Category

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

Andrea Zen[1], Cesira de Chiara[2], Annalisa Pastore[2], Cristian Micheletti[1,*]

[1] SISSA, CNR-INFM Democritos and Italian Institute of Technology, Via Beirut 2, I-34151 Trieste (Italy)

[2] National Institute for Medical Research, The Ridgeway, London NW71AA (UK)

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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 the well characterized nucleic acid binding family. Protein regions consensually involved in statistically-significant dynamics-based alignments were found to correlate with nucleic acids 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.

Contact: michelet@sissa.it

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 (Carnevale et al., 2006; Capozzi et al., 2007). Quantitative algorithms are presently available to detect similar motions in protein pairs. The procedure is termed dynamics-based alignment as 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. β1 is shared by both sheets whilst β3 and β5 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 HMG box transcription factor HBP1 and the polyglutamine-containing ATX1 protein (Lesage et al., 1994; Banfi 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 (Paulson 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 ca. 30% and a homology of ca. 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 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

*To whom correspondence should be addressed.

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). HBP1_AXH (de Chiara et al., 2003; Yue et al., 2001) binds poly(rU) and poly(rA). Weaker or no binding was observed for poly(rG) and poly(rC). 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 (Bahar et al., 1997; Hinsen, 1998; Atilgan et al., 2001; Delarue et al., 2002; Micheletti et al. 2004), we calculated the low energy modes for members of the 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 HBP1_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.

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

#	Structure	bound ligand	PDB id	domain (chain, residue range)
1	RPA70	ssDNA	1jmc	DBD-A (A, 198-289)
2	RPA70	ssDNA	1jmc	DBD-B (A, 305-402)
3	EcSSB	ssDNA	1eyg	(chains A)
4	EcRho	ssRNA	2a8v	(A, 48-118)
5	OnTEBP α1	ssDNA	1jb7	domain 1 (A, 36-204)
6	OnTEBP α1	ssDNA	1jb7	domain 2 (A, 211-314)
7	OnTEBP α2	ssDNA	1kix	domain 1 (A, 36-204)
8	OnTEBP α2	ssDNA	1kix	domain 2 (A, 205-314)
9	OnTEBP β	ssDNA	1k8g	domain 1 (A, 36-203)
10	OnTEBP β	ssDNA	1k8g	domain 2 (A, 205-315)
11	EcAspRS	tRNA anticodon	1c0a	(A 1-104)
12	ScAspRS	tRNA anticodon	1asy	domain 1 (A, 68-204)
13	ScAspRS	tRNA anticodon	1asy	domain 2 (B, 68-204)
14	RecG	Junction DNA	1gm5	(A, 157-245)
15	S12	16S rRNA	1j5e	(L ,26-110)
16	S17	16S rRNA	1j5e	(Q, 3-102)

2 METHODS

2.1 Dataset used for the alignment

AXH domains The first model of the PDB file 1v06 was taken as the reference structure of HBP1_AXH, while for ATX1_AXH we considered the dimer (chains A and B) of PDB file 1oa8.

OB fold representatives A set of canonical OB-fold representatives was compiled based on the OB-fold survey of ref. (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 of amino acids 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. Secondly, 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 ref. (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-monomeric proteins and dealing with distantly related families such as the canonical versus non-canonical OB folds.

Multimeric proteins. For multimeric 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 Fig. 1a-c. Structurally-corresponding β-strands in RPA70 and the AXH domains are shown with the same colour (and same letter in the secondary structure topologies). Strands β_1, β_2 and β_3 of RPA70 match strands β_3, β_4 and β_5 of HBP1. However, strands β_5 and β_4 of RPA70 do not correspond to β_7 and β_6 of HBP1, as expected for preserved β-strands succession, but with β_1 and β_2. 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 generalised by "remapping" the amino acid indices so to achieve a consistent β-strands matching on canonical and non-canonical folds. The procedure is illustrated

