Structural bioinformatics

Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains

Andrea Zen1, Cesira de Chiara2, Annalisa Pastore2 and Cristian Micheletti1,*

1SISSA, CNR-INFM Democritos and Italian Institute of Technology, Via Beirut 2, I-34151 Trieste, Italy and 2National Institute for Medical Research, MRC, The Ridgeway, London NW71AA, UK

Received on February 13, 2009; revised and accepted on May 27, 2009

ABSTRACT
Motivation: We have previously demonstrated that proteins may be aligned not only by sequence or structural homology, but also using their dynamical properties. Dynamics-based alignments are sensitive and powerful tools to compare even structurally dissimilar protein families. Here, we propose to use this method to predict protein regions involved in the binding of nucleic acids. We have used the OB-fold, a motif known to promote protein-nucleic acid interactions, to validate our approach.

Results: We have tested the method using this well-characterized nucleic acid binding family. Protein regions consensually involved in statistically significant dynamics-based alignments were found to correlate with nucleic acid binding regions. The validated scheme was next used as a tool to predict which regions of the AXH-domain representatives (a sub-family of the OB-fold for which no DNA/RNA complex is yet available) are putatively involved in binding nucleic acids. The method, therefore, is a promising general approach for predicting functional regions in protein families on the basis of comparative large-scale dynamics.

Availability: The software is available upon request from the authors, free of charge for academic users.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
The functionality of proteins and enzymes often relies on the capability of these biomolecules to sustain large-scale conformational changes (Frauenfelder et al., 1991). It has been established that these concerted functional movements are typically shared by members of enzymatic superfamilies which may otherwise differ significantly by fold, oligomeric state and even by the details of the catalytic chemistry (Capozzi et al., 2007; Carnevale et al., 2006). Quantitative algorithms are presently available to detect similar motions in protein pairs. The procedure is termed dynamics-based alignment because it allows the establishment of one-to-one correspondences between amino acids that experience similar large-scale movements in the two molecules (Zen et al., 2008). In a previous paper, we have shown that a dynamics-based alignment can result in a remarkable spatial superposition of functionally relevant regions even for structurally dissimilar families of proteins.

These results suggest that specific common concerted movements may have a functional rationale (Carnevale et al., 2006; Zen et al., 2008).

The goal of this study is to demonstrate this concept using as a model system the OB-fold, a well-characterized nucleic acid binding motif for which several structures are available in the PDB database in both their free and bound forms. Most commonly, the OB-fold consists of a closed barrel formed by two three-stranded antiparallel β-sheets. β3 is shared by both sheets, whilst β1 and β2 close the barrel partially or completely by forming a parallel network of hydrogen bonds (Murzin, 1993; Theobald et al., 2003). A relatively distant member of the OB family known to bind nucleic acids is formed by the AXH motif. So far, AXH motifs have been identified in two apparently unrelated human proteins of medical importance (Mushegian et al., 1997): the HMGB transcription factor HBP1 and the polyglutamine-containing ATX1 protein (Banfi et al., 1994; Lesage et al., 1994). Both proteins are thought to be transcription factors (Berasi et al., 2004, Tsai et al., 2004). HBP1, first identified as a target for family members of the retinoblastoma tumor suppressor (Lavender et al., 1997; Tevosian et al., 1997), is involved in cancer signalling pathways (Panhun et al., 2007). Mutations in ATX1 cause the spinocerebellar ataxia type-1 (SCA1), an autosomal-dominant neurodegenerative disorder characterized by ataxia and progressive motor deterioration [reviewed in Orr and Zoghbi (2001)].

The two AXH domains of ATX1 and HBP1 (ATX1 AXH and HBP1 AXH) share a sequence identity of ∼30% and a homology of ∼50% depending on the species. Though evolutionarily related, the two proteins have different domain boundaries and distinct properties (de Chiara et al., 2003). ATX1 AXH, as solved by crystallography (Chen et al., 2004), forms a dimer of asymmetric dimers. The corresponding region of HBP1 AXH is a monomer in solution as assessed by nuclear magnetic resonance (NMR) (de Chiara et al., 2005). Possibly because of their self-association properties and because of a long insertion in HBP1 AXH, the two domains have the same secondary structure, but are not topologically equivalent.

