Nucleic Acids Research Advance Access published August 4, 2010

Nucleic Acids Research, 2010, 1–10
doi:10.1093/nar/gkq683

Using protein design algorithms to understand the molecular basis of disease caused by protein–DNA interactions: the Pax6 example

Andreu Alibés[1,*], Alejandro D. Nadra[1], Federico De Masi[2], Martha L. Bulyk[2,3,4], Luis Serrano[1,5] and François Stricher[1]

[1]EMBL/CRG Systems Biology Research Unit, Center for Genomic Regulation, UPF, Barcelona, Spain, [2]Division of Genetics, Department of Medicine, [3]Department of Pathology, Brigham and Women's Hospital, [4]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA and [5]ICREA professor, Center for Genomic Regulation, UPF, Barcelona, Spain

Received May 30, 2010; Revised July 14, 2010; Accepted July 15, 2010

ABSTRACT

Quite often a single or a combination of protein mutations is linked to specific diseases. However, distinguishing from sequence information which mutations have real effects in the protein's function is not trivial. Protein design tools are commonly used to explain mutations that affect protein stability, or protein–protein interaction, but not for mutations that could affect protein–DNA binding. Here, we used the protein design algorithm FoldX to model all known missense mutations in the paired box domain of Pax6, a highly conserved transcription factor involved in eye development and in several diseases such as aniridia. The validity of FoldX to deal with protein–DNA interactions was demonstrated by showing that high levels of accuracy can be achieved for mutations affecting these interactions. Also we showed that protein-design algorithms can accurately reproduce experimental DNA-binding logos. We conclude that 88% of the Pax6 mutations can be linked to changes in intrinsic stability (77%) and/or to its capabilities to bind DNA (30%). Our study emphasizes the importance of structure-based analysis to understand the molecular basis of diseases and shows that protein–DNA interactions can be analyzed to the same level of accuracy as protein stability, or protein–protein interactions.

INTRODUCTION

The paired box gene 6 (PAX6) is a member of the Pax gene family of transcription factors (TFs) and it is mainly involved in tissue specification during early development (1). Pax6 is required for the multipotent state of retinal progenitor cells (2) and is usually related to the development of the eyes and sensory organs (3,4). Mutations in this TF are linked to eye diseases such as aniridia, foveal hypoplasia, cataracts and nystagmus (5). Because of its importance in human ocular disease and the vast amount of biological information regarding this protein, a database of disease-related mutations of PAX6 is available (6).

Most of the time, a specific disease can be described as the consequence of protein mutations, being a single one or a combination of several. However, establishing the exact effect on the function of protein based on its sequence alone is not trivial. The effects of mutations on protein stability and protein–protein interaction can be reasonably well predicted using protein design tools, as we previously demonstrated in the analysis of the relationship between the stability changes of the human phenylalanine hydroxylase and phenylketonuria (7). Similarly, mutations favoring protein aggregation or amyloid

*To whom correspondence should be addressed. Tel: +34 93 316 0258; Fax: +34 93 316 0099; Email: andreu.alibes@crg.es
Correspondence may also be addressed to François Stricher. Tel: +34 93 316 0258; Fax: +34 93 316 0099; Email: francois.stricher@crg.es
Present addresses:
Alejandro D. Nadra, Departamentos de Fisiología, Biología Molecular y Celular, y Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria (1428), Buenos Aires, Argentina.
Federico De Masi, Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby 2800 Denmark.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

disease in unstructured protein regions can be accurately predicted (8,9). However, similar studies have not been performed on mutations affecting protein–nucleic acid interactions, although studies predicting the effect of mutations on DNA recognition of specific sequences have been published (10–12).

Protein–DNA interactions are a key process in transcriptional regulation and replication. To carry out their function, DNA-binding proteins must find and bind to infrequent and small specific binding sites and discriminate them from a huge excess of non-specific DNA. Protein–DNA complexes involve direct and indirect interactions and there is not a general recognition code to predict base–residue interactions. For some well characterized families [zinc fingers (13,14), homeodomains (11,15) and bHLHs (16)] some general rules can be applied. However, for the majority of DNA-binding proteins the main way to identify the DNA recognition sequence is through experimental methods.

DNA-binding sites are traditionally characterized using a limited number of sequences by biochemical assays. However, in the last few years, several experimental techniques and an increasing number of sequenced genomes allowed a more detailed analysis. Several computational methods for discovering TF binding sites have been described (17,18). Experimental methods that challenge the protein to a library of DNA sequences and successively enrich those with high affinity have been developed, such as *in vitro* selection (19,20) or yeast or bacterial one-hybrid assays (21). Additionally, universal protein-binding microarrays (PBMs) (10,12,22,23) expose the protein to all possible DNA-binding site sequence variants making universal PBMs the only exhaustive technique available.

In the past years, different *in silico* approaches have been developed to predict DNA-binding site motifs for DNA-binding proteins using structures. There have been successful attempts either by using existing crystal structures (24–32), homology modeling (33) or by a docking approach (13). In particular, structure-based predictions were evaluated in zinc fingers (28,34) where a sensitivity to docking geometry was reported (35), and in meganucleases (30–32), highlighting the importance of having multiple templates to enhance the accuracy.

Here we use the protein design algorithm FoldX (http://foldx.crg.es) to do a global analysis of the effect of all described mutations on the paired box domain (PD) of Pax6. FoldX incorporates DNA base mutations, movement of DNA bases, automatic identification of the complementary base and is able to predict the effect of base mutations on DNA stability and binding to a target protein (30–32). First, we validated the DNA capabilities of FoldX by predicting both changes in affinity upon protein or DNA mutation and the DNA-binding specificity from structure of an extensive set of publicly available TF recognition patterns (motifs) and by testing new predictions against novel PBM-determined motifs. We then analyzed all disease-related mutations in Pax6, and looked for structural and energetic reasons that may impair its function and trigger different eye diseases. We looked for both destabilizing mutations and mutations in the interface of DNA that might have an effect on the Pax6 regulation of genes. We find that 88% of the mutations have a negative impact on the stability and/or binding energy of Pax6, thus impairing its function and causing disease.

MATERIALS AND METHODS

Quantitative benchmark sets

Two different data sets were used to quantitatively evaluate the performance of FoldX: one containing protein mutations in the interface with DNA and the other made of wild-type proteins interacting with different DNA sequences.

