar to the canonical TFBSs as well as the highest scoring motifs. However, the majority is still mutationally accessible through point mutations due to high network degree of connectivity (supplementary fig. 10*A* and *B*, Supplementary Material online)

Motifs Acting as Network Hubs as Well as Binding Various Binding Domains

Motifs bound by multiple TFs binding domains and having many functionally neutral neighbors binding the same TF can play an important role in the evolution of regulatory networks, due to their evolutionary potential. In order to identify motifs with high evolutionary potential we considered motifs that we identified as network hubs ("hubs") and the motifs we identified as being able to bind a variety of TF-binding domains ("multidomain"). Because both of the subsets exhibit an abundance of A/T-rich motifs we checked for an overlap between the two sets. Around 23% of motifs from the set "hubs" and 50% of the set "multidomain" are present in both sets, resulting in a third set "shared" (645 motifs, list of motifs forming the sets available in the supplementary material 1, Supplementary Material online). In the network of "shared" motifs the majority of nodes do not form a single connected network, but three smaller connected networks (supplementary fig. 11, Supplementary Material online). Out of these three networks, the largest contains the motifs that center around a nucleotide content of $A/T = 7$. The other two subnetworks contain nodes representing motifs around $A/T = 3$. We compared the distance of the motifs to the

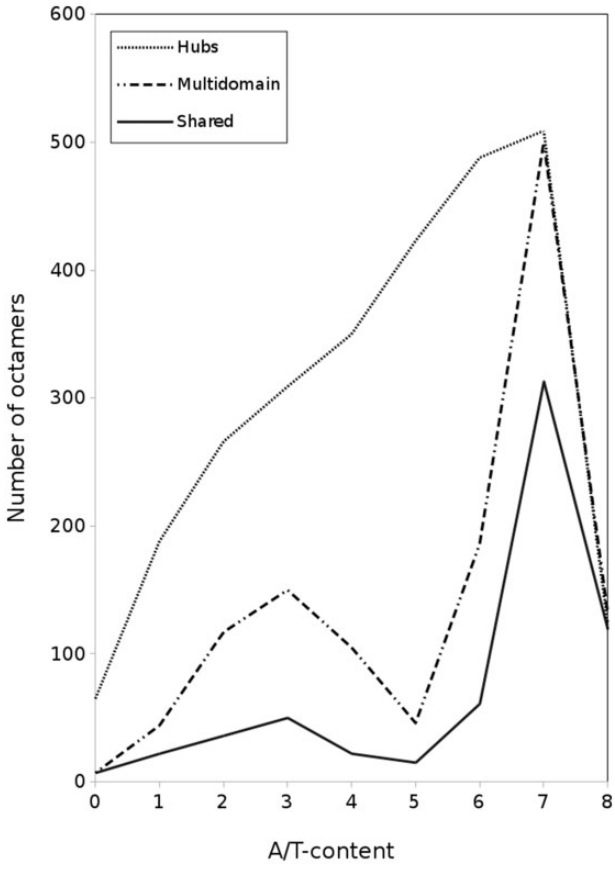

Fig. 4.—Distribution of TFBS by A/T-content. The figure shows the numbers of sites bound by TF utilizing varying binding domains (Competitive binding), sites that have many neighbors in at least one network of binding sites (Degree ≥ 12), and the ones fulfilling both criteria (Shared). All distributions differ from the distribution of available TFBSs. We observe a noticeable shift towards TFBSs with a high A/T-content. The TFBSs bound by various domains also have an unexpected peak number for an A/T-content of 3.

canonical TFBSs and highest scoring octamers for each of the involved TFs. Unexpectedly, the majority of motifs in the subset "shared" are neither similar to the canonical TFBSs nor the highest scoring octamers (supplementary fig. 10*C* and *D*, Supplementary Material online). However, the resulting networks of subsets are still mostly connected. In most cases the motifs are accessible from the canonical or highest scoring motif via point mutations (supplementary fig. 10*C* and *D*, Supplementary Material online).