The AXH motifs seem to play an important role in the function of the respective proteins as most of the interactions of both ATX1 and HBP1 with other molecular partners map into these regions (de Chiara et al., 2003; Yue et al., 2001). Both domains have been shown to bind nucleic acids in vitro, although with different specificities. ATX1 AXH binds RNA homopolymers with preference for poly(rG) and poly(rU) (de Chiara et al., 2003).
This preference corresponds to the same specificity observed for the full-length protein (Yue et al., 2001). HBPI\_AXH (de Chiara et al., 2003; Yue et al., 2001) binds poly(U) and poly(A). Weaker or no binding was observed for poly(G) and poly(C). No structure of an AXH complex with RNA or DNA is available, and the surface of interaction to RNA was hypothesized only on the basis of the combined use of sequence conservation and structure-based analysis. AXH domains, therefore, constitute a paradigmatic example on which to test the possibilities of a dynamics-based alignment approach.

Our analysis, as described in the next sections, is organized in two steps. First, the viability of a dynamics-based alignment as a scheme to predict putative binding sites was investigated by aligning OB-fold members whose interaction surface with RNA or DNA is known. By adopting an elastic network model (Atilgan et al., 2001; Bahar et al., 1997; Delarue and Sanejouand, 2002; Hinsen, 1998; Michetti et al. 2004), we calculated the low-energy modes for members of the full-length OB-fold family and, using the dynamics-based alignment (Zen et al., 2008), identified the regions sharing similar dynamics. These regions were correlated to the surfaces involved in nucleic acid binding and/or recognition. We found that the amino acids involved in several pairwise dynamics-based alignments have a good overlap with the known surface of interaction with nucleic acids. Based on this validation, the dynamics-based alignment was next used to predict the putative DNA/RNA interaction surfaces of HBPI\_AXH and ATX1\_AXH. The predicted sites are a subset of those previously singled-out on the basis of supervised structural alignments (de Chiara et al., 2005) and do not involve positively charged amino acids.

We propose the dynamics-based method as a new approach for predicting functional regions in protein families.

## 2 METHODS

### 2.1 Dataset used for the alignment

#### 2.1.1 AXH domains

The first model of the PDB file 1v06 was taken as the reference structure of HBPI\_AXH, while for ATX1\_AXH we considered the dimer (chains A and B) of PDB file 1oa8.

#### 2.1.2 OB fold representatives

A set of canonical OB-fold representatives was compiled based on the OB-fold survey of Theobald et al. (2003). The detailed list of representatives is shown in Table 1.

### 2.2 Dynamics-based alignment

Dynamics-based alignment establishes one-to-one correspondences between groups of amino acids experiencing similar large-scale motions in two given proteins. The method, described in detail in Zen et al. (2008), is based on an iterative scheme which starts with a tentative selection of the amino acids of two proteins to be put in a one-to-one correspondence. The next steps involve (i) the identification of the large-scale motions of the selected amino acids; (ii) the evaluation of the alignment score. Steps (i) and (ii) are repeated within a stochastic optimization method for maximizing the alignment score over several possible amino acid correspondences. Finally, the statistical significance of the optimal alignment is established. A complete, albeit concise, description of the method follows.

#### 2.2.1 Search rules for amino acid correspondences

The space of possible alignments of two proteins is too large for an exhaustive exploration. The following constraints are accordingly introduced to restrict the search space of matching residues. First, the number of amino acids put in correspondence, n, is limited to multiples of 10, starting from the minimum value n = 70. Second, the amino acids marked for alignment in each protein, numbered sequentially from 1 to n starting from the N-termini, must span blocks of at least 10 consecutive positions, with no intervening gap along the primary sequence.

