doi:10.1016/j.jmb.2005.04.054 J. Mol. Biol. (2005) 350, 402–415

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Correlated Evolutionary Pressure at Interacting Transcription Factors and DNA Response Elements Can Guide the Rational Engineering of DNA Binding Specificity

Michele Raviscioni[1,2,5], Peili Gu[3], Minawar Sattar[3], Austin J. Cooney[3] and Olivier Lichtarge[1,4,5]*

[1]*W. M. Keck Center for Computational and Structural Biology, Houston TX, USA*

[2]*Program in Physiopathology of the Menopause, University of Milan, Italy*

[3]*Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston TX 77030 USA*

[4]*Program in Structural and Computational Biology and Molecular Biophysics and Program in Developmental Biology, and Cellular and Molecular Biology, Baylor College of Medicine, Houston TX 77030, USA*

[5]*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston TX 77030 USA*

**Corresponding author*

Understanding the molecular mechanisms of the specific interaction between transcription factor proteins and DNA is key to comprehend the regulation of gene expression and to develop technologies to engineer transcription factors. Thus far, although there have been several attempts to elucidate protein–DNA interaction through amino acid–base recognition codes, sequence based profiles, or physical models of interaction, the greatest successes in engineering DNA binding specificity remain experimental. Here we present the first systematic evidence of correlated evolutionary pressure at interacting amino acid residues and DNA base-pairs in transcription factors, and show that it can be used to rationally engineer DNA binding specificity. The correlation is between the relative evolutionary importance of protein residues and DNA bases, measured, respectively, in terms of the Evolutionary Trace (ET) rank and information entropy. The evolutionarily most important residues interact with the most conserved base-pairs within the response element while residues of least importance interact with the most variable base-pairs. The correlation averages 0.74 over 12 unrelated families of transcriptional regulators, including nuclear hormone receptors, basic helix–loop–helix, ETS- and homeo-domain family. To test the predictive power of this correlation, we targeted a mutational swap of top-ranked ET residues in a transcription factor, LRH-1. This redirects LRH-1 binding as predicted and showed that, in this case, evolutionary importance and binding specificity are coupled sufficiently strongly for the Evolutionary Trace to guide the computational design of DNA binding specificity. This establishes the existence of evolutionary importance correlation at protein–DNA interfaces, and demonstrates that it is a useful principle for the rational engineering of binding specificity.

Keywords: evolutionary trace; protein–DNA interaction; protein engineering; nuclear receptors; gene expression

Introduction

The specific interaction between transcription factors and their cognate DNA response elements is critical for the regulated expression of genes and is therefore under strong evolutionary pressure. When mutations occur that alter such interaction, the resulting change in gene expression can be incompatible with cell survival, and therefore be quickly eliminated from the population, or can establish an evolutionary advantage that can

Present address: M. Sattar, Department of Immunology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

Abbreviations used: ET, evolutionary trace; ER, estrogen receptor; b-Zip, basic zipper; PDB, Protein Data Bank.

E-mail address of the corresponding author: lichtarge@bcm.tmc.edu

ultimately give rise to a new variant. The rules governing protein–DNA recognition are also of great interest to engineer the DNA binding specificity of transcription factors and to regulate the expression of their target genes, with potential applications in gene therapy.

Early studies, supported by the first crystal structures of protein–DNA complexes, hypothesized that particular amino acid residues would specifically recognize a given DNA base-pair by virtue of their chemical properties, and this led researchers to propose the existence of a degenerate protein–DNA recognition code that governed protein–DNA binding specificity.[1,2] As more protein–DNA structures were solved, however, it became clear that for increasingly diverse mechanisms of DNA sequence recognition, no single universal "code" could provide a complete picture.[3] The structural properties of the protein–DNA interface have since been analyzed in great detail[4–6] and the effects of the structural environment in which amino acid–base interactions occur have been evaluated.[7,8] This revealed both recurrent patterns in protein–DNA interaction as well as, overall, a great diversity of structural and biochemical interactions even among conserved amino acid–base interactions. Thus, on the one hand, Luscombe *et al*. analyzed the conservation of amino acid residues in three groups of DNA binding protein families: non-specific (sequence-independent DNA binding), highly specific (all members of the family bind the same DNA sequence) or multi-specific (different members of the family bind different DNA sequences).[9] They showed that base interacting residues in multi-specific proteins undergo mutation more frequently than in other families, and that this is consistent with the variation in DNA binding specificity. They also showed that even though general principles governing protein–DNA interaction exist, each protein family must be studied separately to understand the determinants of DNA target recognition. On the other hand, Havranek *et al*. have proposed a physical model for protein–DNA interaction based on simple macromolecular energetic calculations and fast algorithms for sampling side-chain conformations.[10] Their method could recover the native DNA binding specificity from the test set, suggesting that it might be useful to redesign protein–DNA interfaces and to predict DNA binding specificity. A probabilistic protein–DNA recognition code has also been proposed to model the DNA interaction of C2H2 Zn-finger proteins. Its ability to predict the DNA binding specificity of this family of transcription factors was tested retrospectively but not applied yet, to our knowledge, to the rational engineering of their DNA binding specificity.[11]

The difficulty in identifying a coherent and general set of protein–DNA binding rules stems from the many components that influence the interaction of a protein with its DNA target, many of which are family-specific. These include local access to DNA, protein folding and dynamics, induced molecular fit, and formation of multi-protein complexes involving transcriptional co-factors. In fact, the most successful examples of DNA binding specificity engineering to date are entirely based on experimental techniques for the affinity selection of C2H2 Zn-finger proteins with custom DNA sequences.[12–15]

A different approach relies on the fact that all these components necessarily evolved to achieve the specificity we see today. Preliminary studies suggest that the analysis of protein–DNA interaction in the context of evolution can identify at least some of the critical determinants of specificity. The Evolutionary Trace (ET) analysis of the nuclear receptors' DNA-binding domain correctly identified key determinants of binding specificity and suggested, based on their patterns of evolutionary variations, that their coordinated exchanges may predictably switch DNA binding specificity.[16] Similarly, the identification of residues whose variation correlates with the grouping of bacterial transcription factors in orthologs and paralogs suggested residues that may determine protein–DNA interaction specificity.[17] These studies are limited, since they focused on a small number of protein families and did not test experimentally the rational engineering of transcription factors, but they suggest two related hypotheses. First, that Evolutionary Trace residues play an important role in determining DNA-binding specificity. This suggests in turn, that their variations cannot be independent of the variations of the DNA binding target itself. For example, in nuclear receptors the more variable trace residues are systematically in structural contact with the more variable bases in the response element.[16] Hence a second, more general hypothesis is that the amino acid and base-pair determinants of interaction specificity evolve at a correlated rate, such that residues under strong evolutionary pressure must interact with bases under a similar pressure, and *vice versa*.

This study aims to directly test for the correlated evolutionary pressure of amino acid residues and their interacting base-pairs at the binding interface of multiple types of transcription factors. The hope is to obtain a family specific evolutionary "code" of protein–DNA coupling that can be used to rationally manipulate the specificity of protein–DNA interaction. Co-evolution will be measured as a correlation between the relative evolutionary importance of protein residues, measured over the protein family evolutionary tree, and their interacting DNA base-pairs, measured over an alignment of cognate DNA response elements. The approach relies on the ET to rank the relative evolutionary importance of protein residues.[18,19] ET identifies sequence variations within a protein family that always correlate with evolutionary divergence: trace residues are invariant within each branch of the evolutionary tree but vary between them and their rank depends on how early in evolution they show this property, starting from the root as top

ranking. The evolutionary importance of the DNA base-pairs was measured from the variability of each position in a multiple sequence alignment of experimentally verified cognate response elements. By using crystal structures of specific protein–DNA complexes we identified the pairs of interacting amino acids and base-pairs and studied the correlation between the respective measures of evolutionary importance.