The set of protein mutants in protein–DNA complexes was initially taken from the ProNIT database (36). However, due to internal inconsistencies of the database, the actual data (physical data of the experiments and the change in interaction energy due to the mutation, $\Delta\Delta G_{int}$) were taken from the papers describing the original experiments. The final set contains 97 mutations, among them 59 conservative mutations (Supplementary Table S1).

The set of proteins binding to different DNA sequences was compiled in ref. (24). We analyzed those for which a crystal structure is available, and we compared the correlation factors obtained between experiments and predictions for each TF with our method and their software.

FoldX DNA force field

The classical four bases -A, C, G and T-, as well as the methylated A and C bases, were incorporated using standard FoldX parameters for defining atoms (37) (charges, Van der Waals radii, volumes, electrostatic interactions, solvation energies and hydrogen bond parameters) without parameter fitting. In order to better take into account the stacking of bases and their preferred conformation, we put an entropy-like term inside the Van der Waals clash term of FoldX for adjacent bases. This term was derived from a statistical analysis of all DNA structures in the Protein Data Bank (PDB) looking at each possible pairs of consecutive bases and looking at the angles made by the planes of each bases. Those angles were discretized and for each bin of 2°, an energy cost was calculated based on the probability P of having two bases in such angles [$\Delta G = -RT\ln(p)$].

Qualitative TF benchmark set

We selected all experimental motifs we found in three different databases (JASPAR database (38), TRANSFAC 7.0 Public 2005 database (39) and the UniPROBE (40) database of PBM-derived motifs). Each motif had to fulfill several requirements in order to be included in the validation set: (i) it had to be connected to a PDB structure through a SwissProt accession or a UniProt ID; (ii) this structure had to come from a crystallization experiment; (iii) contain at least one protein chain and double-stranded DNA that do not include any non-standard base; and (iv) all residues in the interface should be solved.

Mutation protocol for PWMs

For each template considered, the positions of the amino acid side chains and bases in the crystal structure were first energetically optimized using the FoldX *RepairPDB* function, which moves slightly all side chains to eliminate small Van der Waals clashes. Then, each base was mutated to the other three possible bases five times to increase the conformational space analyzed. Each mutation involves searching for the rotamer of the new residue whose energy is better, while slightly moving the side chains of the surrounding residues to accommodate for it. No backbone movement is included. Each single point mutation takes <60 s using a single CPU (Intel Xeon 3.00 GHz, 8GB of RAM).

Using the average value, the difference in the interaction energy with respect to the wild-type ($\Delta\Delta G_{int}$) was calculated, adding the difference in intramolecular clashes if they were higher than for the crystal structure. This extra term penalizes those DNA variants that may have a good binding energy, but are forced into the DNA structure. This function is graphically displayed as information content by means of the R package *seqLogo* (41), where the height of a given nucleotide is proportional to $\exp(-\Delta\Delta G_{int}/RT)$. When more than one structure/chain exists for a given protein, then the one with the better resolution was chosen. In the case of Gcn4, both 2DGC and 1YSA were used as they differ in binding site length. In all the cases, the same physical conditions were assumed: temperature of 298 K, pH of 7.0 and ion strength of 150 mM.

Divergence coefficient

In order to get a quantitative sense of how good or bad a PWM prediction is, we have used a divergence coefficient, *D,* that is the root mean square deviation between the probabilities in the logo:

$$D = \sqrt{\frac{\sum_{i,j} \Delta P_{ij}^2}{N}}$$

where ΔP_{ij} is the difference in probability of having base j in position i between the experimental and the FoldX-predicted PWM, and N is the binding site length. The way this coefficient is defined is so that the closer D is to zero, the better the prediction is. $\sqrt{2}$ is the maximum value and 0.87 is the value obtained when comparing a fully specific PWM and a completely unspecific one. Experimental logos coming from different methods have an average coefficient of 0.2 (see Supplementary Figure S1)

We consider that the prediction for a position (or the whole motif) fails if the coefficient for this position (or for the whole motif) is above 0.58, which is the value we obtain when comparing a PWM where a base in each position has 100% probability against a PWM where this probability has decreased to 50%. Also, as a reference in between the perfect logo ($D < 0.2$) and a failed one ($D > 0.58$), we use the value when the probability has only decreased to 66% ($D = 0.38$) (Supplementary Figure S2).

PBMs

PBM experiments were performed as previously described using 200 nM final concentration for Gcn4, cJun(216–318) and cFos and 'all 10-mer' universal PBMs synthesized on the 4×44 K format Agilent platform (16,23). For the cJun/cFos PBM experiment, both proteins were mixed and pre-incubated for 60 min at room temperature before applying to the microarray slide. Neither cJun(216–318) nor cFos are able to homodimerize (42). Therefore, only those instances where the proteins do form functional DNA-bound heterodimers are being detected by the Alexa488-labeled anti-GST antibody.

Microarray data were quantified using the GenePix 6.0 software (Axon) and data analysis was performed using the Seed and Wobble algorithm (15).

Pax6 thresholds

For the Pax6 analysis, we used a change in energy upon mutation (either of protein stability or of protein–DNA interaction) of 0.8 kcal/mol as the minimum threshold above which the stability or the interaction might be significantly affected. This is the value usually considered as FoldX error (37) and corresponds to a K_{d_mutant}/K_{d_WT} of around 4. The threshold of 1.6 kcal/mol, twice the FoldX error and corresponding to a K_{d_mutant}/K_{d_WT} of 15, was considered as minimum value for a change that very significantly affects binding.

Also, we looked at the divergence coefficient between the wild-type and mutant for each of the DNA positions, and we consider that any value above 0.58 is linked to a significant change in specificity.

Neutral Pax6 mutation set

We have taken all orthologs of the paired domain of Pax6 in the 6PAX structure from SwissProt (see Supplementary Figure S3), aligned their PDs to the human Pax6 and performed the necessary mutation to the 6PAX structure with FoldX to obtain them. We use this set as a neutral mutation set, as these are not mutations causing disease.

Scripts

The Python scripts used to wrap FoldX, speed up and semi-automate the PWM calculations and the creation of logos—that could be used for similar purposes—are available upon request. They take the user from an input PDB structure of a protein–DNA complex to a final recognition pattern.