The simplest strategy for establishing one-to-one correspondences of the marked amino acids in the two proteins is to pair residues with the same marking index, 1 ... n. This intuitive pairing scheme, introduced in Zen et al. (2008), does not enforce a strict one-to-one correspondence at the level of blocks. However, it rules out the possibility of pairing stretches of amino acids that have different block order in the two proteins. The association method was generalized, as next described, to deal with two problems not previously encountered: alignments involving non-canonical amino acids and dealing with distinctly related families such as the canonical versus non-canonical OB-folds.

#### Multiomic proteins:

For multiomic proteins, alignments with all possible orderings of the chains are considered. For a given chain ordering, the amino acids of the entire multimer are numbered consecutively and the simple pairing procedure is applied.

### Canonical and non-canonical OB-folds:

The salient differences of the canonical and non-canonical OB-folds are illustrated in Figure 1a-c. Structurally corresponding β-strands in RPA70 and the AXH domains are shown with the same colour. Strands β3, β2 and β1 of RPA70 match strands β5, β4 and β3 of HBPI1. However, strands β6 and β7 of RPA70 do not correspond to β7 and β6 of HBPI1, as expected for preserved β-strands succession, but with β6 and β7. The latter, in addition, have opposite sequence directionality with respect to RPA70.

Alignments where amino acids are paired sequentially from the N- to C-termini cannot set correspondences of all five β-strands in canonical and non-canonical OB-folds. The pairing scheme was accordingly generalized by ‘remapping’ the amino acid indices so as to achieve a consistent β-strands matching on canonical and non-canonical folds. The procedure is illustrated in Figure 1d. Amino acid reindexing was performed by (i) introducing a single ‘virtual cut’ in HBPI1, and (ii) by changing the order of the two subsequences and the sequence directionality in one of the two (see diagrams at bottom of Fig. 1d). The location of the virtual cut is found by identifying which blocks of residues, and in which sequence order, can be put in loose structural correspondence by local structural alignments. This was done by structurally superposing short segments of 20 amino acids in RPA70 and HBPI1. Such superpositions may induce the spatial proximity (Cα separation below 3 Å) of other amino acids besides those in the two

### Table 1. OB-fold representatives (bold forms) considered in this study

<table>
<thead>
<tr>
<th>No.</th>
<th>Structure</th>
<th>Bound ligand</th>
<th>PDB id</th>
<th>Domain (chain, residue range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RPA70</td>
<td>ssDNA</td>
<td>fjmec</td>
<td>DBD-A (A, 198–289)</td>
</tr>
<tr>
<td>2</td>
<td>RPA70</td>
<td>ssDNA</td>
<td>fjmec</td>
<td>DBD-B (A, 305–402)</td>
</tr>
<tr>
<td>3</td>
<td>ElSSB</td>
<td>ssDNA</td>
<td>6eyg</td>
<td>(A, 1007-1112)</td>
</tr>
<tr>
<td>4</td>
<td>ErRho</td>
<td>ssDNA</td>
<td>2axh</td>
<td>(A, 1089-1112)</td>
</tr>
<tr>
<td>5</td>
<td>OnTTEBP_a1</td>
<td>ssDNA</td>
<td>fjb7</td>
<td>domain 1 (A, 36–204)</td>
</tr>
<tr>
<td>6</td>
<td>OnTTEBP_a1</td>
<td>ssDNA</td>
<td>fjb7</td>
<td>domain 2 (A, 205–314)</td>
</tr>
<tr>
<td>7</td>
<td>OnOB_a2</td>
<td>ssDNA</td>
<td>1k1x</td>
<td>domain 1 (A, 36–204)</td>
</tr>
<tr>
<td>8</td>
<td>OnOB_a2</td>
<td>ssDNA</td>
<td>1k1x</td>
<td>domain 2 (A, 205–314)</td>
</tr>
<tr>
<td>9</td>
<td>OnTTEBP_b</td>
<td>ssDNA</td>
<td>1kkg</td>
<td>domain 1 (A, 36–203)</td>
</tr>
<tr>
<td>10</td>
<td>OnTTEBP_b</td>
<td>ssDNA</td>
<td>1kkg</td>
<td>domain 2 (A, 205–315)</td>
</tr>
<tr>
<td>11</td>
<td>EAcRPS</td>
<td>tRNA anticodon</td>
<td>fctca</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>ScAspRS</td>
<td>tRNA anticodon</td>
<td>1asy</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>ScAspRS</td>
<td>tRNA anticodon</td>
<td>1asy</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>RecG</td>
<td>Junction DNA</td>
<td>1gm5</td>
<td>(A, 187–245)</td>
</tr>
<tr>
<td>15</td>
<td>S12</td>
<td>16S rRNA</td>
<td>1j5c</td>
<td>(L, 26–110)</td>
</tr>
<tr>
<td>16</td>
<td>S17</td>
<td>16S rRNA</td>
<td>1j5e</td>
<td>(Q, 3–102)</td>
</tr>
</tbody>
</table>
The fluctuations are penalized through a quadratic potential:

