SECTION ONE

MATHEMATICAL AND COMPUTATIONAL METHODS
CHAPTER ONE

DNA-BINDING SPECIFICITY PREDICTION WITH FOLDX

Alejandro D. Nadra,* Luis Serrano,†,‡ and Andreu Alibés†

Contents

1. Introduction 4
2. Description of the Force Field, FoldX, and the Implementation of DNA Energy Terms and Base Mutation 5
   2.1. Force field 6
   2.2. DNA parameterization 6
   2.3. Detection of DNA base pairs 6
   2.4. Distance constraints to automatically identify base pairs 8
   2.5. DNA stacking energy 8
   2.6. DNA mutation 8
3. Predicting Specificities 9
   3.1. Prediction of DNA-binding profiles using FoldX 9
4. Changing Binding Capabilities 12
5. Designing New Specificities 13
6. Known Caveats 14
   6.1. Resolution 15
   6.2. Water molecules 15
   6.3. DNA flexibility and base independence 15
   6.4. Clashes 15
7. Conclusions and Future Outlook 16
Acknowledgments 16
References 17

Abstract
With the advent of Synthetic Biology, a field between basic science and applied engineering, new computational tools are needed to help scientists reach their goal, their design, optimizing resources. In this chapter, we present a simple and powerful method to either know the DNA specificity of a wild-type protein...
or design new specificities by using the protein design algorithm FoldX. The only basic requirement is having a good resolution structure of the complex. Protein–DNA interaction design may aid the development of new parts designed to be orthogonal, decoupled, and precise in its target. Further, it could help to fine-tune the systems in terms of specificity, discrimination, and binding constants. In the age of newly developed devices and invented systems, computer-aided engineering promises to be an invaluable tool.

1. Introduction

One of the goals of Synthetic Biology is the design and construction of new biological parts, devices, and systems. Most biological parts have been obtained from nature with slight if any modification, being an exception some experimentally selected mutants. This is particularly true in the case of DNA-binding proteins probably due to the lack of engineering tools to design protein–DNA interfaces until very recently (Alibe’s et al., 2010a,b; Morozov et al., 2005). In the past few years, there have also been successful attempts to predict binding specificity from structure, either by using existing crystal structures (Angarica et al., 2008; Arnould et al., 2006; Endres and Wingreen, 2006; Havranek et al., 2004; Jamal Rahi et al., 2008; Marcaida et al., 2008; Morozov et al., 2005; Paillard et al., 2004; Redondo et al., 2008) or by using a docking approach (Liu et al., 2008). These predictions were evaluated in zinc fingers (Benos et al., 2002; Paillard et al., 2004) where a sensitivity to docking geometry was reported (Siggers and Honig, 2007), and in meganucleases (Arnould et al., 2006; Marcaida et al., 2008; Redondo et al., 2008), highlighting the importance of having multiple templates to enhance accuracy.

Protein–DNA interactions are key to the regulation of cellular activities such as transcriptional regulation and replication. The specificity displayed by a DNA-binding protein toward its DNA target sites is an essential feature for the function of any transcription factor in a given genome where an excess of nonspecific sites exists (Rohs et al., 2010). For Synthetic Biology, binding site should be differentiated from binding profile/motif being the former, an extreme simplification of DNA-binding capabilities. To gain a fine control of the system, a complete description of the behavior of a TF is desired. As an example, it could be desired to have a unique site in the complete genome. Alternatively, the requirement could be to have several but each with different affinities, or with overlapping binding sites, or even a collection of TF with graduated affinity/activity. Further, the goal for certain devices could be to have orthologous systems with the same affinity but varying their \( K_{on}/K_{off} \).

A working definition of Synthetic Biology (http://www.synthetic-biology.info) suggests that Synthetic Biology is the engineering of biological
components and systems that do not exist in nature and the reengineering of existing biological elements; it is determined on the intentional design of artificial biological systems, rather than on the understanding of natural biology. This is particularly challenging in biological systems, where each process may be heavily influenced by regulatory and nonregulatory interactions with the environment. In any case, to achieve this goal for DNA-binding proteins, a specific protein–DNA engineering tool is needed.