Database submissions

GCN4-ATF and Jun-Fos data are deposited in the UniPROBE database at http://the_brain.bwh.harvard.edu/uniprobe/

RESULTS

Validation of the force field for dealing with DNA

We first used the set of entries in the ProNIT database (36) that are linked to a structure in the PDB. After manually

curating the data set, we obtained a quantitative benchmark with 10 protein–DNA complexes containing 97 mutations, 58 of which are conservative (i.e. any amino acid to Ala or Gly, Ile to Val or Thr) or isosteric (Asn-Asp, Gln-Glu, Thr-Val) (Supplementary Table S1). Correlation analysis between the experimental and predicted changes in binding energies for conservative mutations (Supplementary Figure S4) showed a significant correlation ($r = 0.64$), close to the one found for mutations in proteins [set from (37)] ($r = 0.73$). It is important to note that the SD between predicted and experimental values is similar in both cases (0.80 kcal/mol in the case of the protein mutants and 0.86 kcal/mol in the case of protein mutants in complex with DNA). However, it is clear from Supplementary Figure S4 that the scarcity of data for protein–DNA interaction, compared to protein stability (>1000 mutants), should introduce a note of caution in terms of assessment the reliability of mutations on protein–DNA binding energies. More points will be needed to see if the quality is the same.

As an alternative way to assess predictability we also tested for mutations in the DNA bases for the set of protein and DNA complexes that come from X-ray experiments and were used by Morozov *et al.* (24). In all cases, the correlation between predicted and experimental data was good and we obtained significantly better results for 10 of the 13 TFs (Table 1).

Prediction of DNA-binding profiles using FoldX

In order to get a current set of TF binding site motifs for an accurate comparison of predicted versus experimental profiles, we searched several databases of TF DNA-binding site sequence motifs [JASPAR (38), TRANSFAC (39) and UniPROBE (40)]. We selected all proteins in any of these databases with a co-crystal structure available at the PDB (see 'Materials and Methods' section for the selection criteria), in total 23 proteins.

Table 1. Correlation factors for changes in interaction energy upon change in DNA base predicted by FoldX and Morozov *et al.* (24). The cases where FoldX performs better than the dynamic model are shown in green, otherwise in red

PDB	Name	Correlation (24)		Correlation FoldX
		Contact model	Dynamic model	
1LMB	lambdaR	0.42	0.4	0.49
1TRO	trpR	0.052	0.39	0.76
6CRO	CroR	0.12	0.4	0.26[a]
1RUN	Crp	0.38	0.47	−0.03[b]
1MNN	Ndt80	0.63	0.74	0.83
1YRN	MAT_a1_alpha2	0.35	0.37	0.57
1AAY	zif268	0.064	0.1	0.7
1JK1	Zif268 D20 A	0.064	0.1	0.45
1ECR	Tus-Ter	0.71	0.25	0.39
1EFA	LacR	0.72	0.59	0.64
1HCQ	ER	0.33	0.28	0.37
1BHM	BamHI	–	0.24	0.28
1PUE	ETS_domain_of_PU.1	–	0.7	0.47

1AAY and 1JK1 are considered together in Morozov *et al.* (24).
[a]0.63 without two outliers.
[b]0.58 without three outliers.

Then, for these proteins we derived the DNA-binding site motifs using FoldX.

Profile comparison is not trivial, thus to attain an unbiased evaluation of derived DNA-binding profiles, we calculated an index D (based on a root mean square deviation between the experimental and the predicted motifs, see 'Materials and Methods' section) that, together with a visual inspection of the corresponding graphical sequence logos (Figure 1; Supplementary Figures S5 and S6; Table 2), facilitates comparison. The baseline for a perfect base prediction with this coefficient is 0.2, which is the coefficient we get for motifs of the same TF taken from different experimental methods (see Supplementary Figure S1). A D value of 0.87 is the value obtained when comparing a non-degenerate (100% specific) PWM and a completely degenerate one (i.e. no sequence preference at all). To be conservative, we considered D values above 0.58 as base mispredictions (see 'Materials and Methods' section for explanations about the D determination and Supplementary Figure S2 for illustration) even if they still contain some information. For 30% of the motifs, we found almost perfect agreement between predicted and experimental values (Figure 1A and B), while for 65% of the cases the agreement is quite good (i.e. shows a useful prediction power; Figure 1C).

In total, out of the 255 bases in the binding motifs analyzed, we consider that FoldX mispredicts 51 (20%) bases even if some of the information content still remains in the prediction. In fact, there is on average a slight loss of specificity (measured in terms of difference in information content per position) of 0.21, between the experimental and predicted logos (Table 2). For those cases where the experimental specificity is partially or totally lost (85% of all the failing positions), we observe in the crystal structure the absence of any DNA contacting residue or specific water-mediated interaction (e.g. the fourth position in the 1UBD logo, Supplementary Figure S5, where we cannot find the preference towards an adenine that is found experimentally). In all the cases where specificity is gained (15%), the backbone configuration of DNA, through steric hindrance, dictates the change.

We have analyzed the capacity of several experimental and predicted PWMs to find similar hits. We have used as probability threshold for each PWM as the one which made possible to reach all combinations of nucleotides whose individual probability in the experimental PWM was >10%. The predicted PWMs being less specific than the experimental ones have then a lower threshold. With those probability thresholds, we have scanned all possible sequences of the length of the binding site and counted the number of occasions where the same hit was found with both PWMs (Supplementary Table S2A). We can compare these results with what we obtain if we compare the scanning results obtained from a PWM resulting from a non-exhaustive technique, such as the ones found in the TRANSFAC database, and another PWM from an exhaustive technique such as PBMs (some examples in Supplementary Table S2B). So, the same way as PWMs coming from two different experimental methods, predicted PWMs can get as hits almost all