\[ \beta H_{\text{BP1_AXH}}, \] the canonical order and directionality of the

Large-scale motions of the aligned residues are obtained through

2.2.2 Elastic network modeling of large-scale movements of aligned

identification of the virtual cut for both HBP1 (Fig. 1d) and ATX1.

of matching segments. Inspection of the mapping, allows a transparent

structural alignments methods employing various combinatorial explorations

can capture robust global structural correspondences that are elusive to

such global pairings, which being induced by local structural superpositions

a residue in protein HBP1. The matrix in Figure 1d reports the mapping of

segments. Several local superpositions imply global correspondences, in that

they entail more than half of the residues in RPA70 are in proximity with

a residue in protein HBP1. The matrix in Figure 1d reports the mapping of

such global pairings, which being induced by local structural superpositions

can capture robust global structural correspondences that are elusive to

structural alignments methods employing various combinatorial explorations

of matching segments. Inspection of the mapping, allows a transparent

identification of the virtual cut for both HBP1 (Fig. 1d) and ATX1.

2.2.2 Elastic network modeling of large-scale movements of aligned

residues. Large-scale motions of the aligned residues are obtained through

the \( \beta \)-Gaussian model (Micheletti et al., 2004), which adopts an approximate
description of the potential energy controlling the low-energy structural

fluctuations around a given reference structure as an elastic network function.
The fluctuations are penalized through a quadratic potential:

\[ F(k) = \frac{1}{2} \sum_{i,j} k_i M_{ij} k_j \]  

(1)

where \( k_i \) is the displacement of the \( i \)-th \( \beta \)-strand from the position in the reference structure and the sum is over all the \( N \) residues of the protein. The symmetric matrix \( M \) accounts for the chain connectivity (virtual peptide bond between consecutive amino acids) and for pairwise interactions between amino acids, which are described by two interaction centres: one for the backbone and one for the side chain (except for GLY residues and for the two terminal amino acids of each peptide chain). The software implementing the \( \beta \)-Gaussian model can be requested from the authors free of charge for academic purposes.

Large-scale motions occur along the generalized coordinates corresponding to the low-energy modes of the system, that is the eigenvectors of \( M \) associated to the smallest (non-zero) eigenvalues. The 10 lowest energy modes are generally sufficient to account for most of the structural fluctuations occurring at thermal equilibrium. Spatial modulations associated to these modes typically have a collective character and may be related to protein function. Low-energy modes obtained by the \( \beta \)-Gaussian model have also been shown to be well-consistent with the essential dynamical space obtained from the extensive atomistic molecular dynamics simulations for several proteins (Carnevale et al., 2007; Casella et al., 2005; De Los Rios et al., 2005; Micheletti et al., 2004).

This model provides the general framework for calculating the low-energy modes of the \( n \) residues marked for the alignment. This requires the calculation of the effective quadratic free energy obtained by a thermodynamic integration over the possible displacements of the \( N \times n \) residues not taking part to the alignment, as described hereafter. We assume, for simplicity of notation, to have realigned the residues so that the first \( n \) correspond to the ones marked for alignment. Accordingly, the matrix \( M \) describing the quadratic free energy [see Equation (1)] is written as:

\[ M = M^p + V \]  