Protein–DNA interactions are complex phenomena as they involve direct and indirect interactions and there is not a general recognition code to predict base–residue interactions. Thus structural information is required in order to understand the specificity of any DNA-binding protein. Structure-based DNA-binding prediction is a powerful tool to infer protein–binding sites and design new specificities. This approach has developed in the past few years, and there exist some examples of successful binding site prediction for several proteins and particularly for zinc fingers. There is, however, only one example of structure-based design of a new specificity (Ashworth et al., 2006), underlining the state-of-the-art nature of the approach. These techniques usually require a lot of expertise and are difficult to reproduce by nonexperts. The aim of this chapter is to establish a simple protocol for structure-based prediction of proteins binding profiles. This protocol can be used to derive the binding motif of an existing protein and also to modify existing specificities. Designed to be simple, this protocol may be somehow limited, as we tried to make a balance between power and simplicity in the protocol. Some hints on how to fine-tune the predictions to make them more powerful are presented in Section 6.

De novo design is very straightforward with the newly developed ability to optimize and synthesize genes. In the next chapter, we present FoldX protein design algorithm as an example of a tool to evaluate protein–DNA binding specificities that may allow to have a tighter control of your favorite transcription factor or to design new DNA sequence specificities.

2. Description of the Force Field, FoldX, and the Implementation of DNA Energy Terms and Base Mutation

FoldX (Guerois et al., 2002; Schymkowitz et al., 2005; http://foldx.crg.es) is an application that provides a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins, protein–protein complexes, and protein–DNA complexes. The capability of FoldX to deal with protein–DNA complexes has been recently added and its predictions have been validated (Alibés et al., 2010a) and particularly applied to meganucleases (Arnould et al., 2006; Marcaida et al., 2008;
Redondo et al., 2008). Given a template structure, this method generates a PWM/motif displayed by the DNA-binding protein of interest, derived by an energetic criterion.

Its main feature in the framework of this chapter is its ability to mutate both protein and DNA and evaluate the effect of these mutations on the interaction energy and the stability of the complex. FoldX is freely available for academic users upon registration (http://foldx.crg.es).

2.1. Force field

The FoldX force field defines the following terms (Eq. (1.1)) to calculate the free energy:

\[
\Delta G = \Delta G_{vdw} + \Delta G_{solvH} + \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{kon} + T \Delta S_{mc} + T \Delta S_{sc} + \Delta G_{clash}
\]  

(1.1)

\(\Delta G_{vdw}\), the van der Waals term; \(\Delta G_{solvH}\) and \(\Delta G_{solvP}\), the interaction with the solvent, separated between the hydrophobic and polar groups; \(\Delta G_{wb}\), the water bonds term; \(\Delta G_{hbond}\), the term that takes into account hydrogen bonds; \(\Delta G_{el}\), the electrostatic contribution to free energy; \(\Delta G_{kon}\), the electrostatic contribution coming from the interaction of atoms of two different molecules; \(\Delta S_{mc}\) and \(\Delta S_{sc}\), the entropy cost of fixing the main chain or the side chain to a particular conformation; \(\Delta G_{clash}\), the term that takes into account the steric overlaps between atoms in the structure; and finally \(T\), the temperature (Schymkowitz et al., 2005).

2.2. DNA parameterization

The four bases, A, C, G, and T as well as the methylated A and C, were added using standard FoldX amino acid atoms as reference atoms. For each atom of each nucleic acid, we used the standard parameters (van der Waals radii, volumes, solvation energies, hydrogen bond energies, and angles) of an amino acid atom that was closest in nature (type of atom, hybridization). For example, the polar atom N6 of the adenine base which can make two hydrogen bonds use the parameters of the atom ND2 of asparagine, whereas the phosphate and the two oxygen atoms O1P and O2P are taken from the parameters of the phosphorylated serine (Table 1.1).

2.3. Detection of DNA base pairs

To evaluate the preferred base at each position of the binding site, we need to calculate the energy for the four possible nucleotides, and the first step when moving or mutating DNA bases is to take into account that most of
the bases are paired. Therefore, the problem is not just a one-body move but a two-body move. In principle, one could identify the complementary base by looking at the DNA sequence in the structure. However, not always are all bases in a structure paired and in some sequences this could lead to confusion. Thus, a design tool should be able to automatically identify those pairs in order to move (mutate) them at the same time. This is a simple task for the human eye, but not straightforward for a computational algorithm. The classical approach would be to consider two bases as being paired if they make hydrogen bonds (H bonds), but in a test, we found that, considering only high-resolution structures, 7.7% of the base pairs were misassigned. This is due to DNA distortion/opening that may occur upon protein binding and may have two main consequences: (i) bases are too far away from each other to be able to make an H bond; (ii) bases are able to interact with more than one base and as a result, it is difficult to decide, based on hydrogen bonding, which is the correct one.