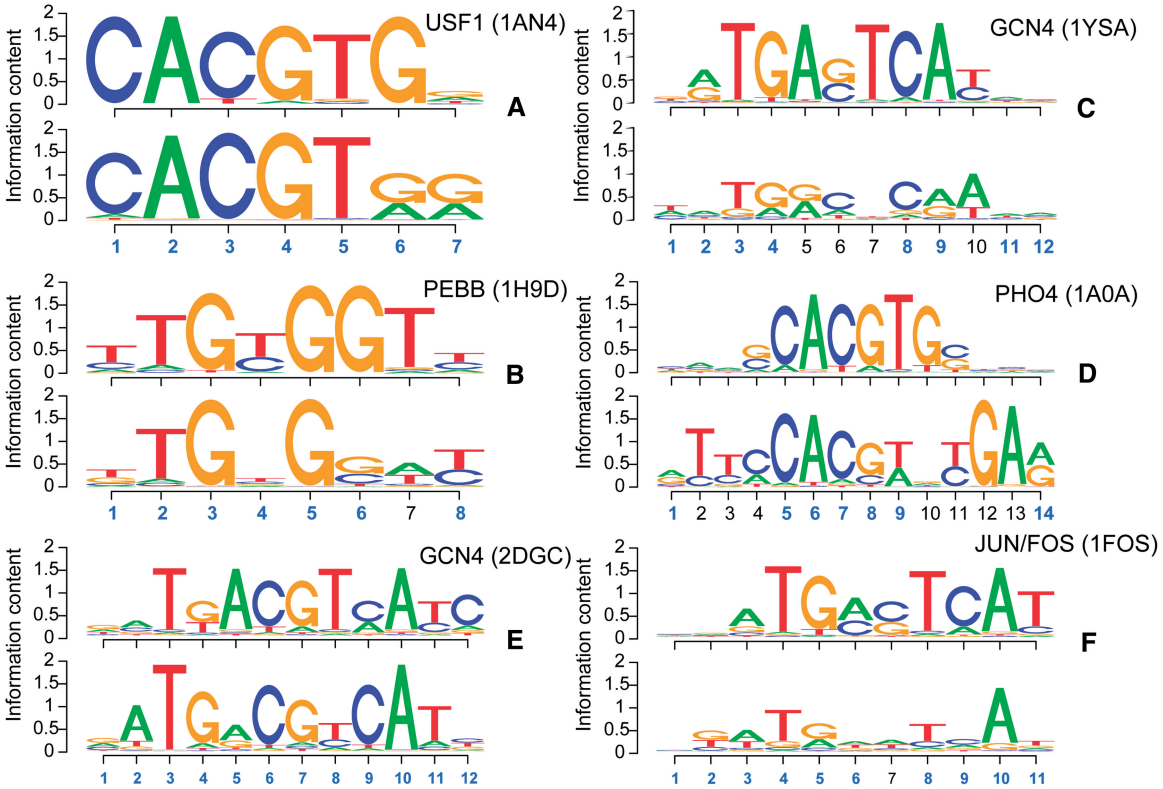

Figure 1. Qualitative evaluation of FoldX-derived DNA-binding profiles. Experimental logos for a subset of proteins displaying different *D* values (see Table 2) are presented above each FoldX prediction. Correctly predicted positions according to our criterion (individual coefficient $D < 0.58$) are shown in blue, while those mispredicted (individual coefficient $D > 0.58$) are shown in black. For the PBM-derived logos (**E** and **F**), Enrichment Scores for the top seeds resulting in the PBM PWMs are 0.485 for Gcn4 (ATF) and 0.497 for Jun/Fos.

experimental hits, but they would be surrounded by a lot of false positives. This finding leads us to suggest the use of PWMs as scanning tools should be restricted to short sequences where it is known to be a binding site, such as the results of a ChIP-on-chip experiment.

Validating predictions with comprehensive DNA-binding profiles

With the aim of testing our predictive power, we decided to predict the DNA-binding specificity for proteins without known DNA-binding site motifs coming from exhaustive experimental evaluations. Among all crystal structures of protein–DNA complexes, we selected Gcn4 bound to the ATF/CREB site (PDB id 2DGC) and the Jun/Fos heterodimer (PDB id 1FOS). The two proteins were produced and their DNA-binding specificity determined by universal PBMs (23) (Figure 1E and F). Comparison with the predicted DNA motifs by FoldX shows that we correctly predict the ATF/CREB site since there are no failing positions and the overall *D* is 0.32. For the second TF, we found also a good prediction ($D = 0.40$) with the base with the highest probability corresponding to the one in the experimentally derived motif in all cases, except the central base that does not contact the protein (Figure 1F).

Structural analysis of Pax6 and its relation to disease

Pax6 is made of a PD, whose structure in complex with DNA has been solved (43) (PDB id 6PAX), a DNA-binding homeo box domain (HD) and a proline-serine-threonine rich domain. The PD contains two globular subdomains: the N-terminal subdomain (residues 4–63) is composed of three α-helices folded like a homeodomain and a β-hairpin, the C-terminal one (residues 80–136) is also composed of three α-helices in a homeodomain-like fold. The two subdomains interact symmetrically with the major groove of the DNA through a Helix-Turn-Helix motif and the rest of the specific contacts come from the binding to the minor groove of the linker (residues 64–79) that connects the two subdomains (Figure 2). The DNA recognition pattern of the PD is known (44) and our *in silico* prediction (Figure 3) yields a good *D* of 0.37. Although an experimental PWM for the HD is available (10), there is no structure of it in complex with DNA that we can use to predict the effect of the four missense mutations that have been found there.

The Human PAX6 Mutation Database (6), single repository of all mutations observed in patients, contains, at the moment of writing, 50 mutations that can be mapped to the structure of the PD of Pax6 (Figure 4). We analyzed the effect of those mutations both on the

Table 2. Quantitative evaluation of FoldX-derived DNA-binding profiles

PDB	TF	Resolution (Å)	Organism	Binding site length	Source	*D*	IC/N exp	IC/N pred
1AN4	**USF1**	**2.9**	**Hsa**	**7**	**J**	**0.22**	1.61	1.63
1MNN	NDT80	1.4	Sce	12	U	0.31	1.12	1.07
2DGC	**GCN4**	**2.2**	**Sce**	**12**	**New**	**0.32**	**1.02**	**1.10**
1R4I	ANDR	3.1	Rno	9	J	0.33	0.85	0.72
1AAY	EGR1	1.6	Mmu	11	U	0.35	1.03	1.29
1PUE	SPI1	2.1	Mmu	11	U	0.36	0.92	0.68
6PAX	PAX6	2.5	Hsa	14	J	0.37	0.99	0.69
1H9D	**PEBB**	**2.6**	**Hsa**	**8**	**J**	**0.37**	**1.32**	**0.98**
1E3O	PO2F1	1.9	Hsa	10	U	0.39	0.99	1.07
1FOS	**JUN/FOS**	**3.05**	**Hsa**	**11**	**New**	**0.40**	**0.96**	**0.49**
3COQ	GAL4	2.4	Sce	19	U	0.40	0.69	0.67
1UBD	TYY1	2.5	Hsa	6	J	0.41	1.35	1.21
1MNM	MCM1	2.3	Sce	20	U	0.46	0.53	0.69
3DFX	GATA3	2.7	Mmu	8	U	0.46	1.44	1.05
1ODH	GCM1	2.85	Mmu	10	U	0.47	1.18	0.74
1APL	MTAL2	2.7	Sce	9	T	0.49	0.77	1.32
1H89	MYB	2.45	Mmu	11	U	0.49	0.99	0.74
1EGW	MEF2 A	1.5	Hsa	10	J	0.50	1.57	0.77
2ERE	LEU3	3	Sce	8	U	0.50	1.22	0.83
1YSA	**GCN4**	**2.9**	**Sce**	**12**	**U**	**0.50**	**1.08**	**0.53**
1IF1	IRF1	3	Mmu	12	T	0.52	1.54	0.89
1KB2	VDR	2.7	Hsa	15	J	0.53	1.36	0.54
1IG7	MSX1	2.2	Mmu	9	U	0.54	1.18	0.83
1MDY	MYOD1	2.8	Mmu	10	T	0.55	1.10	0.71
1A0 A	**PHO4**	**2.8**	**Sce**	**14**	**U**	**0.60**	**0.82**	**1.13**