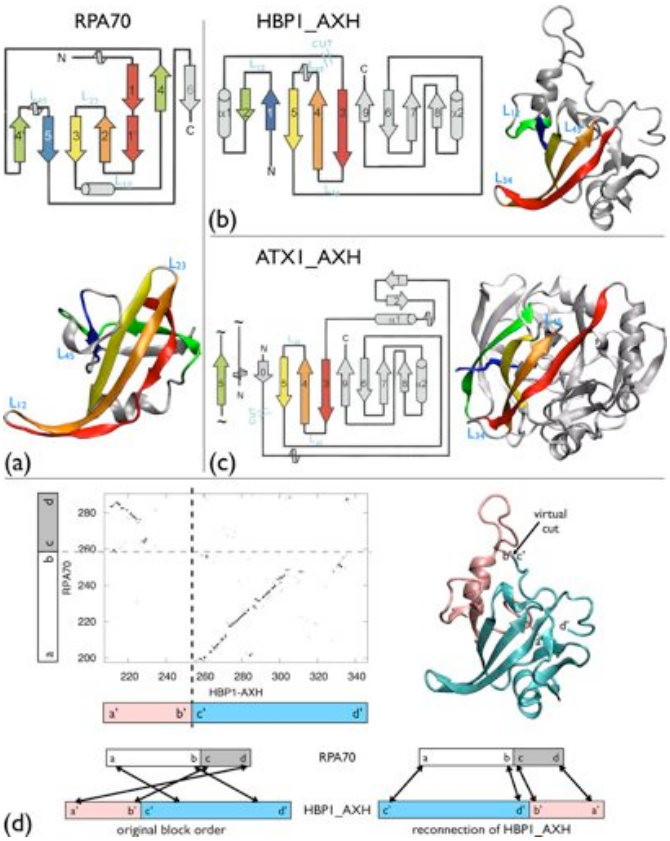

Figure 1 Comparison of topologies and structures of (a) the canonical OB-fold (RPA70, repeat DBD-A) and the non-canonical one of (b) HBP1_AXH and (c) the dimeric ATX1_AXH. Corresponding β-strands are indicated with the same colors thereby highlighting the different sequential order and sequence directionality of matching strands. In ataxin-1, the green strand is from the contiguous monomer (only this element is indicated). It is also worth noting that the symmetry of the dimer breaks around this region and strand β5 in monomer A corresponds to a short 3-10 helix in monomer B. (d) In HBP1_AXH the canonical order and directionality of β-strands is achieved (for alignment convenience) by juxtaposing the two parts separated by the virtual cut, as shown. The procedure to identify the virtual cut is described in Methods, section 2.2.1.

in Fig. 1d. Amino acid reindexing was performed by (i) introducing a single "virtual cut" in HBP1, and (ii) by changing the order of the two subchains and the sequence directionality in one of the two (see diagrams at bottom of panel d). 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 HBP1. Such superpositions may induce the spatial proximity (C_α separation below 3Å) of other amino acids besides those in the two 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 fig. 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 β-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 penalised through a quadratic potential:

$$F(\delta\vec{x}) = \frac{1}{2}\sum_{i,j=1}^{N} \delta\vec{x}_i M_{ij} \delta\vec{x}_j \qquad (1)$$

where $\delta\vec{x}_i$ is the displacement of the ith C_α 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 centers: 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 β–Gaussian model can be requested from the authors free of charge for academic purposes.

Large-scale motions occur along the generalised coordinates corresponding to the low-energy modes of the system, that is the eigenvectors of M associated to the smallest (non-zero) eigenvalues. The ten 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 β-Gaussian model have also been shown to be well-consistent with the essential dynamical space obtained from extensive atomistic molecular dynamics simulations for several proteins (Micheletti et al., 2004; Cascella et al., 2005; De Los Rios et al., 2005; Carnevale et al., 2007).

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

$$M = \begin{bmatrix} M^a & V \\ V^T & M^b \end{bmatrix}$$

where the superscript T denotes the transpose; the symmetric matrices M^a and M^b (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

$$\tilde{F}(\delta\vec{x}^a) = \frac{1}{2}\sum_{i,j=1}^{n} \delta\vec{x}_i^a \tilde{M}_{ij} \delta\vec{x}_j^a \qquad (2)$$

with: $\tilde{M} = M^a - V[M^b]^{-1}V^T$, where $[M^b]^{-1}$ is the pseudoinverse of M^b (Zen et al., 2008). The eigenvectors associated to the smallest nonzero eigenvalues of $\tilde{M}$ give 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 alignment. For detecting similar relative motion of the aligned regions it is required that proteins under comparison have a correct relative orientation (set by performing an optimal structural superposition of the specific residues marked for alignment).