To solve this problem, we examined the set of structures above, analyzed all pairwise distances between atoms of facing DNA bases, and derived some simple distance constraints to automatically identify base pairs (see below).

Table 1.1 One to one correspondence between the atom of adenine and the reference atoms chosen among the amino acids to define their parameters

<table>
<thead>
<tr>
<th>DNA atom</th>
<th>Amino acid</th>
<th>Reference atom</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>SEP</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>O1P</td>
<td>SEP</td>
<td>O1P</td>
<td></td>
</tr>
<tr>
<td>O2P</td>
<td>SEP</td>
<td>O2P</td>
<td></td>
</tr>
<tr>
<td>O5*</td>
<td>HIS</td>
<td>O</td>
<td>Except for H bonds (none possible)</td>
</tr>
<tr>
<td>C5*</td>
<td>GLY</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>C4*</td>
<td>PRO</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>O4*</td>
<td>HIS</td>
<td>O</td>
<td>Except for H bonds (none possible)</td>
</tr>
<tr>
<td>C3*</td>
<td>PRO</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>O3*</td>
<td>HIS</td>
<td>O</td>
<td>Except for H bonds (none possible)</td>
</tr>
<tr>
<td>C2*</td>
<td>PRO</td>
<td>CG</td>
<td></td>
</tr>
<tr>
<td>C1*</td>
<td>PRO</td>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>N9</td>
<td>HIS</td>
<td>N</td>
<td>Except for H bonds (none possible)</td>
</tr>
<tr>
<td>C4</td>
<td>TRP</td>
<td>CE2</td>
<td></td>
</tr>
<tr>
<td>N3</td>
<td>HIS</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>TYR</td>
<td>CD2</td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>HIS</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>TYR</td>
<td>CZ</td>
<td></td>
</tr>
<tr>
<td>N6</td>
<td>ASN</td>
<td>ND2</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>TRP</td>
<td>CE2</td>
<td></td>
</tr>
<tr>
<td>N7</td>
<td>HIS</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td>TYR</td>
<td>CD2</td>
<td></td>
</tr>
</tbody>
</table>
We then tested it against structures not included in the previous set and containing double-stranded DNA in complex with a protein, and structures of double-stranded DNA alone. Of all the base pairs in the protein–DNA set, all were assigned correctly by using the distance constraints, while 5.4% would be misassigned using a hydrogen bonding criteria only. This ratio dropped to 3.6% for the set composed of double-stranded DNA alone because of the higher regularity of the considered structures. In all cases using our distance constraint method, we could correctly identify the corresponding pairing base.

2.4. Distance constraints to automatically identify base pairs

We define the distance between each C1′ atom (ribose atom to which the base is attached), DistC1, and the distance between the atom N1 of the purines and the atom N3 of the pyrimidines, DistN. Their ratio Div = DistC1/DistN is the determining factor for the base pair recognition. To be considered as base pairs, all bases must fulfill the following rules (average values in the dataset were considered for determining the parameters): Div is smaller than 5.0 and strictly either larger than 2.9 or larger than 2.5 if DistC1 is between 10.25 and 11.3 Å. If there was any ambiguity between two possible pairs, we chose the one presenting stronger H bonds. For the few cases that did not fit the criteria, we made a second pass with the following rules: bases which were not paired before and for whom Div is between 2.1 and 5.0 and DistC1 is between 9 and 13 Å were considered as base pairs.

2.5. DNA stacking energy

To 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 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 (ΔG = −RT ln(p)).