Source: J, JASPAR (38); T, TRANSFAC (39); U, UniPROBE (40) and newly reported here.
Divergence coefficient values (see 'Materials and Methods' section) for each structure analyzed and a group is assigned according to its divergence coefficient ($D < 0.38$: white, $0.38 \leq D < 0.58$: light gray, $D > 0.58$: dark gray). The average information content for each position, IC/N, for the experimental and predicted logos is also shown. Logos presented in Figure 1 are in bold and are examples of different divergence coefficients.

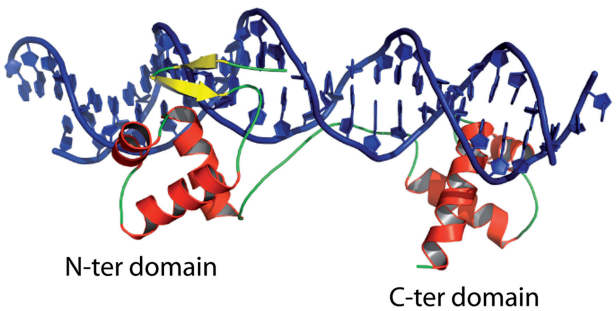

Figure 2. Structure of the Pax6 paired domain (PDB id 6PAX) (43). Cartoon representation showing both N- and C-terminal domains. The figure was done with the molecular visualization software Pymol (57).

capability of the Paired domain to bind DNA and on its intrinsic stability.

Out of the 50 mutations, FoldX is capable to assign an energetic explanation for the relation to disease for 44 of them. Among those, 37 involve significant changes in stability, 6 have their binding with DNA seriously perturbed and 9 may have their specificity towards their DNA-binding site changed (see Figure 5 and Supplementary Table S3). These features are not mutually exclusive, and a destabilizing mutant could also affect DNA binding or specificity (e.g. N9I mutation, see the predicted changes in specificity in Supplementary Figure S7).

These results can be compared to what we obtained by modeling a set of neutral mutations (Supplementary Figure S3 and Supplementary Table S4) made of close orthologs. Only for two (*Drosophila melanogaster* and *Xenopus laevis*) out of nine organisms, the mutations introduce a slightly significant destabilization. In the case of *Xenopus*, it is only slightly above our threshold and for *Drosophila*, the main destabilization comes from the G110 N mutation, which in the 6PAX structure presents a backbone conformation that is only tolerated by Gly. This residue is at the beginning of a long loop that we speculate changes its conformation locally upon mutation.

Only three of these 50 mutations have been analyzed *in vitro* and *in vivo* (45,46). The P68S mutant was analyzed both by a chloramphenicol acetyltransferase assay and by analyzing its ability to repress PAX2 expression (46). Its capacity to bind to the consensus promoter was impaired and the PAX2 repression activity was also significantly decreased. We found that this mutation slightly destabilizes the structure but produces a change in the specificity. The activity of the I87 R and R26 G mutants were assessed with an *in vivo* transcriptional assay and both of them had lost the ability to activate the transcription of the reporter gene (45). However, while *in vitro* I87 R showed a dramatic 14-fold decrease of its binding capacity to the consensus sequence, R26 G showed only a 4-fold decrease. For I87R, we predicted a strong destabilization, confirming the conclusion of the authors who hypothesized a change in the general conformation of the paired domain leading to a loss of activity, while for R26G we predict a smaller effect ($\Delta\Delta G_{int}$ around 1 kcal/mol).

Only for six mutations, R44Q, Q47R, V53L, K55R, E93K and R128C, the predicted changes in stability and binding are within the error of FoldX and the change of specificity is not significant. For the V53L, K55R and R128C mutations, our method does not predict any effect that could be linked to disease and carefully looking at its potential effects in the structure does not give any insight. We can hypothesize that even if they were found in patients suffering from eye disease, these mutations might not be direct cause and other factors not described may be at play. Finally, residues R44, Q47 and E93 are on the surface of the protein and could be involved in interactions with other proteins, such as SOX2 (47). Indeed, in the case of Q47R, most members of the Pax family contain Arg47 while Pax6 has Gln47. We suggest that this mutation will not affect overall stability, as an Arg at that position is structurally and evolutionarily accepted. In the case of E93K, the Pax family position 93 is quite variable, accepting Asp, Arg, Gln, Glu, Gly and Ser, but not Lys. E93 may interact with the HD domain (48) and its mutation to Lys affects this interaction, or more likely is a miss assignment and the mutation responsible for the disease is in another gene.

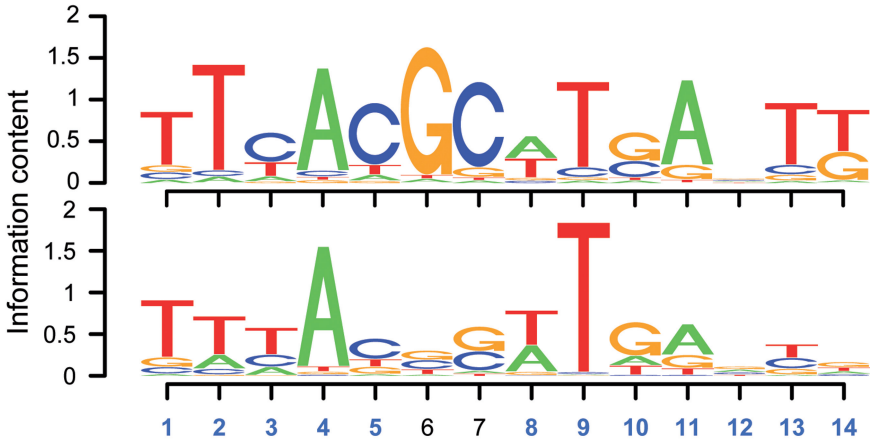

Figure 3. Comparison of the experimental logo (44) (top panel) and the predicted logo for the wild-type Pax6 (bottom panel).