We find recurring correlations at the DNA interface between evolutionarily important residues and the DNA bases they contact in their cognate response elements. These correlations vary somewhat depending on the type of transcription factor and on the availability of experimentally proven DNA sequence data, but they are statistically significant and support the hypothesis of a correlated evolution between transcription factors and their DNA target sequences. Moreover, to test whether such a correlation is strong enough to support a family-specific and evolutionary trace-based code to guide rational protein engineering, we let our evolutionary study guide the computational design of DNA binding specificity in LRH-1, a transcriptional regulator of stem cell proliferation and differentiation.[20] We mutated two DNA interacting residues in the transcription factor LRH-1 into cognate residues from the steroid receptors branch of the evolutionary tree. As anticipated, the DNA binding specificity of LRH-1 changed accordingly. This work provides the first large-scale evidence for correlated evolutionary pressure at the molecular interface in multiple transcription factors. It also demonstrates, at least for LRH-1, that these correlations are sufficient to guide the rational design of binding specificity. Thus, the evolutionary tree in combination with ET is likely to provide general computational design guidelines to manipulate a transcription factor's DNA binding specificity with possible applications to gene expression for cellular engineering and gene therapy.

Results

The evolutionary importance of interacting amino acids and DNA bases is correlated

To investigate whether a protein residue and its interacting DNA base-pairs are under similar evolutionary pressure, we defined three parameters: a measure of evolutionary importance for the amino acid residues, a measure of evolutionary importance for the DNA base-pairs, and a set of rules to identify the interacting residue-base partners at the interface of a protein–DNA complex. The ET was used as a measure of the evolutionary importance of protein residues in a multiple sequence alignment.[18] The same quantity cannot be used to measure the evolutionary importance of base-pairs from response elements because these sequences are too short and too few to build robust evolutionary trees. Furthermore, in most cases, the different architecture of the response elements between ortholog transcription factors makes it impossible to align DNA sequences across the entire evolutionary tree. Therefore, we initially measured the evolutionary pressure experienced by each base-pair by means of the information content, or entropy, a simpler measure of sequence variability,[21–23] that we applied to an alignment of

Table 1. The dataset used in this study includes 22 transcription factors belonging to ten evolutionary unrelated families

Protein	Family	PDB	No. DNA seqs	No. prot seqs
GR	Nuclear rec.	1glu	72	375
ER	Nuclear rec.	1hcq	37	375
RAR	Nuclear rec.	1dszA	28	375
THB	Nuclear rec.	2nllB	12	375
RXR	Nuclear rec.	2nllA	13	375
USF-1	bHLH	1an4	14	142
MyoD	bHLH	1mdy	11	142
Myc	bHLH	1nkp	11	142
Elk-1	ETS	1dux	15	81
Ets-1	ETS	1k79	34	81
Pu-1	ETS	1pue	12	81
p53	P53	1tsr	17	54
Nfat	REL	1a02N	10	84
NF-kB	REL	1vkx	48	84
Exd	Homeod.	1b8iB	14	181
Gal4	Gal4	1d66	33	118
Pax6	Paired-Pax	1k78	10	72
Cap	CAP	1j59	28	136
Jun	b-ZIP	1a02J	31	365
Fos	b-ZIP	1a02F	20	365
GCN4	b-ZIP	1ysa	18	365
CEBP	b-ZIP	1h89	21	365

All proteins included satisfy the following conditions: (1) availability of a crystal structure of the specific protein–DNA complex; (2) availability of at least ten experimentally verified DNA response element sequences in Transfac 6.0 public release;[26] (3) being analyzable by Evolutionary Trace, which requires building of the family evolutionary tree.

cognate response elements. Successively, for the two families where it was possible, we tested whether the result is affected by the evolutionary span of the response element alignment by performing the same analysis with an alignment that includes several different orthologs. Finally, for every amino acid residue we defined the interacting DNA base-pairs by calculating hydrogen bonds and non-bonded interactions across the protein–DNA interface in the crystal structure of specific protein–DNA complexes with the publicly available programs HBplus[24] and Nucplot,[25] which are based on the distance and geometrical relationship between residue and base.

Our analysis was conducted on 22 protein–DNA complexes from ten of the most important transcription factors' families. These choices were based primarily on the availability of both protein and DNA sequence data and a crystal structure of the specific protein–DNA complex (Table 1). But we also focused only on transcription factors whose ortholog members specifically recognize different DNA response elements, in order to identify the determinants of such specificity. For example, protein families with high specificity for a single response element (e.g. Trp repressor, TATA binding proteins) do not allow for the correlation between protein and DNA sequence variation because all members of the family tree specifically recognize the same response element sequence. Also not included were, at the other extreme, protein families that bind DNA promiscuously and without specificity (polymerases, DNases, integration host factors). The dataset comprises all the transcription factors listed in Transfac 6.0 public release[26] which: (i) have at least ten experimentally verified DNA response elements (we defined ten as the minimum number of DNA sequences necessary for a meaningful entropy calculation); (ii) have a crystal structure of the specific protein–DNA complex; (iii) can be analyzed with the ET method. The C2H2 zinc-finger proteins were excluded because the short length of the sequences (25–30 amino acid residues, on average), the high sequence identity, and the huge number of artificial sequences makes the evolutionary tree overly uncertain. The evolutionary importance of protein residues and DNA base-pairs were normalized on a 0–1 scale, with 0 being the top ranking, evolutionarily most important.

An example of how the correlations between amino acid ET rank and base-pair entropy are obtained is shown in Figure 1 for the estrogen receptor's (ER) DNA binding domain. The information entropy of each DNA base is computed from an alignment of 37 ER cognate response elements and is shown in red. The average evolutionary rank of the interacting protein residues is in blue. The panel shows the correlation between the two measures of evolutionary importance of DNA base-pairs and amino acid residue, both rescaled on a 0–1 scale, with 0 being the most important evolutionary rank. A Pearson's correlation coefficient of 0.87, although calculated only on seven data pairs, represents a positive quantitative measure of a correlation that can also be visually appreciated.

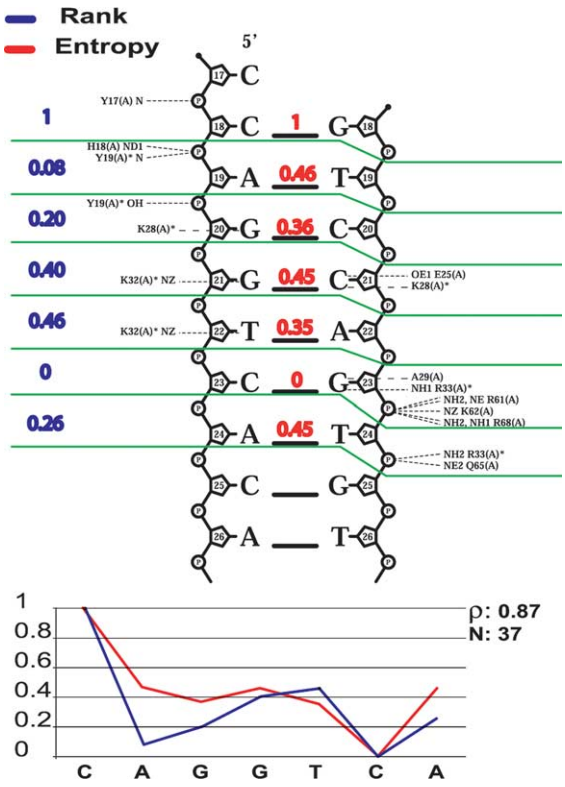

Figure 1. Correlation of the information entropy and the ET rank values for the estrogen receptor and its interacting DNA response element. The protein–DNA interactions, including hydrogen bonds and neighbors interactions, are calculated with HBplus and represented with Nucplot. The entropy values (red) for each base-pair and the average ET rank (blue) for their interacting residues are then correlated. When one base-pair interacts with more than one amino acid residue, the average rank value for the interacting residues is considered. ρ indicates the Pearson's correlation coefficient and *N* the number of DNA response elements retrieved from Transfac 6.0 public release.