2.6. DNA mutation

To mutate DNA, we replace a base in the crystal by its mutant, superposing the two bases on the N1 atom and the C1′–N1–C2 plane for the pyrimidines and the N9 atom and C1′–N9–C4 plane for the purines. Based on a statistical analysis of the available DNA structures in the PDB and in order to simulate movements of the bases relative to the ribose, we allow two degrees of freedom: the angle C1′–N9–C4 (125° ± 5°) and the dihedral
angle O4’–C1’–N9–C4 (±5° around the original dihedral angle form the structure) for the purines and the angle C1’–N1–C2 (118° ± 5°) and the dihedral angle O4’–C1’–N1–C2 (±5° around the original dihedral angle form the structure) for the pyrimidines. We replace in one step both the desired base and its base pair (when present). The side chains of the surrounding residues and bases were then moved sequentially to adapt to the new mutated bases.

3. Predicting Specificities

The binding profile of a transcription factor can be known using in silico methods and the structure of its complex with DNA (Alibés et al., 2010a). The general strategy is to generate mutations in the DNA and evaluate the energies for each mutant toward the protein. This way, we evaluate the binding energy as well as the discrimination for related DNA sequences. The first step is to find a template structure. Then, the DNA is mutated at each position to all four bases from which the interaction energy will be evaluated and specificity calculated.

3.1. Prediction of DNA-binding profiles using FoldX

The only requirement is to have a good quality structure (X-ray cocrystal up to 2.2 Å resolution are preferred over NMR structures) for the desired protein in complex with DNA. Structures may be obtained from the Protein Data Bank (Berman et al., 2000). Then, DNA-binding-site motifs may be derived for this protein using FoldX. To do so, one should mutate every base in the structure to the other three in an exhaustive manner and determine the predicted binding energy between protein and mutant DNA. Additionally, it is useful to consider the van der Waals intramolecular clashes of the DNA molecule that could appear upon base mutation. This extra term penalizes those DNA variants that may have a good binding energy, but are forced into the DNA structure. Considering these energy terms, one can derive a DNA sequence profile by doing a partition function for each position. This procedure requires the following steps (Fig. 1.1):

1. The first step is to look for the best template in a database. Look at the RCSB-PDB Web site for crystal structures of the candidate bound to DNA (structures solved by NMR can also be used but usually crystal structures give better results). As the prediction results can vary depending on the template, only high-resolution structures should be considered. Using structures with resolutions better than 2.2 Å is a reasonable cutoff. If the structure had poorly determined regions, they should not occur in the protein–DNA interface.
2. PDB structures contain a lot of information and, usually, more structures than those needed. Delete everything but the molecules you want to model (i.e., DNA strands, protein chain, metal ions, and water molecules in the interface). This can be done with any text editor or a structure visualization program, such as VMD, PyMol, or Swiss-PdbViewer.

3. The next step requires preparing the reference structure for the simulations. To optimize the crystal structure, use the FoldX RepairPDB function to minimize the energy according to the FoldX force field, removing small van der Waals clashes; generating missing side chains; and choosing the correct histidine, glutamine, and asparagine rotamers. After running the RepairPDB function, check the generated PDB file for the removal of relevant unrecognized atoms. This function creates a new file called “RepairPDB_,” followed by the name of the original structure file. This file will be used in the next step.

Crystal structures are not perfect and sometimes interacting atoms are placed too close (van der Waals clashes) or too separate (missing a hydrogen bond). As a result, FoldX can move out a side chain that makes specific contacts with a base, or another amino acid. The user

Figure 1.1 Scheme of the procedures described. In italics, FoldX function used in each step.
should check the resulting structure after repair to see if this is the case. There is an option in FoldX that will fix those problematic side chains if needed. As an example, we could mention the case of GCN4 crystal (PDB: 1YSA; Fig. 1.2). When FoldX repairs this crystal, arginine 243 side chain moves outward and does not make a specific contact with the bases. The distance involved is 3.2 Å, from the NH1 from the arginine to the C2 from the adenine in front, which results in van der Waals clashes and consequently Arg243 is flipped out. This behavior is not frequent but the user should be aware of it. When such a case is detected, the user should “fix” the residue, to force the side chain to remain contacting the ligand (available for the RepairPDB and BuildModel commands).

4. Using the optimized structure obtained in the previous step, mutate the DNA sequence in the template structure to all possible bases at each position. The number of mutants will be four times the length (avoid mutating bases too far away that will not contact the protein, i.e., distances to protein larger than 6 Å).