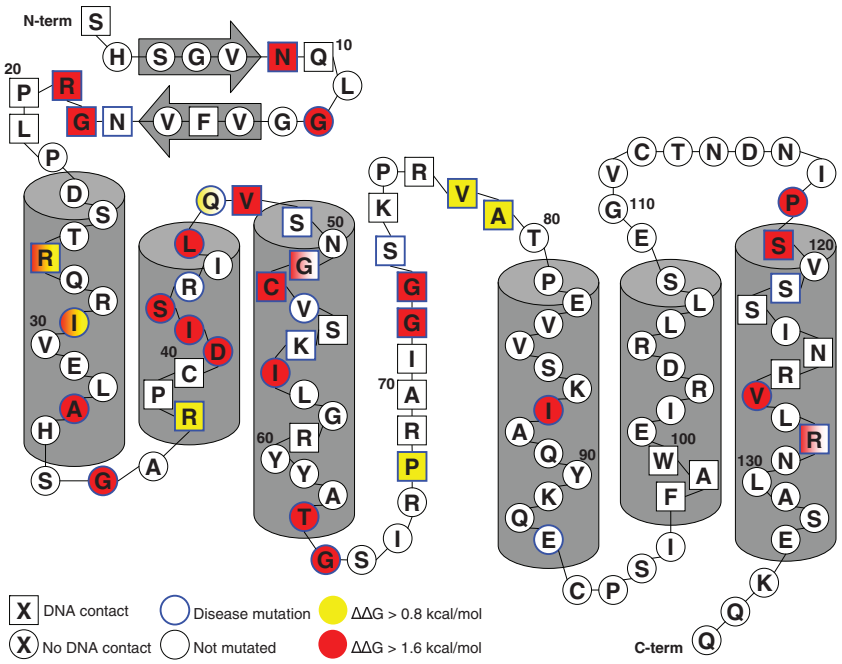

Figure 4. Schematic view of the distribution of Pax6 mutations and the energy results. The residues are color coded according to its change in stability. Residues with two colors represent the results for different mutations in the same position. Residue numbering through out the article is based on the Uniprot numbering (Isoform 1) that is three positions shifted from the PDB one.

DISCUSSION

We have shown that using a self-consistent force field, not specifically trained for protein–DNA interactions, it is possible to accurately predict protein–DNA binding specificity using a crystal structure. Both speed and accuracy can be improved at the expense of one another, but when applying FoldX into a high throughput analysis, a balance can be made between them. To benchmark our predictions, we contrasted our results with sets of protein or DNA mutations, achieving a good correlation with the experimental results. We also compared them to experimentally determined DNA-binding specificity motifs. From a total of 25 proteins evaluated (23 existing in public databases plus two new ones reported here), we accurately predict a third of them ($D < 0.38$), and reasonably well another 65% ($D < 0.58$). The lack of any specific protein–DNA contacts at some DNA positions is the source of most of the failed predictions. Incorrect predictions of specificity in positions where there is no contact may be the result of intrinsic local conformation of the DNA, emphasizing the importance of local backbone moves. Adding DNA and protein backbone flexibility, especially at the edges of the binding site, should improve the prediction. Also problems in the crystal structure, like low resolution or local artifacts, may be affecting our predictions.

We showed that FoldX can be used as a tool to predict DNA-binding specificities if there is a structure of the

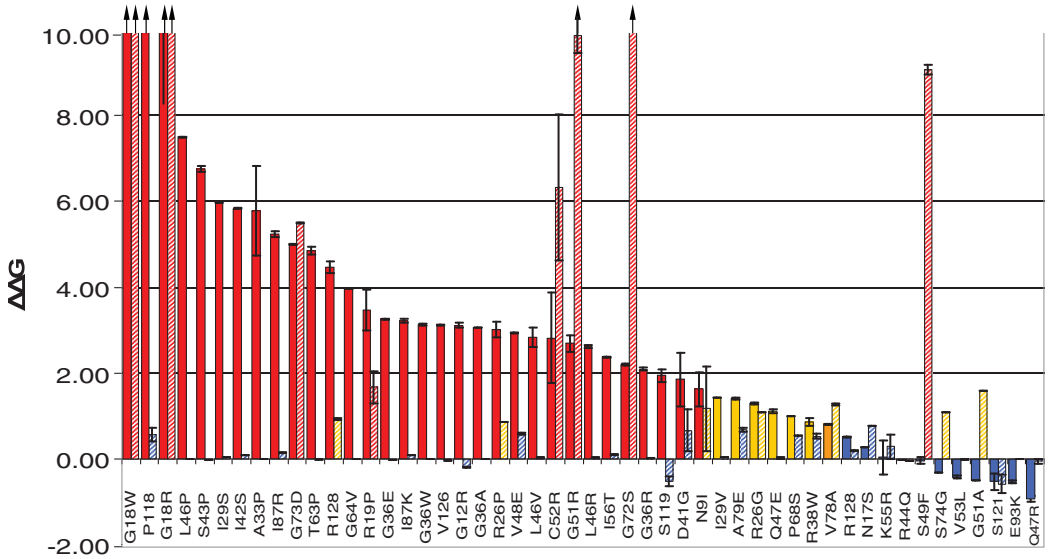

Figure 5. Energy changes upon mutation. Red for changes >1.6 kcal/mol, orange for changes between 1.6 and 0.8 kcal/mol and blue for those whose effect is <0.8 kcal/mol. The changes in stability are displayed as solid colors and the changes in interaction energy as hash-bars. Values >10 kcal/mol are shown with an arrow. Mutations are sorted according to their stability changes. Values can be found in Supplementary Table S3.

protein–DNA complex available, as well as for designing changes in specificity (31,32). In this work, we have only used crystal structures, but structures coming from NMR or homology modeling with high percentage identity (especially in the interface with DNA) could also be used. However, it has been shown (31) that structures coming from high percentage identity DNA-binding proteins can have important differences in the DNA conformation, making predictions based on models of even very similar proteins more unreliable than in the case of those based on crystal structures.