(2)

where the superscript \( T \) denotes the transpose; the symmetric matrices \( M^p \) and \( M^p \) with linear size \( n \) and \( N-n \), respectively describe the effective interactions between the residues that are, respectively, marked and not marked for alignment, and the rectangular matrix \( V \) accounts for the interaction between the two sets. The effective free-energy controlling the equilibrium fluctuations of the \( n \) marked residues is given by

\[ F(k) = \frac{1}{2} \sum_{i=1}^{n} \left( k_i M_{ij} k_j + V_{ij} \right) \]  

with: \( M = M^p - V (M^p)^{-1} V^T \), where \( (M^p)^{-1} \) is the pseudoinverse of \( M^p \) (Zen et al., 2008). The eigenvectors associated to the smallest non-zero eigenvalues of \( M \) gives the directions of the sought lowest energy modes for the marked residues.

2.2.3 Alignment score: definition, maximization and statistical significance

The quality of a dynamics-based alignment is measured through a score that measures the correspondence of the low-energy displacements of matching residues along with their good space proximity after an optimal structural superposition.

The structural/dynamical consistency of \( n \) aligned amino acids in two proteins A and B, having low-energy modes \( \{v^{\alpha}, \alpha = 1, \ldots, 10\} \) and \( \{w^{\beta}, \beta = 1, \ldots, 10\} \), respectively, is measured through the following quantity:

\[ q_{ij} = \max \left\{ 0, 1 - \frac{1}{10} \sum_{\alpha=1}^{10} \sum_{\beta=1}^{10} \left( \frac{1}{2} \sum_{i=1}^{10} \sum_{j=1}^{10} v^{\alpha}_{i} w^{\beta}_{j} (d_{ij}) \right) \right\} \]  

(3)

where \( u \) and \( v \) run over the indices of the modes, \( i \) and \( j \) run over the indices of the aligned amino acids, \( d_{ij} \) is the distance between the \( C_\alpha \) positions of the \( i \)-th aligned residue of the two proteins, \( f(d)=\tanh(d-a)-d/2 \) is a distance weighting factor interpolating the asymptotic values of 0 and 1 for distances, respectively, much larger and smaller than \( d_0=4\AA \). We remark that Equation (3) can be viewed as a distance-weighted generalization of the root mean square inner product (RMSIP), which is customarily used to measure correspondences of two sets of essential dynamical spaces.
Amino acids are divided in those predicted to interact with nucleic acids. The dynamics-based prediction of nucleic acid binding amino acids is carried out in a systematic analysis to identify the key aligned residues that are, or are not, contacting nucleic acids and is defined as whose accessible surface area (ASA) changes by more than 1 Å² upon observation, the alignment consensus score of residue \(k\) compared, for validation purposes, against the sites that actually bind DNA. For each protein, we calculated the reference protein in a given alignment is measured as following the notation of Equation (3):

\[
\xi_k = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i - \beta_j \left( \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_i \right)
\]

where \(i\) is the index of the matching pair to which amino acid \(k\) takes part to. The physical meaning of \(\xi_k\) is transparent as, apart from a multiplicative factor, it represents the local contribution to the mean square inner product of the modes of the aligned residues. For a perfect matching of the modes \(\alpha\) and \(\beta\), the average value of \(\xi_k\) per aligned residue is 1. Based on this observation, the alignment consensus score of residue \(k\) is defined as \(\xi_k\), where the brackets denote the average of \(\xi_k\) over significant alignments having same length, \(n\). Since most of the significant alignments between the OB-fold domains in Table 1 have length \(n\approx70\), we have used only the alignments of 70 residues to calculate the consensus values.

The consensus score is used to predict a set of residues putatively involved in the binding of the nucleic acids.

2.4 Definition of DNA/RNA-binding interface

The dynamics-based prediction of nucleic acid binding amino acids is compared, for validation purposes, against the sites that actually bind DNA or RNA. As in Jones et al. (2003), they are identified as the amino acids whose accessible surface area (ASA) changes by more than 1 Å² upon omitting the nucleic acid from the available structure of the protein/DNA (or RNA) complex. The calculation of the ASA was performed with NACCESS (Hubbard, 1993). For most of the proteins in Table 1, the typical fraction of residues contacting nucleic acids is \(\sim 20\%\).