This approximation requires considering each protein–DNA base pair contact as independent and additive. In general, this could be the case, but when having a densely packed area, it is quite probable that context effects due to the neighboring bases could be important. If the user suspects this is the case, she can proceed by mutating the base pairs as triplets centered in the target base. So, instead of considering four variants for
each position, we should analyze 64 base combinations. This kind of analysis may be more relevant when it is known that the flexibility and/or significant DNA deformation is important either in intermediates or in the final complex. However, the user should be aware that it will significantly increase the computation time.

5. Once the DNA variants have been generated, we have to evaluate the changes in interaction energy with respect to the original structure. When mutating residues or bases, some variants that give good interaction energies but introduce internal van der Waals clashes (either with the DNA or with the protein) can be wrongly selected as a good option. To take this into account, add the intraclashes energy when the mutant has higher values than the reference (optimized structure) to the difference in interaction energy (note that it should never be subtracted the intraclashes value if the mutant displays a lower value than the crystal as it is reporting a crystallographic problem and not a gain in interaction energy). As a rule of the thumb, changes in intraclashes smaller than 0.6 kcal/mol could be omitted (0.6 kcal/mol is the standard FoldX error). To avoid disrupting the interaction with the mutants introduced, the interaction energy should not be much higher than that of the crystal structure. In case you have fixed a side chain because it has bad contacts in the crystal structure, you should not consider the changes in van der Waals clashes in the overall interaction energy upon mutating the interacting base unless they become worse than the reference.

6. The changes in interaction energy corresponding to nucleotides in each position are converted to probabilities proportional to \( \exp(-\Delta G_{\text{int}}/RT) \) and then, using the seqLogo package from R Bioconductor, graphically represented as logos (see examples in Fig. 1.3).

### 4. Changing Binding Capabilities

Another important issue to consider is that DNA-binding capabilities could be changed quantitatively, qualitatively, or both. This means that it could be changed by the interaction energy without affecting significantly sequence discrimination. This can be done by varying electrostatics interactions between the protein and DNA backbone, without affecting the side chain–base contacts that dictate sequence specificity. However, one could desire to modify base discrimination but keeping relatively invariant the interaction strength. This task is quite more challenging and involve introducing very carefully specific contacts that either favors (or disfavors) one particular base or its size (to specify a purine or pyrimidine in a particular position). Combination of both preceding aspects can be made to completely change a binding site, slightly modify it, extend it, restrict it, introduce variable bases in the middle of a conserved binding region, etc.
If we want to modify the binding profile of a transcription factor, we can do, before doing the experimental work, a first screen of the possible mutants \textit{in silico}. To do so:

1. Select which residues are close to the DNA positions where the specificity needs to be changed.
2. Mutate those residues to alanine and then use the alanine mutant structure to scan all possible DNA combinations at the selected DNA positions as done above.

\textit{Figure 1.3} Comparison between the experimental logos (top) and the FoldX predicted ones (bottom) for Usf1 and Gcn4. Usf1 experimental logo from Jaspar (Sandelin \textit{et al.}, 2004); Gcn4 experimental logo from UniPROBE (Newburger and Bulyk, 2009). (See Color Insert.)
3. Remove all DNA combinations that have bad internal energies due to van der Waals clashes (incompatibility of the DNA sequence with the DNA structure).

4. Mutate the alanine in each of the structures with its specific combination of DNA bases to the selected new residues (remember you can explore as many protein positions and amino acids as you want, but this comes with a computational cost). As a rule of the thumb, more than three positions with 20 amino acids each will make the calculation too long. It is useful in many cases to learn from nature. Try to find all proteins related to the one you are mutating and see if there is enough sequence variability at the positions you desire. If this is the case, then use those residues at each position and explore all combinations. If this is not possible, first mutate each protein position to the 20 amino acids and select the residues that will have favorable interaction energies with your DNA template, and then do all combinations. When mutating to residues with larger side chains, it is necessary to repeat the mutations several times (3–5). It is accomplished with the option \(<\text{NumberOfRuns}\>5\). Then, for a statistical analysis, the average value should be taken, while for a particular design, the mutant with the best value should be taken. The reason being that the algorithm could be trapped in a minimum due to the starting residue and rotamer in the mutagenesis procedure. By using the multiple run option, it will start with a different position and rotamer each time, thus increasing the chance of finding the real energy optimum.