The same analysis repeated for 22 different transcription factors confirms the generality of this correlation, as shown in Figure 2. Each panel shows on the *x*-axis the reference DNA response element sequence from the crystal structure of the protein–DNA complex. As before, the *y*-axis shows the entropy value of each base-pair (in red) and the average ET rank of the residues interacting with that base-pair (in blue). Evolutionarily more important base-pairs (low entropy value) generally interact with top ranking, evolutionarily more important amino acid residues (low ET rank average value). Conversely, less important base-pairs (high entropy value) are in contact with poorer ranking, less important amino acid residues (high ET rank average value). The Pearson's correlation coefficient

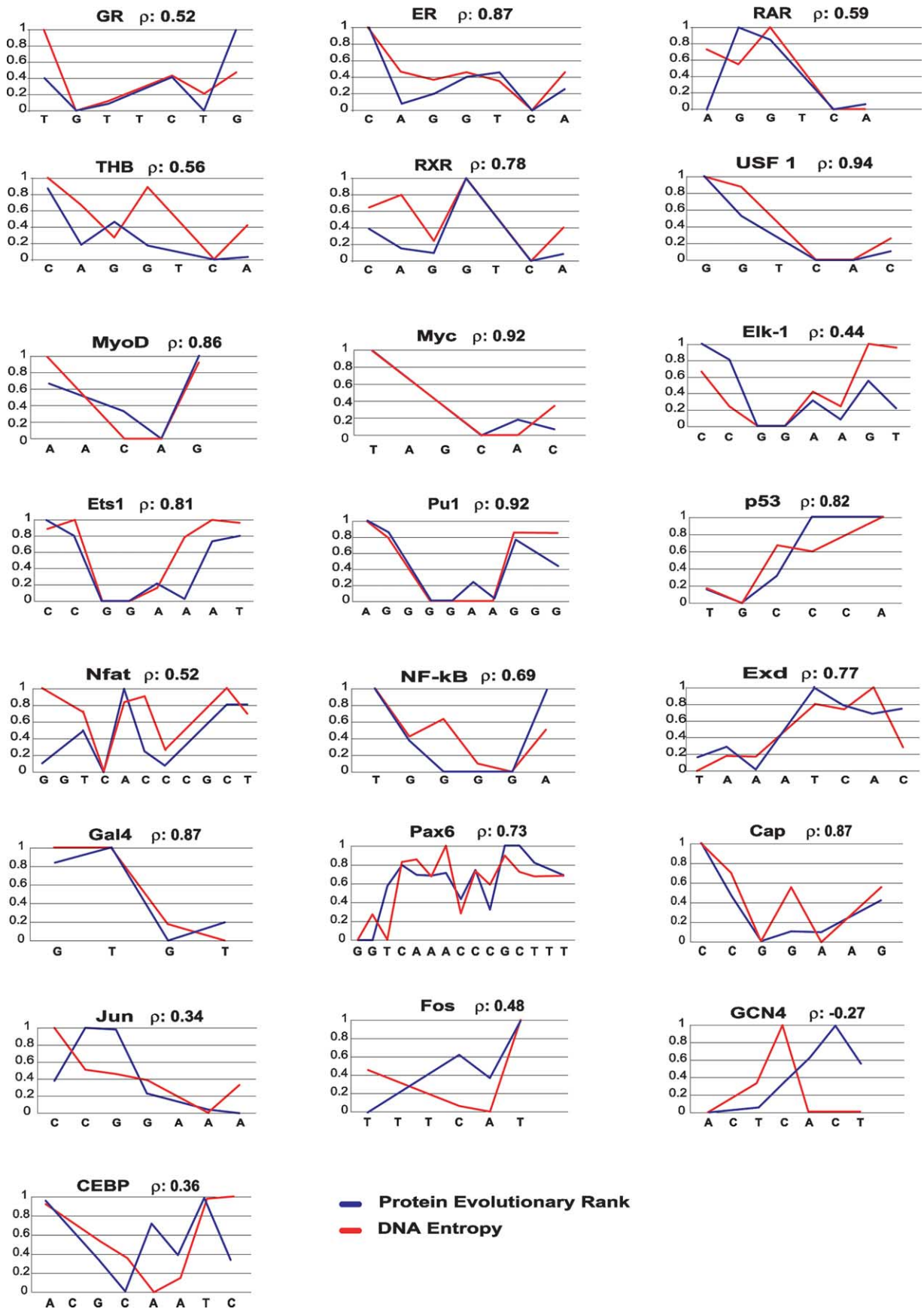

Figure 2. Protein ET rank and DNA entropy values are correlated in all but four of the specific protein–DNA complexes analyzed. For the 22 transcription factors analyzed, on the *x* axis the sequence of the DNA response element is indicated. The bases that do not make contact with the protein, as calculated by HB-plus, are shown in grey. The ET rank

is below 0.5 for only four proteins (Jun, Fos, GCN4, CEBP, last four panels of Figure 2) that are all members of the basic zipper (b-Zip) family, but it is greater than 0.7 for 12 proteins (55%), and greater than 0.8 for eight proteins (36%).

As pointed out above, evolutionary importance is measured over the entire evolutionary span of the protein family, but only over the narrower span of ortholog cognate response elements that bind the same protein. To test whether the observed correlation depends on the evolutionary span of the DNA sequence alignment, we compared the results obtained using alignments of cognate DNA response elements with the correlation obtained by aligning all the available DNA sequences for the orthologs included in this study. This comparison cannot be done for all transcription factors. Sometimes not only the sequences recognized by ortholog transcription factors vary, but also the architecture of the response elements, making the alignment impossible. Nuclear receptors, for example, can bind DNA response elements organized as directed repeats, inverted repeats or palindromic sequences, with spacer sequences that can vary from 0 bp to 9 bp.[27] Even when an alignment of DNA response elements can be obtained for the entire transcription factor family, the issue remains of which representative crystal structure to use to define the map of protein–DNA interactions. Thus, different orthologs establish different structural couplings with DNA and the bonds formed by one can switch to different bases in another, or not be formed at all. Another limitation derives from the poor availability of experimentally verified natural DNA response elements: in our dataset, only four transcription factor families are present with more than two ortholog members. We analyzed the bHLH and ETS families, because in each case a signature sequence motif in the DNA response elements (nCAn and (C/A)GGA(A/T), respectively) facilitates the alignment of ortholog response elements. The upper panels of Figure 3(a) show the correlations obtained for each ortholog member of the bHLH family using the alignment of cognate DNA response elements and the cognate contact map. The lower panels show, for each ortholog, the correlations calculated by using a single DNA sequence alignment that spans the DNA sequences recognized by all three ortholog proteins. The cognate contact map was used in order to preserve the correct structural coupling. Figure 3(b) shows the same comparison for the ETS family. The results show that the correlation between the evolutionary importance of interacting protein residues and DNA base-pairs is detected (Pearson's correlation coefficient >0.7) both when using a smaller alignment of cognate sequences or a larger alignment including sequences recognized by ortholog members of the same protein family. Thus, we show that correlated evolution for different variants of a protein can be detected when entropy is calculated for an alignment that includes DNA response elements specific for each variant. This also strengthens the result obtained for protein families for which a "global" alignment of ortholog DNA response elements cannot be obtained.

To reliably assess the statistical significance of these correlations we pooled pairs of ET rank and DNA entropy values for the entire dataset to generate a large data sample, excluding the four b-Zip proteins. Figure 4 represents the average DNA entropy value for 0.1 protein rank intervals for the entire dataset. The parametric Pearson's correlation coefficient is 0.73 and the non-parametric Spearman's rank order correlation coefficient is 0.75. With both methods the correlation measured is statistically significant (p-value $<10^{-5}$).