2.5 Performance of the dynamics-based prediction scheme

Amino acids are divided in those predicted to interact with nucleic acids or not according to whether their consensus score is, respectively, above or below a given threshold. All possible values for the threshold were considered and the performance of the prediction was assessed by comparison against the sets of amino acids that are known to interact (or not interact) with DNA/RNA. For a given threshold value, the prediction is characterized, as customary, in terms of the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The TP are the amino acids that are correctly predicted to interact with DNA or RNA, while TN are those correctly predicted not to interact. The FP are the amino acids that are incorrectly predicted to interact with DNA or RNA, while FN are those incorrectly predicted not to interact. These basic quantities are used to define the accuracy, specificity and selectivity of the prediction (Baldi et al., 2000).

The accuracy is the fraction of correct prediction for amino acids that are, or are not, contacting nucleic acids and is defined as (TP + TN)/(TP + TN + FP + FN). The specificity, defined as TP/(TP + FP), represents the fraction of correct hits among residues predicted. The sensitivity, TP/(TP + FN), is the fraction of residues known to interact with DNA/RNA which are predicted to do so.

The predictive performance of the method as a function of the consensus score threshold is aptly summarized by the receiver operating characteristic curve (ROC curve) obtained by plotting ‘hit rate’ (sensitivity, TP/(TP + FN)) versus the ‘false alarm rate’ (false positive rate, FP/(FP + TN)).

3 RESULTS

3.1 Alignment of the OB-fold family

Dynamics-based alignments were carried out among all 120 distinct pairings of the 16 canonical OB-fold representatives constituted by all the domains listed in Table 1. The quality of each alignment is conveyed by an alignment score which rewards correspondences between amino acids that have (i) similar geometric relationships in the two proteins and (ii) sustain similar large-scale movements. The combined consideration of structural and dynamical features ensures that high-scoring alignments reflect genuine correspondences of large-scale rearrangements in two given proteins. The statistical significance of each alignment is quantified by comparing the score against a reference distribution of scores from a heterogeneous set of enzymes. From this comparison, we could calculate a P-value (or equivalently a z-score). Given the limited size of the database considered, we assumed as indicative of a significant alignment a z-score > 2.3, corresponding to a P-value < 0.01. The dynamics-based scores for all pairwise alignments among the proteins in Table 1 are provided in the density maps of Figure 2a. The accompanying graph, see Figure 2b, summarizes the dynamics-based correspondences having a statistical significance higher than the above mentioned threshold. Inspection of the graph reveals the existence of several triangular relations (i.e. protein A is in relation with proteins B and C, and also B is in relation with C). Proteins OnTEBP, RPA70 and RecG, for instance, form a completely connected subgraph. These circular relationships suggest the existence of a common alignable core among these proteins. This can be verified by inspecting Figure 2a that shows pileup representations of the alignments involving OnTEBP a2 (domain 1), RecG and RPA70 (repeat DBD-B) and the alignable partners. The structural superposition of OnTEBP a2 (domain 1) with RecG and RPA70 (repeat DBD-B) is shown in Figure 3b and c, respectively. The alignable regions involve amino acids that are flexible and in proximity of the bound nucleic acid, as can be appreciated by comparison with the complexes in Figure 3d-f.

This observation suggests that the set of amino acids of a given OB-fold that can be significantly aligned with several other OB-fold partners are typically located in regions involved in nucleic acid binding. This hypothesis was quantitatively verified with the following analysis (Baldi et al., 2000). We computed the consensus alignment score for all amino acids of proteins RPA70 (repeat DBD-A), EcSSB, EcRho, OnTEBP a1 (domain 1), OnTEBP a2 (domain 1), EcAspRS, ScAspRS (domain 1), RecG Notice that proteins S12 and S17, which are largely surrounded by nucleic acids, were not considered for the test and that, to limit redundancy, only the N-terminal domain was retained for multidomain proteins.