The alanine mutation step is very important and it should be done before mutating the DNA bases, so the newly introduced ones can move more freely and adopt a favorable conformation that may be sterically blocked by the wild-type residue. By deleting the contact between the residue and the nucleotide, alternative bases can accommodate to the best conformation according to the DNA structure only.

5. Finally, the change in interaction energy with respect to the original wild-type structure should be calculated. We have to consider as possible candidates only those where the change of interaction energy between the wild-type complex and the mutant is not too large. Failure to consider this factor could yield nonrealistic logos.

6. **Known Caveats**

The procedures and examples described above rely on having “the best case scenario”: a high-resolution structure, very stable and rigid upon mutation, whose binding sites are not affected by the chromatin structure. However, that is not the usual case and several factors have to be taken into account.
6.1. Resolution

Preferably, high-resolution structures should be used (resolution < 2.2 Å), and low-resolution structures may have nonsolved side chains that limit the prediction power of the procedure described and errors in backbone conformation or side chain placement.

6.2. Water molecules

Proteins whose contact with DNA depends on water-mediated interactions are more challenging to predict its specificities accurately. Even if FoldX is capable of predicting water molecule positions, the energetics of water-mediated interactions are difficult to predict.

6.3. DNA flexibility and base independence

For complexes where the intrinsic DNA structure or flexibility exerts an important role, nucleotide mutations could trigger important structural changes. FoldX is not capable of moving backbones upon mutation, and cannot take these cases into account. Further, in our procedure, we assume base independence, that is, that a nucleotide in one position does not change the probability of having another nucleotide in the neighboring positions. Although it is clearly not true in all cases, it is usually a good approximation. For those structures where changes in the DNA sequence could affect the neighboring bases, we recommend mutating the DNA in triplets. Also if there are several structures of the same protein, or a closely related protein with DNA, use all available templates; this will increase the number of possible hits when changing specificity, as well as the accuracy of the logo.

6.4. Clashes

It is frequent to find that the FoldX informs of high van der Waals clashes between bases and DNA backbone or residues in the wild-type structure. When this happen and introducing a mutation relieves considerably the clashes present in the wild type, it could be assumed that there was a crystallographic problem and the clashes should be dismissed.

Further, small increments in van der Waals clashes could be omitted if they are lower than 0.6 kcal/mol. Fine analysis of interaction energies require to analyze both, the gain in energy due to specific and nonspecific contacts and the repulsive forces due to close contacts. Indeed, this analysis could be done discriminating between the effect in each ligand (i.e., intraclashes) and the interface interclashes. It is quite common to introduce a mutation that enhances binding—thus, interaction energy—significantly but introduces high intraclashes that destabilize the protein, maybe, in a bigger extent.
7. Conclusions and Future Outlook

We have described a procedure where using a self-consistent force-field protein–DNA interaction specificity can be accurately predicted using a crystal structure. The protocol can also be used to change protein specificity and affinity. 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.

FoldX in combination with a protein–DNA structure can provide the user with a PWM (Position weight matrix) to be used for the purpose of scanning the genome or locus of interest for putative binding sites. Although we only discuss using X-ray structures, structures coming from NMR or homology modeling with high percentage identity (especially in the interface with DNA) could also be used. When dealing with a known flexible protein/interaction, incorrect prediction of specificity in positions where there is no contact may be the result of specific 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, could improve the prediction. In the near future, we plan to incorporate backbone moves to FoldX with the use of Brix (Baeten et al., 2008), a collection of protein fragments.

ACKNOWLEDGMENTS

We would like to thank François Stricher for the continuous development of FoldX during these past years and for insight on the characterization of DNA.

Appendix. Mutation Protocol

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. Then, each base was mutated to the other three possible bases five times to increase the conformational space analyzed. Each single point mutation takes less than 60 s using a single CPU (Intel Xeon 3.00 GHz, 8 Gb 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 function is graphically displayed in Fig. 1.3 as information content by means of the R package seqLogo (Bembom, 2007), 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 should be chosen.

In general, the same physical conditions may be used unless particularities of the system. Standard values may be: temperature of 298 K, pH of 7.0, and ion strength of 150 mM (in proteins from extremophiles, changing these parameters may be required).

REFERENCES