The number of available TFs is limited, creating competition for TF binding amongst TFBSs (Karreth et al. 2014). This is especially important in eukaryotes as their TFs can be prone to binding wrong TFBSs (Wunderlich and Mirny 2009). Though for the complete set of binding motifs we could not detect a general trend regarding A/T-content and promiscuity we still observed a higher than expected rate of A/T-richness in the motifs in the subsets labeled as "hub" and "multidomain." Common TF domain types involved in those

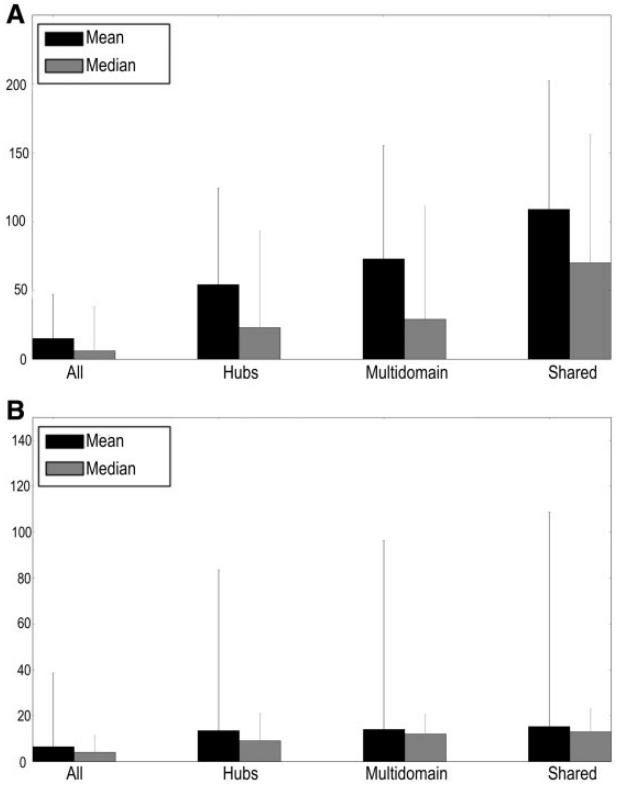

Fig. 5.—Comparison of binding specificity of different subsets of motifs. The figure shows the mean and median numbers of TFs binding to different subsets of motifs. The compared subsets are: 1) all available motifs involved in binding interactions, 2) motifs with high network connectivity (hubs), 3) motifs interacting with various binding domains (multidomain) and 4) the overlap between the sets hub and multidomain. All subsets show higher promiscuity compared with the set of all motifs involved in binding interactions. The mean in all sets is higher than the median, indicating an abundance of promiscuous TFBSs. (*A*) Shows interactions between binding motifs and all TFs available in the data set. (*B*) Shows the same data with TFs utilizing a homeobox binding domain removed to control for their highly promiscuous binding behavior. The differences of the subsets (hubs, multidomain, shared) to the main set are significant (respectively: *A*, 0.00, 0.00, 8.878571 e-272; *B*, 4.38480 e-297, 4.06469799 e-294, 4.236987 e-175).

interactions are zincfingers, forkheads, and homeoboxes. The nucleotide preferences of the three TF-binding domain types also seem to determine the peaks we observed for motifs of A/T = 3 and A/T = 7 (fig. 4).

Binding Specificity of Motifs with Possible Evolutionary Potential

We identified motifs that might play a role in regulatory network evolution. Considering the possible connection between multifunctionality and evolvability we compared binding interactions of the whole data set with the subsets "hubs", "multidomain", and "shared."

When calculating the mean and median percentage of the number of TFs bound per motif in the different subsets we observe that the motifs present in all three subsets tend to bind more promiscuously. In all sets the binding specificity varies strongly between motifs (fig. 5), leading to a high standard deviation. We tested the significance between the subsets and the main set using a Wilcoxon–Mann–Whitney test. Despite the high standard deviations the motifs in the three subsets exhibit a significantly higher level of promiscuity. Furthermore, the mean of all subsets is higher than the median, indicating an abundance of unspecific motifs (fig. 5). To ensure the pattern is universal in the data set and not caused by the abundant and promiscuously binding homeobox TF we repeated the analysis excluding all homeobox TFs. The previously observed trend of the motifs in the subsets acting more promiscuously is less pronounced without the homeoboxes, but we were still able to detect significantly higher levels of unspecific binding (fig. 5). The specificity of the investigated motifs regarding number of bound TFs as well as variety of binding domains further strengthen the possibility of them playing an important role in network evolution.