Thus, the correlated evolution at the molecular interface between protein residues and their contact bases can be detected and is statistically significant for all cases with the exception of the members of the b-Zip family. If these proteins are included, the Pearson's and Spearman's correlation coefficients decrease to 0.64 and 0.66, respectively, but remain significant with a p-value $<10^{-5}$. However, we believe that other factors determine DNA binding in these proteins besides the interface considered here, explaining both the consistent lack of correlation in that family and justifying their exclusion.

Top ranking DNA interacting trace residues interact preferentially with the purine and pyrimidine rings

Next, in order to characterize the properties of DNA interacting trace residues, we analyzed their physico-chemical features. The purpose is not to provide a detailed and comprehensive characterization of DNA interacting amino acid residues: other groups have already performed such thorough analyses on more representative datasets.[6,9,28] Our goal is to test whether DNA interacting residues have rank-specific properties. We divided all of the DNA interacting residues in our dataset (288) in four bins, ranging from the top 25% to the bottom 25% evolutionary ranking, and calculated the frequency of basic, acidic, hydrophobic and aromatic DNA interacting residues in each bin. We also calculated the frequency of residues interacting with the sugar–phosphate backbone.

(shown in blue) and the DNA entropy (shown in red) have been rescaled on a 0–1 scale, with 0 being the evolutionarily most important and 1 the least important. When one base-pair interacts with more than one amino acid residue, the average rank value for the interacting residues was considered. p is the Pearson's correlation coefficient and N the number of DNA response elements in the alignment.

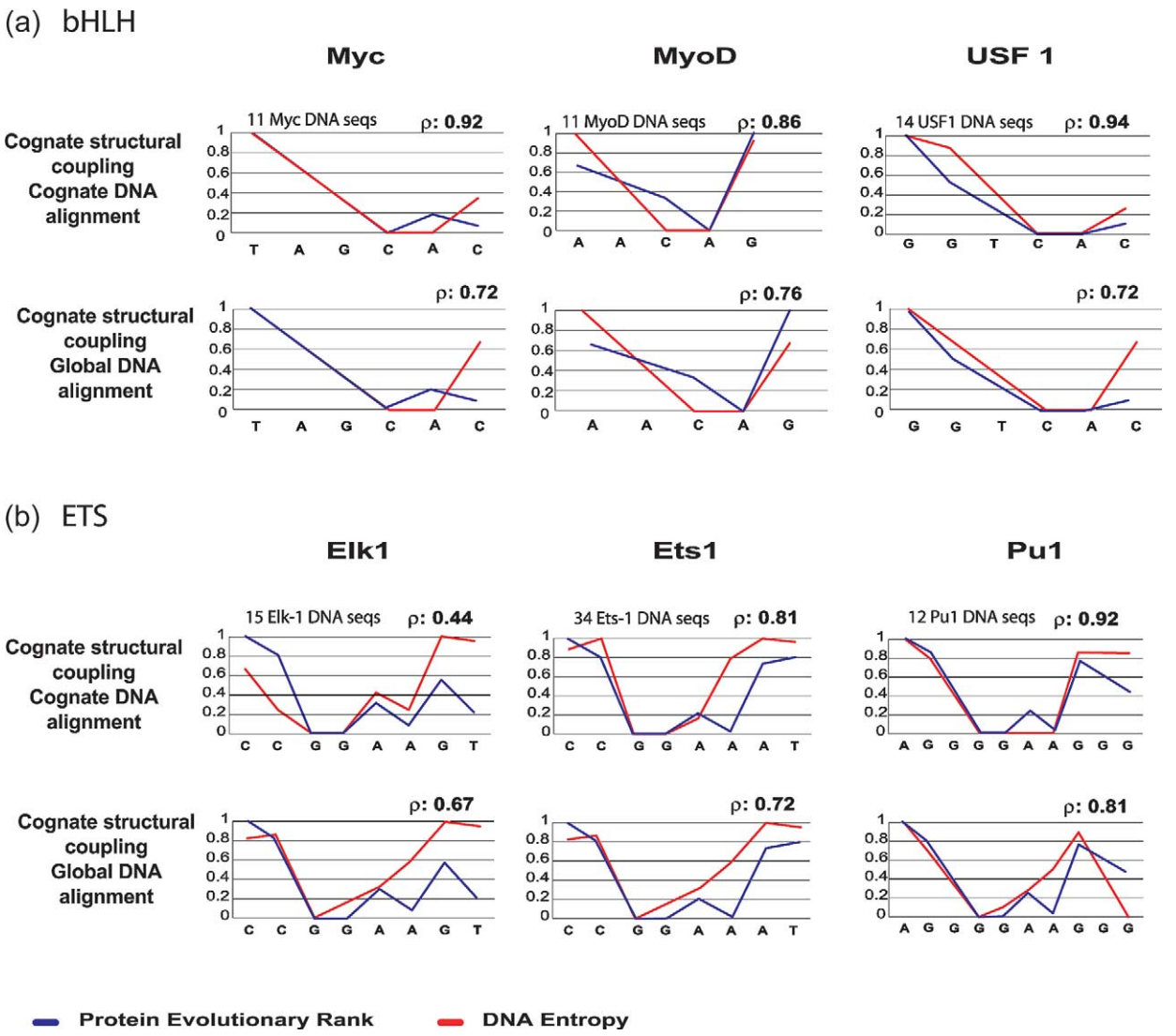

Figure 3. The correlated evolution between interacting amino acids and base-pairs is independent from the span of the DNA sequence alignment and the variation of the structural couplings between ortholog transcription factors. To test whether the observed correlation is dependent on the evolutionary span of the DNA sequence alignments, we compared the results obtained using alignments of cognate DNA response elements with the correlation obtained by aligning all the available DNA sequences for the orthologs included in this study. (a) The upper panels show the correlations obtained for each ortholog member of the bHLH family using the alignment of cognate DNA response elements and the cognate contact map. The lower panels show, for each ortholog, the correlations calculated by using a single DNA sequence alignment that includes the DNA sequences recognized by all three proteins (36 sequences total). (b) The same comparison is shown for the ETS family (61 sequences are present in the single DNA alignment). The correlation between the evolutionary importance of interacting protein residues and DNA base-pairs can be detected both by using a smaller alignment of cognate sequences or a larger alignment including sequences recognized by ortholog members of the same protein family.

Figure 5 shows that residues with different physicochemical features are equally distributed in the four bins. Interestingly, backbone-interacting residues are more frequent at poorer ranks, while purine/pyrimidine ring-interacting residues (corresponding to the complementary fraction) are more frequent among the top ranks. This suggests that the evolutionarily most important residues are more likely to be involved in interactions responsible for the direct read-out of the DNA sequence, and therefore for the definition of specificity.

Evolutionary identification of the DNA binding specificity determinants in LRH-1

Next, we wished to test whether the correlated evolution suffices to identify the determinants of DNA binding specificity. We combined the evolutionary trace of nuclear receptors, to which the orphan receptor LRH-1 belongs, with that family's evolutionary tree to generate a key that describes how top-ranked, DNA interacting residues match response element variations, and hence binding specificity. The evolutionary tree (Figure 6)

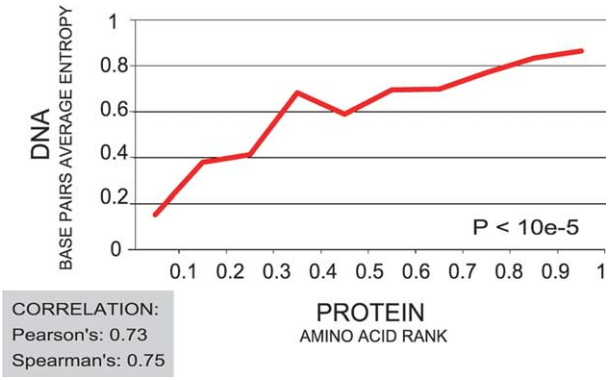

Figure 4. The overall correlation between protein ET rank and DNA entropy is statistically significant. In order to obtain a number of rank and entropy data pairs large enough to perform a significant statistical analysis, their values for each transcription factor were pooled and the Pearson's and Spearman's correlation coefficients were calculated. The *x* axis shows the protein ET rank on a 0–1 scale in 0.1 intervals and the *y* axis the average value of entropy of the corresponding DNA base-pairs.