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:

$$q_n = \sqrt{\max\{0, \frac{1}{10}\sum_{\alpha,\beta=1}^{10} |\sum_{j=1}^{n} \vec{v}_j^\alpha \cdot \vec{w}_j^\beta| |\sum_{i=1}^{n} \vec{v}_i^\alpha \cdot \vec{w}_i^\beta f(d_i)|\}} \qquad (3)$$

where α and β run over the indices of the modes, i and j run over the indices of the aligned amino acids, d_i is the distance between the C_α positions of the ith aligned residue of the two proteins, $f(d)=[1-\tanh((d-d_c)/2)]/2$ is a distance weighting factor interpolating the asymptotic values of 0 and 1 for distances respectively much larger and smaller than $d_c=4$Å. We remark that eq. (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.

Following Zen et al. (2008), the statistical significance of an alignment is assigned comparing the score of eq. (3) against a Gaussian reference distribution for alignment scores involving n residues in unrelated protein pairs. In this way, to each alignment it is associated a z-score or, equivalently, a p-value. The former is a measure of how distant (in terms of standard deviations) is the obtained score from the average random reference case. The p-value, instead, corresponds to the probability that an alignment of n residues of two unrelated proteins returns a score higher than the one actually observed. The lower is the p-value (i.e. the higher the z-score), the more atypical, and hence significant, is the alignment.

2.3 Consensus profile of dynamics-based alignment

We carried out a systematic analysis to identify the key aligned residues that recurrently appear in significant alignments. For each protein we calculated the consensus profile of dynamical accord, that is the residue-wise average contribution to the statistically-significant alignment with other OB-fold members. The degree of dynamical involvement of the kth amino acid of the reference protein in a given alignment is measured as (following the notation of eq. 3):

$$\xi_k = \frac{n}{10}\sum_{\alpha,\beta=1}^{10} \vec{v}_i^{\,\alpha}\cdot\vec{w}_i^{\,\beta}\,[\sum_{j=1}^{n}\vec{v}_j^{\,\alpha}\cdot\vec{w}_j^{\,\beta}] \qquad (4)$$

where i is the index of the matching pair to which amino acid k takes part to. The physical meaning of ξ_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 $\{v\}$ and $\{w\}$, the average value of ξ_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 ξ_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=70, 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 Å^2 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 ~20%.

2.5 Performance of the dynamics-based prediction scheme

Amino acids are divided in those predicted to interact or not to interact with nucleic acids 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/false positive and true/false negatives. The true positives, TP, [true negatives, TN] are the amino acids that are correctly predicted [not] to interact with DNA or RNA. The false positives, FP, [false negatives, FN] are the amino acids that are incorrectly predicted [not] to interact with DNA or RNA. 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 Fig. 2a. The accompanying graph, see Fig. 2b, summarises 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 Fig. 3a which shows pileup representations of the alignments involving OnTEBP α2 (domain 1), RecG and RPA70 (repeat DBD-B) and the alignable partners.

The structural superposition of OnTEBP α2 (domain 1) with RecG and with RPA70 (repeat DBD-B) is shown in Figs. 3b and 3c, 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 Fig.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-

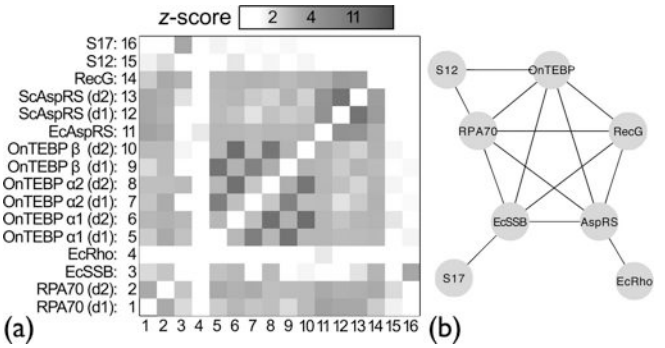

Figure 2 (a) Density map of the *z*-score for all pairwise dynamics-based alignment of canonical OB-fold representatives (indexing according to the Table 1). (b) Graph representation of significant pairwise alignments (*z*-score > 2.3).

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 α1 (domain 1), OnTEBP α2 (domain 1), OnTEBP β (domain 1), EcAspRS, ScAspRS (domain 1), RecG. Notice that proteins S12 and S17, that 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 summarised in the plots in Fig. 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 figs.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-acids binding sites. For instance, an accuracy of 71%, a specificity of 35% and a 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 the specific dataset considered here, the on-line sequence-based method of Hwang et al. (2007) for DNA binding-sites prediction had an accuracy of 63%, a specificity of 23% and a sensitivity of 45% (further details on the difference between the sequence-based and dynamics-based predictions are reported in Supp. Info.). 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 rationalised 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 to 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