Amino acids with a sufficiently high consensus score are expected to be relevant for the functional dynamics and hence to correlate with...
sites involved in nucleic acid binding. To assess the extent to which the consensus score can be used to predict interaction sites with DNA/RNA, we carried out the performance analysis of Section 2.5.

The results are summarized in the plots in Figure 4. The plots can be used to set the threshold for the consensus score so to have a balanced predictive performance in terms of accuracy, specificity and selectivity. In fact, excessively large threshold values correspond to very few predictions for amino acids interacting with DNA/RNA and this reflects in a poor coverage of the sites that are known to interact with nucleic acids. Conversely, very small threshold values result in predicting that almost all amino acids interact with DNA/RNA, thus, leading to a large fraction of false positives. A balance between these two limiting situations is achieved by setting the consensus score threshold to 0.7. Examples of the consensus regions are given in Figure 3d–f. The corresponding overall accuracy of the algorithm is 79%, specificity is 38% and sensitivity is 24%. A useful term of reference for these values is provided by advanced sequence-based techniques for the prediction of nucleic acid binding sites. For instance, an accuracy of 71%, specificity of 35% and sensitivity of 53% was calculated for the method implemented by Yan et al. (2006), in a different dataset of DNA-binding proteins. In addition, on

![Fig. 2.](image1) ![Fig. 3.](image2) ![Fig. 4.](image3)
The above results indicate that, within the limits of binding specificity, the consensus residues point at regions involved in nucleic acid binding and are therefore not common to all OB-folds. The specific region considered here, the online sequence-based method of Hwang et al. (2007) for DNA binding sites prediction had an accuracy of 63%, specificity of 23% and sensitivity of 45% (further details on the difference between the sequence- and dynamics-based predictions are reported in the Supplementary Material). It therefore emerges that the dynamics-based approach compares well in terms of accuracy and specificity, while returns appreciably smaller values for sensitivity. This aspect is rationalized by the observation that the dynamics-based alignment will be especially promoted in correspondence of flexible amino acids, and consequently the residues close to the nucleic acid chain and with a low mobility are likely to have a low consensus score. The dynamics-based predictions are, therefore, particularly targeted at a specific subset of nucleic acid binding sites (the mobile ones) and this reflects in a diminished sensitivity of the algorithm compared with the complementary sequence-based methods. Additionally, regions which cannot be aligned and that are therefore not common to all OB-folds may be also involved in binding and be the ones responsible for recognition specificity.

3.2 Prediction of the nucleic acid binding surface of the AXH domains

The above results indicate that, within the limits of binding specificity, the consensus residues point at regions involved in nucleic acid binding. The approach was used as a predictive tool for representatives of the AXH-domain family.

Prediction of the nucleic acid binding surface based on sequence and structural comparison with other members of the OB-fold was previously attempted (de Chiara et al., 2005). However, the two families are too divergent to extract useful hints from sequence conservation, whereas a structure-based analysis was inconclusive. It was only through a combined use of sequence and structural conservation that two distinct patches of conserved or semiconserved residues could be identified. Only one of them corresponds to the surface involved in nucleic acid binding in other OB-folds. We therefore reasoned that this example would be an appropriate case for attempting a dynamics-based prediction.

Since HBPI_AXH is monomeric and therefore easier to deal with, we aligned it (1v06) first against OB-fold representatives using their dynamics properties. HBPI_AXH can be significantly aligned with two distinct regions of RPA70 (z-score 3.5) (Fig. 5a). It also aligns with RecG with a z-score of 2.5 (data not shown). The single-stranded DNA-binding domain of human RPA70 (residues 183–420) contains two tandem OB-fold repeats. Dynamics-based alignments of HBPI_AXH against both repeats are highly consistent and involve residues 212–237 and 214–235 (including f1 and f2 with a reversed backbone orientation to regions f1 and f2 (Fig. 1a) of DBD-A and DBD-B. The consensus regions emerging from such alignments strongly suggest that nucleic acid binding involves HBPI residues N228, K229, E230, S270, V271, S272, F273, G274, E275, T286, V287 and E288 which correspond to the cavity formed by loops f1/f2, f1/f2 and f1/f2 of HBPI (Fig. 5a, left). These residues correspond to residues in direct contact with DNA in the holo form of RPA70 (Fig. 5a, right). The predicted residues are not positively charged, suggesting that the interaction would not be electrostatically driven, but rather sequence or structural specific. They are well consistent with those previously predicted on the base of a structural alignment (Fig. 5b; de Chiara et al., 2005).
ATX1 AXH aligns with RecG with a z-score of 3.3 (Fig. 5c). The aligned side chains are all exposed and do not interfere with dimer formation (Fig. 5d).