shows five different types of consensus DNA response elements in the nuclear receptors family. All of the target DNA sequences share a common AG – – CA pattern, but significant variations occur at the two middle bases and at the 5′ terminus. In particular, the tree reveals that the variation between the DNA response element of LRH (consensus responsive element TCAAGGTCA) and the steroid receptors group (consensus responsive element TCAAGAACA) correlates with the

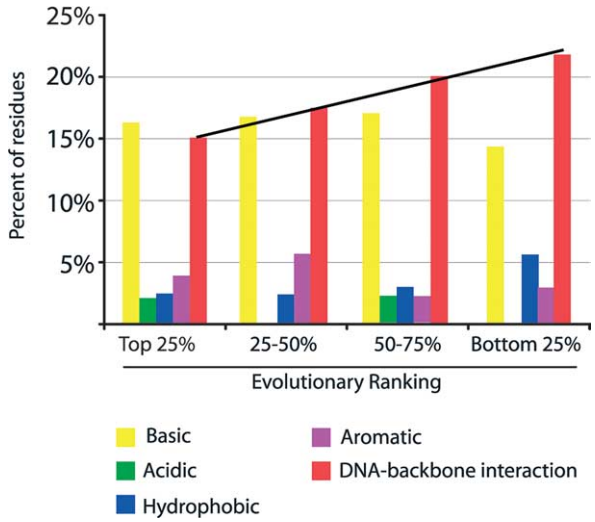

Figure 5. Top ranking DNA interacting trace residues interact preferentially with the purine and pyrimidine rings. The histogram shows the frequency of basic, acidic, aromatic and hydrophobic DNA interacting residues for four groups of DNA interacting residues, from the top 25% evolutionarily most important to the bottom 25% least important. It also shows the frequency of residues interacting with the sugar–phosphate backbone.

variation of two amino acid residues (G462 and E458 according to the numbering in PDB 1glu, or G129 and E125 in Swiss-Prot entry P45448) that in the representative crystal structure appear to interact with the differentiating DNA base-pairs. These two residues, ranking 14 and 20, respectively, are the top ranking residues that vary between the LRH-1 and the steroid receptors branches, and are therefore most likely involved in the formation of the intermolecular couplings responsible for binding specificity. Residues 474, 511 and 512 also vary between the two branches, but their poor evolutionary rank (154, 227 and 164, respectively) suggests only a secondary role in the definition of binding specificity. Based on our model, therefore, the experimental mutation of residues 462 and 458 should be sufficient to switch the DNA binding specificity of LRH-1 to a steroid-like response element.

ET guided experimental engineering of LRH-1 DNA binding specificity

To test experimentally whether the correlated evolution between protein residues and their interacting DNA base-pairs can rationally guide the re-design of a transcription factor's DNA binding specificity, we engineered the orphan receptor LRH-1. LRH-1 is a monomeric C4 zinc finger protein member of the Fushi Tarazu factor I subfamily and binds a consensus CAAGGTCA sequence:[29,30] an alpha-helix of the zinc-finger domain is inserted into the DNA major groove where it interacts with the core AGGTCA and a C-terminal extension of the DBD binds the 5′ CA in the minor groove.[27,31]

We mutated residues G462V and E458G in the LRH-1 Zn-finger domain, so as to change the DNA binding specificity from the natural TCAAGGTCA consensus element (DR0) to the mutant TCAA GAACT sequence (DR0m), which is recognized by steroid responsive transcription factors (Figure 6). To compare the DNA binding affinities of wild-type and mutant LRH-1 for the mutant DR0(m) probe, unlabeled DR0 and DR0(m) probes were used to compete for the binding to ^{32}P-labeled probe DR0(m) (Figure 7). The binding signal of wild-type LRH-1 (lanes 1–3) was dramatically decreased with increasing unlabeled probe DR0, while unlabeled DR0 did not compete on the binding of mutant LRH-1 (lanes 7–9), indicating that wild-type LRH-1 has much higher affinity for DR0 than for the DR0(m) sequence. Conversely, the increasing unlabeled DR0m probe sharply reduced the binding signal of mutant LRH-1 (lanes 10–12) but not of the wild-type receptor (lanes 4–6), thus demonstrating the specificity of binding of the mutant LRH-1 for the DR0m probe. These experiments show that we have generated a mutant LRH-1 with ortholog DNA binding specificity by targeting mutations to only two top ranked DNA interacting amino acid residues that are different in LRH and steroid receptors.

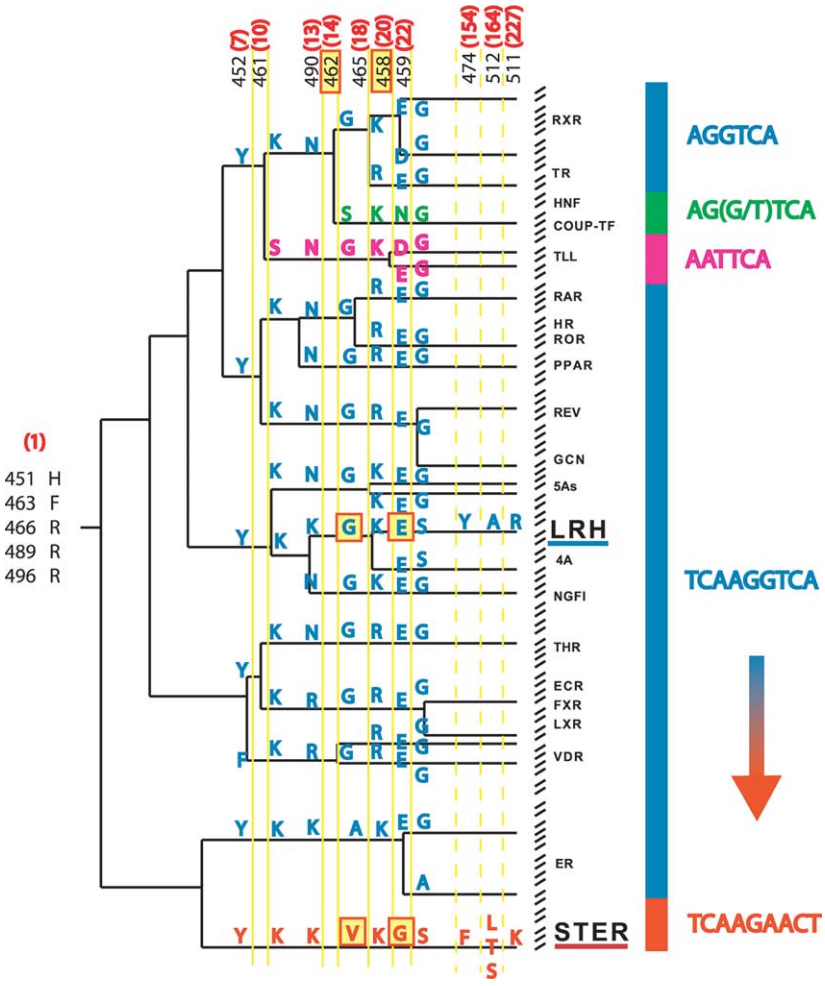

Figure 6. The evolutionary tree of the nuclear receptor's family can be used as a key to guide the rational engineering of DNA binding specificity. The vertical yellow bars define the ET rank and the sequence variation of the LRH-1 DNA-interacting residues are shown at the intersection with the tree. The residue numbers are indicted in black and their ET rank in red. The top ranking residues whose evolutionary variation correlates with the variation in the response element, form the interfacial couplings primarily responsible for binding specificity. Residues 462 and 458 (shown in a yellow box) are the top ranking residues that vary between the LRH-1 (G,E) and the steroid receptors branch (V,G), and are therefore identified as the determinants of DNA binding specificity between these two groups. The structure of the glucocorticoid receptor (PDB, 1glu) was used to model LRH-1.