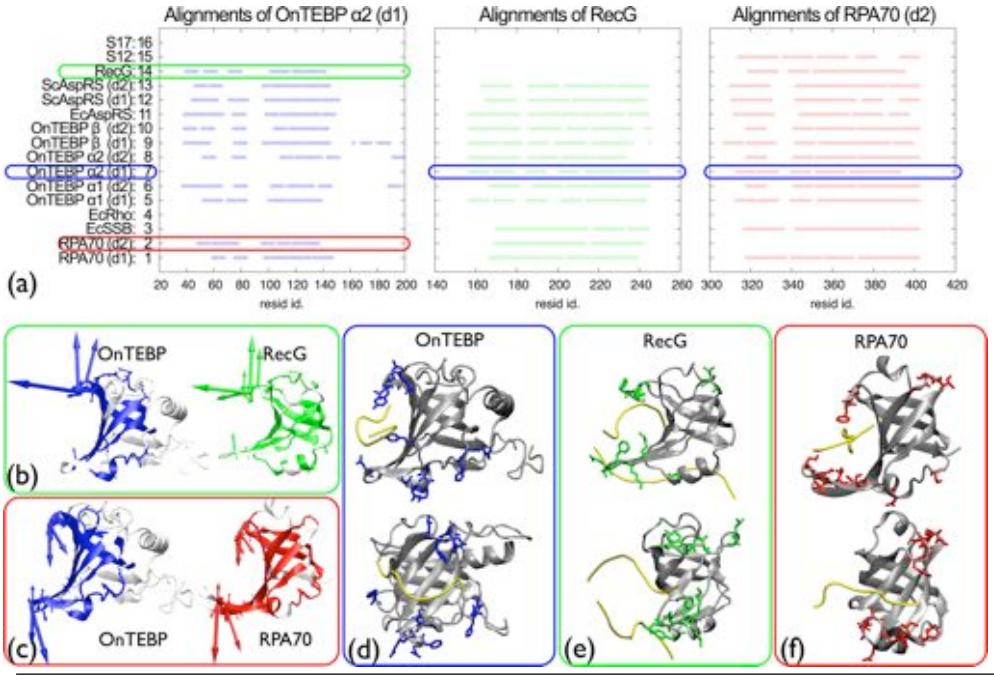

Figure 3 (a) Pileup representations of the significant alignments involving: OnTEBP α2 domain 1, RecG and of RPA70 repeat DBD-B. The dynamics-based alignment of OnTEBP α2 domain 1 and RecG is shown in (b), while the one between OnTEBP α2 domain 1 and RPA70 repeat DBD-B is shown in (c). Amino acids involved in alignments are colored. The arrows represent the three best corresponding (Zen et al., 2008) lowest-energy modes for the aligned regions. Panels (d), (e) and (f) illustrate, respectively, the consensus residues of OnTEBP α2 domain 1, RecG and RPA70 repeat DBD-B. Two different views are displayed, the upper one is the same adopted in panels (b) and (c), the lower is rotated of 90° around the z-axis. Nucleic acid strands are shown as yellow tubes and the sidechains of consensus residues are highlighted in color.

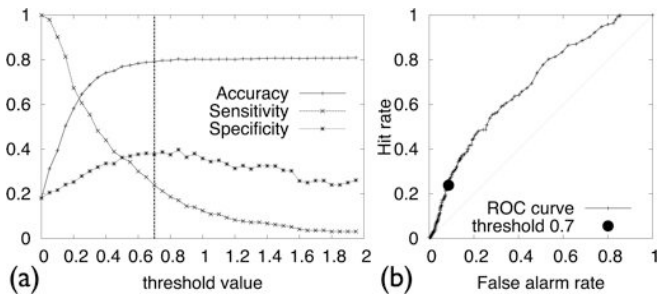

Figure 4 (a) Trend for the accuracy, sensitivity and specificity of dynamics-based predictions of amino acids at the protein/nucleic acid interface are shown as a function of the consensus score threshold. (b) Corresponding receiver operating characteristic (ROC) curve.

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 semi-conserved 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 HBP1_AXH is monomeric and therefore easier to deal with, we aligned it (1v06) first against OB-fold representatives using their dynamics properties. HBP1_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 HBP1_AXH against both repeats are highly consistent and involve residues 212-237 and 214-235 (including β1 and β2) with a reversed backbone orientation to regions β4 and β5 (fig.1a) of DBD-A and DBD-B. The consensus regions emerging from such alignments strongly suggest that nucleic acid binding involves HBP1 residues N228, K229, E230, S270, V271, S272, F273, G274, E275, T286, V287 and E288 which correspond to the cavity formed by loops β1/β2, β3/β4 and β4/β5 of HBP1 (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 sidechains are all exposed and do not interfere with dimer formation (fig. 5d).