Finally, the dynamics-based alignment between HBPI AXH and ATX1 AXH comprises residues 257–271, 274–288, 290–339, 222–231 and 609–623, 624–638, 639–688, 565–574, respectively (Fig. 5e). It is worth noting that the region 222–231 of HBPI AXH, which is not topologically equivalent in the two proteins, aligns with a reverse orientation in sequence with the corresponding region of ATX1 AXH (Fig. 1b and c). This could suggest that despite their difference, the two regions share a functional role within the context of the domain.

ATX1 AXH (monomer A) and HBPI AXH can be superposed by structural criteria (Fig. 5f) with an RMSD of 3.8 Å over 84 amino acids. The two folds differ for the topology of an N-terminal P1, P2 and P3 motif which packs differently in the two structures. Concomitantly, the spacing between these three elements of secondary structure is different and only the regions 260–335 of HBPI AXH and 612–684 of ATX1 AXH can be meaningfully aligned. These regions are a subset of the residues alignable on structural considerations (de Chiara et al., 2005). Exposed conserved and semiconserved residues of the AXH subfamily (corresponding to K217, E235, D236, E268, G285, P324, N344, K225, E230, W231, R239, A240, E246, E269, K307, E327, L328, I330 and N341 in HBPI, Fig. 5b) cluster near the two exposed patches that comprise or are directly contiguous to those predicted by dynamics-based alignment.

Interestingly, as for the alignment of ATX1 AXH with other OB-folds, the two AXH folds would not lead to interference of nucleic acid binding with the dimerization interface of the ATX1 AXH domain, thus being well compatible with the knowledge that this domain is an obliged dimer in solution (de Chiara et al., 2005).

4 CONCLUSIONS

Several methods, both sequence- and structure-based, exist that provide predictions for nucleic acid binding sites in proteins. While sequence-based techniques have the advantage of being applicable when structural models are not available, it is commonly recognized that exploiting structure-based information (such as surface shape, solvent accessibility, interatomic interaction potentials, etc.) can significantly improve prediction. Here, we introduce and discuss a new method that, while not making use of primary sequence information, identifies putative binding sites on the basis of similarities in the dynamics of a family of proteins. The new approach may be used (possibly in conjunction with other criteria) to predict the interaction surface within a protein family.

We have shown here a specific application to the OB-fold, selected because of the large plethora of data available. By comparing the dynamics of a comprehensive subset of members of the family known both in their free and bound forms, we observed that nucleic acid binding sites share common dynamical properties. This observation prompts the consideration that the large-scale movements that putatively accompany/assist biological functionality may be conserved among protein families and that can be detected using dynamics-based alignments. We then applied this information to a non-canonical OB-fold, for which the putative nucleic acid binding surface could not be easily predicted from sequence or structural (static) considerations.

While still in need of further validation using different and even more divergent examples, for which sequence- and structure-based alignments may be not obvious, our present results encourage us to believe that our method may develop into a useful and powerful predictive tool. Natural applicable avenues for the method, which we plan to validate in other contexts, are structure/function genomics studies.

Funding: Italian Ministry for Education (grant PRIN-2006025255 and FIRB RBNE03PX83).

Conflict of Interest: none declared.

REFERENCES


Hubbard, S.J. (1993) NACCESS. Department of Biochemistry and Molecular Biology, University College, London.


Laekema, P. et al. (1997) The HMG-Box transcription factor HBPI is targeted by the pocket proteins and EIA. Oncogene, 14, 2721–2728.


Predictions of nucleic acid binding sites

Yan, C. et al. (2006) Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics, 7, 262.