Discussion

Protein–DNA correlated evolutionary pressure is a key component of specific DNA binding

This study reveals a surprisingly large and common degree of evolutionary correlation at the molecular interface between transcription factors and their response elements. This likely reflects that the timely and precise control of gene expression is vital to development and homeostasis, so that the biophysical and biochemical properties of transcriptional regulatory proteins and their target DNA are exquisitely tuned to one another. Yet, many factors can take part in determining their interaction: the chromatin structure and DNA accessibility, the protein and DNA structures and their flexibility for induced fit, and the formation of higher-order protein complexes on DNA. The transcription factor–DNA contact by itself may therefore only account for a small part of the overall interaction and hence of its specificity. Nevertheless, our findings suggest that few direct protein–DNA contacts undergo random mutation without consequences for gene expression that result ultimately in elimination or in natural selection and evolution (i.e. speciation).

This points to a model of evolution at macromolecular interfaces in which interfacial couplings can be treated as evolutionary units. This is reasonable, since it is the structural couplings that are the basis for binding complementarity. As with any set of evolutionary units, some will be under more selective pressure than others. But since the rate at which couplings change depends on the rate at which either of their components changes, it is logical to guess that the couplings for which changes are most likely to be selected against, would in turn involve residues and bases that are relatively equally unlikely to mutate. And conversely, couplings that are evolutionarily permissive, in the sense that they are not always rapidly eliminated from the population upon alteration, should involve residues and bases that are relatively equally tolerant to change and where mutations are associated with evolutionary divergence. Thus, over all evolutionary instances, this model predicts that the mutation tendency of structural contact partners at an interface should be matched, even when the structural couplings change, as they do, in different orthologs. In this

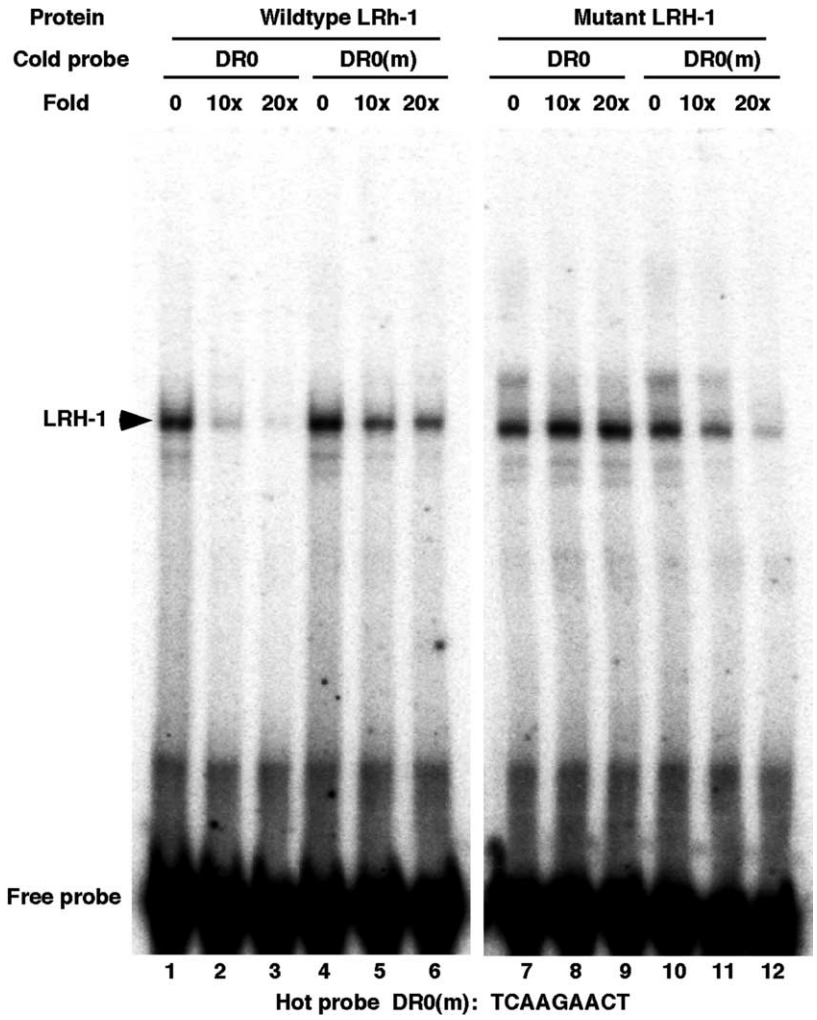

Figure 7. The mutant LRH-1 displays ortholog DNA binding specificity *in vitro*. The DNA binding specificity of the wild-type and mutant LRH-1 was assessed by measuring the amount of cold wild-type (DR0) or mutant (DR0m) DNA probe bound to the proteins in the presence of a constant amount of radiolabeled mutant (DR0m) probe. Radioactivity is displaced from wild-type LRH-1 by the wild-type DR0 probe and not by the mutant DR0m. Conversely, radioactivity is displaced from mutant LRH-1 only by cold DR0m and not by cold DR0, thus showing that mutant LRH-1 has acquired the new, desired ortholog DNA binding specificity.

model no distinction is made among the many ways in which a coupling may vary: whether it is the residue, the base or both that vary the coupling is altered regardless. In other words, the evolutionary correlation as defined here extends beyond compensatory mutations (rationalized as able to preserve the integrity of an interaction) to also include all cases where either the residue, or the base, or both change and that change is associated with speciation.

This argument is the basis for our hypothesis that correlated evolution occurs at the DNA interface of transcription factors, and it is closely related to evolutionary tracing (which identifies residues where mutations systematically correlate with divergence). Thus, it is logical to organize our search for correlation around the concept of evolutionary importance measured by ET. Even though for practical reasons we are forced to simply measure information entropy for the bases themselves, it is striking the degree to which the importance of the residues and bases are correlated during evolution of the protein–DNA interface of the ten most important families of transcription factors. Thus, indeed, evolutionary pressure is closely mirrored on either side of the structural protein–DNA interface. Furthermore, this symmetric distribution of evolutionary pressure can be detected regardless of the evolutionary span of the DNA sequence alignment and of the structural couplings, confirming that the differences between ortholog structural couplings have evolved in parallel with the sequence variations at their sides.

As expected, this correlation is not entirely uniform, although it is significant. Different protein folds imply different modes of interaction, such that mutations on either the protein or DNA might affect a coupling differently, and might be compensated through structural rearrangements. The one exception is for the proteins from the b-Zip family (Jun, Fos, GCN4, CEBP). Previous studies have described DNA recognition by b-Zip transcription factors as "fuzzy" because a single protein can bind different DNA target sequences with similar affinity.[9] To account for this non-specificity, it has been shown that the Asn235 side-chain of GCN4 (PDB code, 1ysa) can rotate to recognize two different base-pairs that are positioned diagonally on opposite strands (TC′ or CT′).[32] Thus, this protein family can

tolerate DNA sequence variations that do not affect binding specificity and therefore the evolutionary pressure is not distributed symmetrically on the protein and DNA sides. For this reason, in the b-Zip family, the correlated evolution between protein and DNA target sequence might represent a smaller contribution to the specificity of DNA interaction, as we observe.