Finally, the dynamics-based alignment between HBP1_AXH and ATX1_AXH comprises residues 257-271, 274-288, 290-339, 222-213 and 609-623, 624-638, 639-688, 565-574 respectively (fig. 5e). It is worth noting that the region 222-213 of HBP1_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,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 HBP1_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 β1, β2 and α1 motif which packs differently in the two structures.

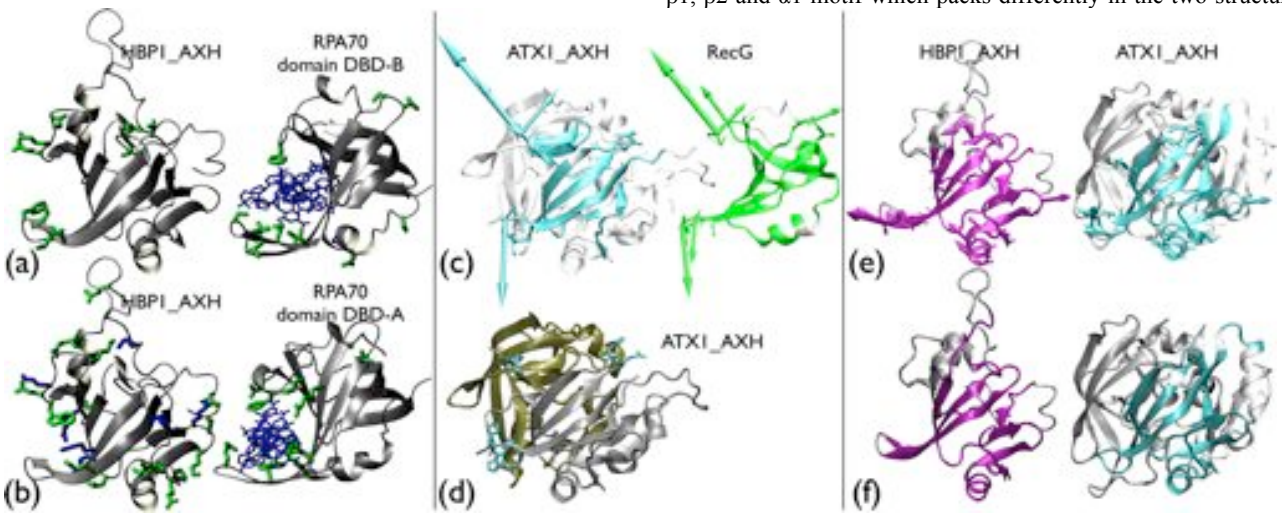

Figure 5 (a) Ribbon representations of HBP1_AXH (left) and of the DBD-B repeat of RPA70 (right) as dynamically aligned. The side chains of consensus residues are explicitly reported. (b) Comparison of HBP1_AXH (left) and of the DBD-A repeat of RPA70 (right) in complex with DNA (in blue) as aligned structurally (adapted from fig. 5 of de Chiara et al., 2005). The side chains of completely and semiconserved residues of HBP1_AXH are indicated in green, additional lysines and arginines that could contribute to binding are shown in blue. DNA and the side chains of residues of RPA70 DBD-A in contact with DNA are indicated explicitly in blue and green respectively. (c) Dynamics-based alignment of the ATX1_AXH dimer (left) with RecG. (right). Aligned regions are shown in cyan and green respectively. Arrows represent the three best corresponding (Zen et al., 2008) lowest-energy modes for the aligned residues. (d) The sidechians of the consensus residues are shown on the ATX1_AXH dimer in cyan. The two subunits forming the dimer are in gold and silver. (e) Dynamics-based alignment of HBP1_AXH (left) and ATX1_AXH (right). Aligned regions are colored in purple and cyan respectively. The dynamics-based alignment involves 90 residues with an RMSD of 3.5 Å. The RMSIP of the ten lowest-energy modes (the best corresponding three are shown as arrows) as calculated using the βGaussian network model, is 0.77. (f) Structurally-based alignment of the same proteins as achieved by DALIlite. 84 residues were aligned with an RMSD of 3.8 Å.

Concomitantly, the spacing between these three elements of secondary structure is different and only the regions 260-335 of HBP1_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, L298, K307, E327, L328, I330 and N341 in HBP1, 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 used 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 applicative avenues for the method, which we plan to validate in other contexts, are structure/function genomics studies.