Conclusion

Here, we identified binding motifs that possess high evolutionary potential. We followed the hypothesis that motifs with many functionally neutral neighbors and binding multiple domains possess a high evolutionary potential. This hypothesis is based on four considerations: 1) motifs with many neighbors in their neutral network can be considered gateways to a multitude of evolutionary pathways; 2) motifs interacting with multiple domains are multifunctional, which creates the potential for further specialization through EAC; 3) possible competition between activators and repressors; 4) different TF availability in cell types and development times. Such changes in binding interactions introduce further regulatory nuances and potentially introduce adaptive changes in phenotype. The evolution of binding interaction hinges on changes in TFBSs (Arnone and Davidson 1997; Stone et al. 2001); therefore, we tried to identify traits that might help characterize evolutionary important binding motifs. We found a set of binding motifs that we consider to be potential points of evolutionary change in regulatory networks, because they act as bridges to new functions. The motifs are unexpectedly high in A/T-content. This is especially noteworthy as A/T-rich-binding motifs like the TATA-box are the slowest emerging motifs (Behrens and Vingron, 2010). This is likely related to their importance in developmental processes and therefore high level of conservation. For example, homeoboxes are known to be involved in developmental processes, and binding to TATA-boxes. Their abundance in the data set could be a possible explanation for the abundance of A/T-rich motifs in the subsets. However, the observed trends do not disappear after excluding homeoboxes from the analysis. The possibility

of the identified motifs possessing high evolutionary potential persists and is worth to be experimentally analyzed and further investigated.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Andreas Lange and Brennen Heames for valuable comments and corrections on the article. We acknowledge support from the Open Access Publication Fund of the University of Münster.

Literature Cited

Aguilar-Rodriguez J, et al. 2017. A thousand empirical adaptive landscapes and their navigability. Nat Ecol Evol. 1:45.
Aharoni A, et al. 2005. The 'evolvability' of promiscuous protein functions. Nat Genet. 37(1):73–76.
Albert R, Barabasi L. 2002. Statistical mechanics of complex networks. Rev Mod Phys. 74(1):47.
Altschul SF, et al. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.
Amoutzias GD, et al. 2004. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. EMBO Rep. 5(3):274–279.
Anderson TA, et al. 2005. Sequence determinants of a conformational switch in protein structure. Proc Natl Acad Sci U S A. 51:18344–18349.
Arnone MI, Davidson EH. 1997. The hardwiring of development: organization and function of genomic regulatory systems. Development 124:1851–1864.
Badis G, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. Science 324(5935):1720–1723.
Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet. 5:101–113.
Behrens S, Vingron M. 2010. Studying the evolution of promoter sequences: a waiting time problem. J Comput Biol. 17(12):1591–1606.
Berg J, et al. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol Biol. 4:42.
Bettstetter C. 2002. On the minimum node degree and connectivity of a wireless multihop network. In: Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing; 2002 June 09–11; Lausanne, Switzerland: ACM. p. 80–91.
Billeter M, et al. 1993. Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. J Mol Biol. 234(4):1084–1093.
Bornberg-Bauer E. 1997. How are model protein structures distributed in sequence space? Biophys J. 73(5):2393–2403.
Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol. 3(7):e245.
Chaudhary J, Skinner MK. 1999. Basic helix-loop-helix proteins can act at the E-box within the serum response element of the c-fos promoter to influence hormone-induced promoter activation in Sertoli cells. Mol. Endocrinol. 13(5):774–786.
Crocker J, et al. 2015. Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell 160(1–2):191–203.
Darieva Z, et al. 2010. A competitive transcription factor binding mechanism determines the timing of late cell cycle-dependent gene expression. Mol Cell. 38(1):29–40.
Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. Science 311(5762):796–800.
Depristo MA. 2007. The subtle benefits of being promiscuous: adaptive evolution potentiated by enzyme promiscuity. HFSP J. 1(2):94–98.
DeVal S, et al. 2008. Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. Cell 135:1053–1064.
Dror I, et al. 2015. A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res. 25(9):1268–1280.
Elena SF, Carrera J, Rodrigo G. 2011. A systems biology approach to the evolution of plant-virus interactions. Curr Opin Plant Biol. 14(4):372–377.
Farley EK, et al. 2015. Suboptimization of developmental enhancers. Science 350(6258):325–328.
Gadau J, et al. 2012. The genomic impact of 100 million years of social evolution in seven ant species. Trends Genet. 28(1):14–21.
Gamsjaeger R, et al. 2007. Sticky fingers: zinc-fingers as protein-recognition motifs. Trends Biochem Sci. 32(2):63–70.
Gehring WJ. 1992. The homeobox in perspective. Trends Biochem Sci. 17(8):277–280.
Gehring WJ. 1993. Exploring the homeobox. Gene 135(1–2):215–221.
Graves BJ, Petersen JM. 1998. Specificity within the ets family of transcription factors. Adv Cancer Res. 75:1–57.
Gutierrez-Hartmann A, et al. 2007. ETS transcription factors in endocrine systems. Trends Endocrinol Metab. 18(4):150–158.
Hagberg AA, et al 2008. Exploring network structure, dynamics, and function using networkX. In: p 11–15.
Hittinger TC, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449(7163):677–681.
Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng. 9:90–95.
Karreth FA, et al. 2014. Target competition: transcription factors enter the limelight. Genome Biol. 15(4):114.
Kaufmann E, Knochel W. 1996. Five years on the wings of fork head. Mech Dev. 57(1):3–20.
Krishna SS, et al. 2003. Structural classification of zinc fingers: survey and summary. Nucleic Acids Res. 31(2):532–550.
Laity JH, et al. 2001. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 11(1):39–46.
Lenhard B, et al. 2012. Metazoan promoters: emerging characteristics and insights into tn sequence space? ranscriptional regulation. Nat Rev Genet. 13(4):233–245.
Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454:762–765.
Moore AD, et al. 2008. Arrangements in the modular evolution of proteins. Trends Biochem Sci. 33(9):444–451.
Murre C, et al. 1994. Structure and function of helix-loop-helix proteins. Biochim Biophys Acta 1218(2):129–135.
Mustonen V, Lässig M. 2005. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc Natl Acad Sci U S A. 102(44):15936–15941.
Nakagawa S, et al. 2013. DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc Natl Acad Sci U S A. 110(30):12349–12354.
Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 37:77–82.
Newman MEJ. 2003. The structure and function of complex networks. SIAM Rev. 45(2):167–256.
Oikawa T, Yamada T. 2003. Molecular biology of the Ets family of transcription factors. Gene 303:11–34.