The magnitude of the correlations between the remaining transcription factors and their respective target DNA response elements is perhaps surprisingly large. First, since most bind DNA response elements as part of larger complexes, few would be expected to bear a strong signature of binding specificity. Second, attempts to detect interfacial correlations based on the concept of simultaneous, compensatory mutations have generally been much weaker. Here, however, we obtain a strong signal from a simpler question: what is the relative evolutionary pressure on the residues and bases, and are these pressures structurally correlated. The fact they are may suggest a general feature of structural systems in biology, namely, that the evolutionary importance of a structural coupling tends to be reflected equally on its different components. This fits in with basic mechanical engineering principles that structural couplings depend on the load of the connected parts,[33] and it is reflected within proteins by the general clustering property of evolutionarily important amino acids.[34] Here the distributed load on a structural coupling is reflected by the mirrored evolution across the protein and DNA interfaces from transcriptional regulatory networks. It could be interesting in the future to determine whether this is a general principle across macromolecular interfaces.

The evolutionarily most important DNA-interacting residues are key determinants of binding specificity

The observation that residues interacting with the purine/pyrimidine ring experience little sequence variation and therefore stronger evolutionary pressure than those interacting with the sugar–phosphate backbone is not entirely novel.[9] However, it serves as a further control for the ability of the Evolutionary Trace to identify the key determinants of DNA binding specificity. The interactions with the sugar–phosphate backbone are important for DNA binding in that they provide stabilization to the complex and contribute to the read-out of the DNA conformational features. However, base-interacting residues are responsible for the read-out of the DNA sequence, and therefore the major determinants of DNA binding specificity. Thus, the top ranking trace residues play a major role in defining a protein's DNA binding specificity, suggesting that their rational mutation can redirect binding to an ortholog DNA sequence.

Protein–DNA co-evolution can guide the engineering of transcription factors' DNA binding specificity

This study demonstrates not only that coupled protein residues and DNA base-pairs co-evolve, but also that this is immediately useful for the rational design of binding specificity. On the one hand, ET has already been successfully applied to engineer functional specificity at a protein–protein interface[35] and to swap developmental pathways[36] by exchanging top-ranked residues between members of different evolutionary branches. On the other hand, DNA binding specificity switch experiments have already been carried out by simply replacing amino acids identified by pair-wise sequence comparison.[37] However, based on the widespread observation of evolutionary correlations across interfaces of transcriptional factors, for the first time we suggest that the systematic analysis of top-ranked trace residue variations across different evolutionary branches provides a simple, family-specific code for protein–DNA interaction specificity.[16] For example, in the nuclear receptor's evolutionary tree, all the branches bind to an AG – – CA like element, suggesting that these amino acid–base-pair interactions to the terminal bases define essential DNA binding features for that entire family. But the interfacial couplings to the middle bases show greater sequence variability, both on the DNA and protein side. These variations are associated with five different types of response elements, and the corresponding trace residues difference can be read off directly from the evolutionary tree. These correlations correspond to parallel clips in the red and blue graphs of Figure 6 that define DNA binding specificity. Thus, swapping these residues should exchange DNA binding specificity. Indeed, by mutating only two residues (G129V and E125G) we were able to redirect LRH-1 DNA sequence binding specificity to that of another branch of the tree, proving the predictive power of an evolution-based computational design strategy to engineer protein–DNA interaction.

Evolution-guided design of tools for functional genomics

This analysis of co-evolution creates a new tool to study functional genomics and gene expression engineering. LRH-1 is a key transcription regulator involved in development, metabolism and steroidogenesis.[20] In a recent study, Gu *et al*. showed that during early mouse development it activates the expression of the gene Oct4, a key player in embryonic development and cellular differentiation.[38,39] The LRH-1 mutant, therefore, will help investigate Oct4 regulation and its role in the regulation of embryonic stem cell proliferation and differentiation *in vitro*.

More generally, this study extends from protein–protein to protein–DNA the demonstration of

ET-guided computational re-design of functional specificity. Unlike wholesale modular design of proteins to manipulate pathways, these experiments target mutational swaps to only a minimum number of top-ranked trace residues.[40] As a result, this is sufficient to alter one specific activity while leaving other functional sites intact. Others have also used targeted mutations to transplant a new function onto a protein scaffold[41] or proposed the use of protein building blocks.[42] In one sense, our approach is less general, since it requires evolutionary relatedness between "donor" and "acceptor" proteins, but it can also be seen as more general, in the sense that a functional exchange can be conceived among related proteins of any family that can be traced. Thus, in the context of gene therapy or cell engineering, one may be able to apply ET-based computational design to multiple protein components of a signaling or transcription pathway, in order to gain control over a complex "physiological" regulatory effect.[43]

In summary, this work provides the first strong and widespread evidence for correlated evolutionary pressure across molecular interfaces. This phenomenon was observed in several and diverse families of transcription factors and their target response elements. Thus, it emerges as a general principle rooted in the consistency of the evolutionary load on individual structural couplings resulting from the force that natural selection exerts on molecular recognition and specificity. As shown experimentally, the Evolutionary Trace can identify the key components of this correlation to guide the computational design of a transcription factor's DNA binding specificity: this further establishes a new systematic approach to rationally engineer proteins and the transcriptional pathways they control. It is tempting to hypothesize that correlated evolution as defined here could, in the future, be detected in larger multi-component gene regulatory complexes, and provide important clues for their reconstruction or manipulation. Here, the reduced dimensionality of the DNA molecule in terms of its sequence, structure, and interaction with proteins, facilitates the task of correlating the relative evolutionary importance across the protein interface to the DNA bases. However, the observation described here leads us to believe that similar studies might be conducted to generalize the observation of evolutionary consistent structural couplings to protein–protein interfaces, thereby opening new possibilities for the study of macromolecular complexes.

Experimental Procedures

Sequence alignments and determination of the relative evolutionary importance

For each transcription factor, the protein sequences were retrieved with a BLAST[44] search using the sequence of the crystal structure as a query. The sequences were aligned with CLUSTALW,[45] the alignments were pruned to eliminate gaps and fragments and traced in order to determine the relative evolutionary importance of each residue.[18,19] The sequences of DNA response elements specifically recognized by each transcription factor and naturally occurring in gene promoters were retrieved from Transfac 6.0 public release[26] and aligned with CLUSTALW. When the DNA sequences present in Transfac 6.0 exceeded in length the annotated response element contained in it, the alignments of the annotated response element were adjusted manually. The relative evolutionary importance of each base in the DNA multiple sequence alignment was determined based on the entropy, or information content, value, calculated as:

$$-\sum_{n=1}^{4} p \log p$$

where p represents the frequency of each of the four DNA bases.

The values of entropy and ET rank were normalized on a 0–1 scale, with 0 being the evolutionary most important (lower entropy or rank value), according to the transformation:

$$\frac{X - X\mathrm{min}}{X\mathrm{max} - X\mathrm{min}}$$

Structure-based identification of the interacting residues and correlation of DNA entropy and ET rank

The programs HBplus[24] and Nucplot[25] were used to calculate the hydrogen bonds and non-bonded interaction between protein residues and DNA bases in the crystal structures of specific complexes. If at least one type of interaction was detected between an amino acid residue and a DNA base, the residue and base-pairs involved were defined as interacting. A map of pairs of interacting DNA base-pairs and protein residues was obtained and the values of ET rank and entropy of the interacting partners were correlated. When one base-pair interacted with more than one amino acid residue, the average rank value for the interacting residues was considered. Statistical analysis of the correlation was carried out with the *R* statistical environment†.

Characterization of DNA interacting trace residues

DNA-interacting residues were classified as basic (K,H,R), acidic (D,E), hydrophobic (L,P,V,A,M,W,F,I), aromatic (W,F,Y,H). Sugar-backbone interacting residues were defined according to the HBplus/ Nucplot output when atoms in the deoxyribosephosphate were involved in a H-bond or a non-bonded interaction with the protein. The frequencies are calculated over the total number of residues in each group.