FoldX in combination with a protein–DNA structure could provide the user with a PWM to be used in combination with ChIP-on-chip experiments, to find the binding sites within the ~1000-bp sequences resulting from this technique. Given the (small) loss of information content reported, which is related to the predicted specificity, it would be advisable to rely only on the top hits found. With over 1500 protein–DNA complexes currently in the PDB (a number that steadily increases by 15–20% each year) and only a few hundred PWMs available from the literature, structure-based prediction is a promising tool for filling the gap between structural and genomic fields.

Using this validated protein design tool, we have analyzed the effects of mutations found in humans on the PD of Pax6. Pax6 is part of a complex system as it has been described as the master gene in eye development, but it also has an important role in the development of the central nervous system (49). With the Pax6 example, we have seen that we are now capable of finding structural and energetic explanations for the majority of the reported missense mutations in its PD (44 out of 50 mutations). We observed that most of the Pax6 mutations in the paired domain leading to ocular developmental diseases are driven by a loss of stability of the domain, which is coherent with the early concept that aniridia was due to the haploinsufficiency. Indeed, a destabilization of the domain could lead to its spatial reorganization and a total loss of binding as a truncated protein would do. It has been described (50) that in the complete Human PAX6 Mutation Database, most of the non-aniridia phenotypes were caused by missense mutations. However, this is non-reciprocal and most of the observed missense mutations still lead to aniridia-related phenotypes. As for the general mutations, we did not find here any correlation between the energetic and structural effects of the missense mutations and their phenotypic outcome. The only part worth notifying is that all but one of the foveal hypoplasia cases here included are caused by mutations that do not affect the general stability of the paired domain, but rather the binding affinity and/or specificity. This finding may indicate, as already suggested (5,51), that the number of missense mutations in PAX6 is greatly underestimated, as their effects could be more subtle. As Pax6 in general and its PD in particular are highly conserved in chordates, any mutation should give rise to a phenotypical trait different from the wild-type, but only relatively few mutations have been reported. Indeed if most of the detected mutations abolish the activity, either by unfolding the domain or by making its interface improper to DNA binding, leading to classical aniridia, it is reasonable to think that some of the possible missense mutations would have mild effects and rather change the specificity of binding. Therefore, the expressed phenotype would differ greatly from classical aniridia, and as most of the PAX6 analyses have been done on aniridia patients, this data is missed. A broader look at the pathologies that may be caused by Pax6 mutations could increase the knowledge about this TF if more biochemical data is analyzed.

To further assess the disease effects of TF in general, and Pax6 in particular, it would be interesting to do transcriptomics analyses not only on null mutants as has been done up to now (52–54), but also on point mutations that have proven to give rise to developmental (or cancer) problems.

The approach shown here could be applied to other DNA-binding proteins, by assigning a potential effect to mutations detected *in utero* so that their potential involvement in a future disease could be known in advance. Moreover, analyzing precisely the molecular effects of such single point mutations in the interface with DNA represents a way to differentiate between neutral or functionally important non-synonymous SNPs in DNA-binding proteins. As such, this is complementary of the more traditional genomic approach. Those diseases caused by mutations that destabilize the protein could be treated using specifically designed small molecules that would help stabilize it back (55,56), while those that are caused by a loss of binding or a change in specificity would need gene therapy to be cured. As such, knowledge derived from structure-based protein design analysis could be a key factor to fully develop personalized medicine.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Raik Grünberg, Almer van der Sloot, Trevor Siggers and Sheldon Rowan for helpful discussions, and Justine Leigh-Bell for helping correct the language in the manuscript.

FUNDING

CRG-Novartis fellowship (to A.A.); EU project EMERGENCE (NEST – 043338 to L.S.); EU project ZNIP (LSHB-CT-2006-037783 to L.S.); EU project PROSPECTS (HEALTH-F4-2008-201648 to L.S.); National Institutes of Health/National Human Genome Research Institute (R01 HG003985 to F.D.M. and M.L.B.). Funding for open access charge: The EU project PROSPECTS and the Spanish CONSOLIDER – INGENIO project Centrosome 3D.

Conflict of interest statement. None declared.

REFERENCES

1. Chi,N. and Epstein,J.A. (2002) Getting your Pax straight: Pax proteins in development and disease. *Trends Genet*., **18**, 41–47.
2. Marquardt,T., Ashery-Padan,R., Andrejewski,N., Scardigli,R., Guillemot,F. and Gruss,P. (2001) Pax6 is required for the multipotent state of retinal progenitor cells. *Cell*, **105**, 43–55.
3. Tsonis,P.A. and Fuentes,E.J. (2006) Focus on molecules: Pax-6, the eye master. *Exp. Eye Res*., **83**, 233–234.
4. van Heyningen,V. and Williamson,K.A. (2002) PAX6 in sensory development. *Hum. Mol. Genet*., **11**, 1161–1167.
5. Kokotas,H. and Petersen,M.B. (2010) Clinical and molecular aspects of aniridia. *Clin. Genet*., **77**, 409–420.
6. Brown,A., McKie,M., van Heyningen,V. and Prosser,J. (1998) The Human PAX6 Mutation Database. *Nucleic Acids Res*., **26**, 259–264.
7. Pey,A.L., Stricher,F., Serrano,L. and Martinez,A. (2007) Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. *Am. J. Hum. Genet*., **81**, 1006–1024.
8. Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol*., **22**, 1302–1306.
9. Maurer-Stroh,S., Debulpaep,M., Kuemmerer,N., de la Paz,M.L., Martins,I.C., Reumers,J., Morris,K.L., Copland,A., Serpell,L., Serrano,L. *et al*. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
10. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pena-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al*. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
11. Noyes,M.B., Christensen,R.G., Wakabayashi,A., Stormo,G.D., Brodsky,M.H. and Wolfe,S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.
12. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
13. Liu,Z., Guo,J.T., Li,T. and Xu,Y. (2008) Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. *Proteins*, **72**, 1114–1124.
14. Persikov,A.V., Osada,R. and Singh,M. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, **25**, 22–29.
15. Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc*., **4**, 393–411.
16. Grove,C.A., De Masi,F., Barrasa,M.I., Newburger,D.E., Alkema,M.J., Bulyk,M.L. and Walhout,A.J. (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell*, **138**, 314–327.
17. Tompa,M., Li,N., Bailey,T., Church,G., De Moor,B., Eskin,E., Favorov,A., Frith,M., Fu,Y., Kent,J. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol*., **23**, 137–144.
18. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol*., **5**, 201.
19. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
20. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
21. Meng,X., Brodsky,M.H. and Wolfe,S.A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol*., **23**, 988–994.
22. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet*., **36**, 1331–1339.
23. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol*., **24**, 1429–1435.
24. Morozov,A., Havranek,J., Baker,D. and Siggia,E. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res*., **33**, 5781–5798.
25. Havranek,J.J., Duarte,C.M. and Baker,D. (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol*., **344**, 59–70.
26. Endres,R.G. and Wingreen,N.S. (2006) Weight matrices for protein-DNA binding sites from a single co-crystal structure. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., **73**, 061921.