Otting G, et al. 1990. Protein–DNA contacts in the structure of a homeodomain–DNA complex determined by nuclear magnetic resonance spectroscopy in solution. EMBO J. 9(10):3085–3092.
Payne JL, Wagner A. 2014. The robustness and evolvability of transcription factor binding sites. Science 343:875–877.
Reiter F, et al. 2017. Combinatorial function of transcription factors and cofactors. Curr Opin Genet Dev. 43:73–81.
Schmitz JF, et al. 2016. Mechanisms of transcription factor evolution in Metazoa. Nucleic Acids Res. 44(13):6287–6297.
Schofield PN. 1987. Patterns, puzzles and paradigms: the riddle of the homeobox. Trends Neurosci. 10(1):3–6.
Schulz KN, et al. 2015. Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo. Genome Res. 25(11):1715–1726.
Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond B Biol Sci. 255(1344):279–284.
Seidman SB. 1983. Network structure and minimum degree. Soc Netw. 5(3):269–287.
Sikosek T, et al. 2012. Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. Proc Natl Acad Sci U S A. 109(37):14888–14893.
Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface. 11(100):20140419.
Simola DF, et al. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome Res. 23(8):1235–1247.
Slattery M, et al. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell 147(6):1270–1282.
Slattery M, et al. 2014. Absence of a simple code: how transcription factors read the genome. Trends Biochem Sci. 39(9):381–399.
Song L, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21(10):1757–1767.
Spitz F, Furlong EE. 2012. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 13(9):613–626.
Stampfel G, et al. 2015. Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature 528:147–151.
Stewart AJ, Hannenhalli S, Plotkin JB. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics 192(3):973–985.
Stone JR, Wray GA. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol. 18(9):1764–1770.
Verger A, Duterque-Coquillaud M. 2002. When Ets transcription factors meet their partners. Bioessays 24(4):362–370.
Veron AS, Kaufmann K, Bornberg-Bauer E. 2006. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. Mol Biol Evol. 24(3):670–678.
Wagner A. 2008. Robustness and evolvability: a paradox resolved. Proc Biol Sci. 275(1630):91–100.
Wei G-H, et al. 2010. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. 29(13):2147–2160.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8:206–216.
Wroe R, Chan HS, Bornberg-Bauer E. 2007. A structural model of latent evolutionary potentials underlying neutral networks in proteins. HFSP J. 1(1):79–87.
Wunderlich Z, Mirny LA. 2009. Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. 25(10):434–440.
Yanez-Cuna JO, et al. 2012. Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. Genome Res. 22(10):2018–2030.

Associate editor: Josefa Gonzalez