Experimental mutation of LRH-1

The construction of *in vitro* translation plasmid pT7-Myc-LRH-1 and mammalian expression vector pCMV-Myc-LRH-1 was described.[46] The mutant plasmids pT7-Myc-LRH-1(m) and pCMV-Myc-LRH-1(m)

† Team, R. D. C. (2004). R: A language and environment for statistical computing http://www.r-project.org)

were produced using the QuickChange XL site-directed mutagenesis kit (Stratagene) and performed according to the manufacturer's protocol.

In vitro DNA binding assay

In vitro translated LRH-1 proteins were incubated with ^{32}P-labeled DNA probes in binding buffer (25 mM Hepes (pH 7.9), 75 mM KCl, 0.2 mM EDTA, 1 mM DTT, 10% (v/v) glycerol, 1 mM PMSF and 1× Proteinase Inhibitor Cocktail (Roche)) and separated in 5% (w/v) polyacrylamide gels in 0.5×TBE buffer (Tris–borate, EDTA). The sequences of DNA probes are DR0: sense: 5′CTGACTGGGTAAGGTCAAGGCTATTCTAAAGTCG A3′ and antisense 5′TCAGTCGACTTTAGAATAGCC TTGACCTTACCCAG3′ and DR0(m) sense 5′CTGAC TGGGTAAGGTCAAGAATATTCTAAAGTCGA3′ and antisense 5′TCAGTCGACTTTAGAATATTCTTGACCT TACCCAG3′.

Acknowledgements

The authors thank Hui Yao for insightful discussion. M.R. thanks the Lichtarge laboratory for introducing him to the world of computational biology, Adriana Maggi and Paolo Ciana for inspiration and encouragement. This work was supported by grants from the National Science Foundation (DBI-0318415), National Institutes of Health (R01 GM066099), and March of Dimes (1-FY03-93) (to O.L.) and by the NIH NURSA orphan receptor program, grant U19DK62434-01 (to A.C.). M.R. was supported by a Fontana-Lionello fellowship from the Italian Association for Cancer Research and by a training fellowship from the W. M. Keck Center for Computational and Structural Biology.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.04.054

References

1. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
2. Pabo, C. O. & Sauer, R. T. (1984). Protein–DNA recognition. *Annu. Rev. Biochem.* **53**, 293–321.
3. Matthews, B. W. (1988). Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
4. Chen, S., Vojtechovsky, J., Parkinson, G. N., Ebright, R. H. & Berman, H. M. (2001). Indirect readout of DNA sequence at the primary-kink site in the CAP–DNA complex: DNA binding specificity based on energetics of DNA kinking. *J. Mol. Biol.* **314**, 63–74.
5. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
6. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
7. Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
8. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
9. Luscombe, N. M. & Thornton, J. M. (2002). Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **320**, 991–1009.
10. Havranek, J. J., Duarte, C. M. & Baker, D. (2004). A simple physical model for the prediction and design of protein–DNA interactions. *J. Mol. Biol.* **344**, 59–70.
11. Benos, P. V., Lapedes, A. S. & Stormo, G. D. (2002). Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.* **323**, 701–727.
12. Isalan, M., Klug, A. & Choo, Y. (2001). A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature Biotechnol.* **19**, 656–660.
13. Tan, S., Guschin, D., Davalos, A., Lee, Y. L., Snowden, A. W., Jouvenot, Y. *et al*. (2003). Zinc-finger protein-targeted gene regulation: genome wide single-gene specificity. *Proc. Natl Acad. Sci. USA*, **100**, 11997–12002.
14. Ordiz, M. I., Barbas, C. F., 3rd & Beachy, R. N. (2002). Regulation of transgene expression in plants with polydactyl zinc finger transcription factors. *Proc. Natl Acad. Sci. USA*, **99**, 13290–13295.
15. Kim, J. S. & Pabo, C. O. (1998). Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. *Proc. Natl Acad. Sci. USA*, **95**, 2812–2817.
16. Lichtarge, O., Yamamoto, K. R. & Cohen, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325–337.
17. Mirny, L. A. & Gelfand, M. S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol Biol.* **321**, 7–20.
18. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
19. Madabushi, S., Gross, A. K., Philippi, A., Meng, E. C., Wensel, T. G. & Lichtarge, O. (2004). Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.* **279**, 8126–8132.
20. Fayard, E., Auwerx, J. & Schoonjans, K. (2004). LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis. *Trends Cell Biol.* **14**, 250–260.
21. Shenkin, P. S., Erman, B. & Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins: Struct. Funct. Genet.* **11**, 297–313.
22. Pei, J. & Grishin, N. V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

23. del Sol Mesa, A., Pazos, F. & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.
24. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
25. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (1997). NUCPLOT: a program to generate schematic diagrams of protein–nucleic acid interactions. *Nucl. Acids Res.* **25**, 4940–4945.
26. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I. *et al*. (2001). The TRANSFAC system on gene expression regulation. *Nucl. Acids Res.* **29**, 281–283.
27. Khorasanizadeh, S. & Rastinejad, F. (2001). Nuclear–receptor interactions on DNA-response elements. *Trends Biochem. Sci.* **26**, 384–390.
28. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
29. Nitta, M., Ku, S., Brown, C., Okamoto, A. Y. & Shan, B. (1999). CPF: an orphan nuclear receptor that regulates liver-specific expression of the human cholesterol 7alpha-hydroxylase gene. *Proc. Natl Acad. Sci. USA*, **96**, 6660–6665.
30. Sirianni, R., Seely, J. B., Attia, G., Stocco, D. M., Carr, B. R., Pezzi, V. & Rainey, W. E. (2002). Liver receptor homologue-1 is expressed in human steroidogenic tissues and activates transcription of genes encoding steroidogenic enzymes. *J. Endocrinol.* **174**, R13–R17.
31. Zhao, Q., Khorasanizadeh, S., Miyoshi, Y., Lazar, M. A. & Rastinejad, F. (1998). Structural elements of an orphan nuclear receptor–DNA complex. *Mol. Cell*, **1**, 849–861.
32. Ellenberger, T. E., Brandl, C. J., Struhl, K. & Harrison, S. C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein–DNA complex. *Cell*, **71**, 1223–1237.
33. Mischke, J. & Schgley, C., *Mechanical Engineering Design: Classic* (5th 2002 edit.) *Part III*.
34. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
35. Sowa, M. E., He, W., Wensel, T. G. & Lichtarge, O. (2000). A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl Acad. Sci. USA*, **97**, 1483–1488.
36. Quan, X. J., Denayer, T., Yan, J., Jafar-Nejad, H., Philippi, A., Lichtarge, O. *et al*. (2004). Evolution of neural precursor selection: functional divergence of proneural proteins. *Development*, **131**, 1679–1689.
37. Umesono, K. & Evans, R. M. (1989). Determinants of target gene specificity for steroid/thyroid hormone receptors. *Cell*, **57**, 1139–1146.
38. Pan, G. J., Chang, Z. Y., Scholer, H. R. & Pei, D. (2002). Stem cell pluripotency and transcription factor Oct4. *Cell Res.* **12**, 321–329.
39. Pesce, M. & Scholer, H. R. (2001). Oct-4: gatekeeper in the beginnings of mammalian development. *Stem Cells*, **19**, 271–278.
40. Lim, W. A. (2002). The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr. Opin. Struct. Biol.* **12**, 61–68.
41. Marvin, J. S. & Hellinga, H. W. (2001). Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc. Natl Acad. Sci. USA*, **98**, 4955–4960.
42. Tsai, H. H., Tsai, C. J., Ma, B. & Nussinov, R. (2004). *In silico* protein design by combinatorial assembly of protein building blocks. *Protein Sci.* **13**, 2753–2765.
43. Hardy, J. A. & Wells, J. A. (2004). Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* **14**, 706–715.
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
45. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
46. Gu, P., Goodwin, B., Chung, A. C., Xu, X., Wheeler, D. A., Price, R. R. *et al*. (2005). Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell Biol.* **25**, 3492–3505.

Edited by J. Thornton

(Received 15 February 2005; received in revised form 18 April 2005; accepted 22 April 2005)