ACKNOWLEDGEMENTS

We acknowledge financial support from the Italian Ministry for Education (grant PRIN-2006025255 and FIRB RBNE03PX83).

REFERENCES

Atilgan,A.R. *et al.* (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–515.

Bahar,I. *et al.* (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, *Folding & Design*, **2**, 173-181.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412-424.

Banfi,S. *et al.* (1994) Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nature Genet.*, **7**, 513-520.

Berasi,S.P. *et al.* (2004) HBP1 repression of the p47phox gene: cell cycle regulation via the NADPH oxidase. *Mol. Cell. Biol.*, **24**, 3011-3024.

Capozzi,F. *et al.* (2007) Essential dynamics of helices provide a functional classification of EF-hand proteins. *J. Proteome Res.*, **6**, 4245-4255.

Carnevale,V. *et al.* (2006) Convergent dynamics in the protease enzymatic superfamily. *J. Am. Chem. Soc.*, **128**, 9766-9772.

Carnevale,V. *et al.* (2007) Large-scale motions and electrostatic properties of Furin and HIV-1 protease. *J. Phys. Chem. A*, **111**, 12327-12332.

Cascella,M. *et al.* (2005) Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases. *J. Am. Chem. Soc., 127, 3734-3742.*

Chen,Y.W. *et al.* (2004) The structure of the AXH domain of spinocerebellar ataxin-1. *J. Biol. Chem.*, **279**, 3758-3765.

de Chiara,C. *et al.* (2003) The AXH module: an independently folded domain common to ataxin-1 and HBP1. *FEBS Lett.*, **551**, 107-112.

de Chiara,C. *et al.* (2005) The AXH domain adopts alternative folds the solution structure of HBP1 AXH. *Structure*, **13**, 743-753.

De Los Rios,P. *et al.* (2005) Functional dynamics of PDZ binding domains: a normal-mode analysis. *Biophys. J.*, **89**,14-21.

Delarue,M. and Sanejouand,Y.H. (2002) Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: The elastic network model. *J. Mol. Biol.*, **320**, 1011–1024.

Frauenfelder,H. *et al.* (1991) The energy landscapes and motions of proteins. *Science*, **254**, 1598-1603.

Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.

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

Hwang,S. *et al.* (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634-6.

Jones,S. *et al.* (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins, *Nucleic Acids Res.*, **31**, 7189-7198.

Lavender,P. *et al.* (1997) The HMG-box transcription factor HBP1 is targeted by the pocket proteins and E1A. *Oncogene*, **14**, 2721-2728.

Lesage,F. *et al.* (1994) Expression cloning in K+ transport defective yeast and distribution of HBP1, a new putative HMG transcriptional regulator. *Nucleic Acids Res.*, **22**, 3685-3688.

Micheletti,C. *et al.* (2004) Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and gaussian models. *Proteins*, **55**, 635–645.

Murzin,A.G. (1993) OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *Embo J.*, **12**, 861-867.

Mushegian,A.R. *et al.* (1997) Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 5831-5836.

Orr,H.T. and Zoghbi,H.Y. (2001) SCA1 molecular genetics: a history of a 13 year collaboration against glutamines. *Hum. Mol. Genet.*, **10**, 2307-2311.

Paulson,K.E. *et al*. (2007) Alterations of the HBP1 transcriptional repressor are associated with invasive breast cancer. *Cancer Res.*, **67**, 6136-6145.

Tevosian,S.G. *et al.* (1997) HBP1: a HMG box transcriptional repressor that is targeted by the retinoblastoma family. *Genes Dev.*, **11**, 383-396.

Theobald,D.L. *et al.* (2003) Nucleic acid recognition by OB-fold proteins. *Annu. Rev. Biophys. Biomolec. Struct.*, **32**, 115-133.

Tsai,C.C. *et al.* (2004) Ataxin 1, a SCA1 neurodegenerative disorder protein, is functionally linked to the silencing mediator of retinoid and thyroid hormone receptors. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 4047-4052.

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

Yue,S. *et al.* (2001) The spinocerebellar ataxia type 1 protein, ataxin-1, has RNA-binding activity that is inversely affected by the length of its polyglutamine tract. *Hum. Mol. Genet.*, **10**, 25-30.

Zen,A. *et al.* (2008) Correspondences between low-energy modes in enzymes: dynamics based alignment of enzymatic functional families. *Protein Sci*., **17**, 918–929.