27. Jamal Rahi,S., Virnau,P., Mirny,L.A. and Kardar,M. (2008) Predicting transcription factor specificity with all-atom models. *Nucleic Acids Res.*, **36**, 6209–6217.
28. Paillard,G., Deremble,C. and Lavery,R. (2004) Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.*, **32**, 6673–6682.
29. Angarica,V.E., Perez,A.G., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, **9**, 436.
30. Arnould,S., Chames,P., Perez,C., Lacroix,E., Duclert,A., Epinat,J.C., Stricher,F., Petit,A.S., Patin,A., Guillier,S. *et al.* (2006) Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *J. Mol. Biol.*, **355**, 443–458.
31. Redondo,P., Prieto,J., Munoz,I.G., Alibés,A., Stricher,F., Serrano,L., Cabaniols,J.P., Daboussi,F., Arnould,S., Perez,C. *et al.* (2008) Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature*, **456**, 107–111.
32. Marcaida,M.J., Prieto,J., Redondo,P., Nadra,A.D., Alibés,A., Serrano,L., Grizot,S., Duchateau,P., Paques,F., Blanco,F.J. *et al.* (2008) Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl Acad. Sci. USA*, **105**, 16888–16893.
33. Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. *Proc. Natl Acad. Sci. USA*, **104**, 7068–7073.
34. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
35. Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.
36. Kumar,M.D., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–206.
37. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
38. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
39. Matys,V., Fricke,E., Geffers,R., Gößling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
40. Newburger,D.E. and Bulyk,M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **37**, D77–D82.
41. Bembom, O. (2007) seqLogo: An R package for plotting DNA sequence logos. http://bioconductor.org/packages/2.6/bioc/html/seqLogo.html.
42. Doi,N., Takashima,H., Kinjo,M., Sakata,K., Kawahashi,Y., Oishi,Y., Oyama,R., Miyamoto-Sato,E., Sawasaki,T., Endo,Y. *et al.* (2002) Novel fluorescence labeling and high-throughput assay technologies for in vitro analysis of protein interactions. *Genome Res.*, **12**, 487–492.
43. Xu,H.E., Rould,M.A., Xu,W., Epstein,J.A., Maas,R.L. and Pabo,C.O. (1999) Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. *Genes Dev.*, **13**, 1263–1275.
44. Epstein,J., Cai,J., Glaser,T., Jepeal,L. and Maas,R. (1994) Identification of a Pax paired domain recognition sequence and evidence for DNA-dependent conformational changes. *J. Biol. Chem.*, **269**, 8355–8361.
45. Tang,H.K., Chao,L.Y. and Saunders,G.F. (1997) Functional analysis of paired box missense mutations in the PAX6 gene. *Hum. Mol. Genet.*, **6**, 381–386.
46. Azuma,N., Yamaguchi,Y., Handa,H., Tadokoro,K., Asaka,A., Kawase,E. and Yamada,M. (2003) Mutations of the PAX6 gene detected in patients with a variety of optic-nerve malformations. *Am. J. Hum. Genet.*, **72**, 1565–1570.
47. Kamachi,Y., Uchikawa,M., Tanouchi,A., Sekido,R. and Kondoh,H. (2001) Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes Dev.*, **15**, 1272–1286.
48. Bruun,J.A., Thomassen,E.I., Kristiansen,K., Tylden,G., Holm,T., Mikkola,I., Bjorkoy,G. and Johansen,T. (2005) The third helix of the homeodomain of paired class homeodomain proteins acts as a recognition helix both for DNA and protein interactions. *Nucleic Acids Res.*, **33**, 2661–2675.
49. Osumi,N., Shinohara,H., Numayama-Tsuruta,K. and Maekawa,M. (2008) Concise review: Pax6 transcription factor contributes to both embryonic and adult neurogenesis as a multifunctional regulator. *Stem Cells*, **26**, 1663–1672.
50. Tzoulaki,I., White,I.M. and Hanson,I.M. (2005) PAX6 mutations: genotype-phenotype correlations. *BMC Genet.*, **6**, 27.
51. Hanson,I., Churchill,A., Love,J., Axton,R., Moore,T., Clarke,M., Meire,F. and van Heyningen,V. (1999) Missense mutations in the most ancient residues of the PAX6 paired domain underlie a spectrum of human congenital eye malformations. *Hum. Mol. Genet.*, **8**, 165–172.
52. Visel,A., Carson,J., Oldekamp,J., Warnecke,M., Jakubcakova,V., Zhou,X., Shaw,C.A., Alvarez-Bolado,G. and Eichele,G. (2007) Regulatory pathway analysis by high-throughput in situ hybridization. *PLoS Genet.*, **3**, 1867–1883.
53. Wolf,L.V., Yang,Y., Wang,J., Xie,Q., Braunger,B., Tamm,E.R., Zavadil,J. and Cvekl,A. (2009) Identification of pax6-dependent gene regulatory networks in the mouse lens. *PLoS ONE*, **4**, e4159.
54. Holm,P.C., Mader,M.T., Haubst,N., Wizenmann,A., Sigvardsson,M. and Gotz,M. (2007) Loss- and gain-of-function analyses reveal targets of Pax6 in the developing mouse telencephalon. *Mol. Cell Neurosci.*, **34**, 99–119.
55. Basse,N., Kaar,J.L., Settanni,G., Joerger,A.C., Rutherford,T.J. and Fersht,A.R. (2010) Toward the rational design of p53-stabilizing drugs: probing the surface of the oncogenic Y220 C mutant. *Chem. Biol.*, **17**, 46–56.
56. Foss,T.R., Kelker,M.S., Wiseman,R.L., Wilson,I.A. and Kelly,J.W. (2005) Kinetic stabilization of the native state by protein engineering: implications for inhibition of transthyretin amyloidogenesis. *J. Mol. Biol.*, **347**, 841–854.
57. DeLano W.L. (2002) The PyMOL molecular graphics system. Schrödinger, LLC. http://www.pymol.org.