

Structural bioinformatics

Natively unstructured regions in proteins identified from contact predictions

Avner Schlessinger^{1,2,*}, Marco Punta^{1,2} and Burkhard Rost^{1,2}¹Department of Biochemistry and Molecular Biophysics, Columbia University and ²Columbia University Center for Computational Biology and Bioinformatics (C2B2), NorthEast Structural Genomics Consortium (NESG), New York, NY, USA

Received on April 28, 2007; revised on June 26, 2007; accepted on June 27, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Natively unstructured (also dubbed *intrinsically disordered*) regions in proteins lack a defined 3D structure under physiological conditions and often adopt regular structures under particular conditions. Proteins with such regions are overly abundant in eukaryotes, they may increase functional complexity of organisms and they usually evade structure determination in the unbound form. Low propensity for the formation of internal residue contacts has been previously used to predict natively unstructured regions.

Results: We combined PROFcon predictions for protein-specific contacts with a generic pairwise potential to predict unstructured regions. This novel method, *Ucon*, outperformed the best available methods in predicting proteins with long unstructured regions. Furthermore, *Ucon* correctly identified cases missed by other methods. By computing the difference between predictions based on specific contacts (approach introduced here) and those based on generic potentials (realized in other methods), we might identify unstructured regions that are involved in protein–protein binding. We discussed one example to illustrate this ambitious aim. Overall, *Ucon* added quality and an orthogonal aspect that may help in the experimental study of unstructured regions in network hubs.

Availability: http://www.predictprotein.org/submit_ucon.html

Contact: as2067@columbia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

1.1 Many different flavors of unstructured regions

Regions in proteins that do not adopt well-ordered 3D structures under physiological conditions are often dubbed natively unstructured, disordered, intrinsically unstructured or unfolded. Typical are proteins that adopt stable 3D structures only upon binding to substrates to carry out their function (Fig. 1A–C), or proteins that perform a particular function in their ‘unstructured state’ (Dyson and Wright, 2002, 2005; Oldfield *et al.*, 2005b). The better our experimental and computational means of identifying such proteins, the more

we realize that they come in a great variety: some adopt regular secondary structure (helix or strand) upon binding, some remain loopy; some proteins are almost entirely unstructured, others have only short unstructured regions (Demchenko, 2001; Dunker *et al.*, 2001; Fink, 2005; Mohan *et al.*, 2006; Namba, 2001; Romero *et al.*, 2004; Tompa, 2005; Uversky *et al.*, 2000, 2005; Wright and Dyson, 1999; Esnouf *et al.*, 2006). There is no single way to define ‘unstructured regions’. Here, we refer to a region as unstructured if it appears to lack a defined 3D structure by either of the following experimental techniques: circular dichroism spectroscopy (CD), nuclear magnetic resonance spectroscopy (NMR), X-ray crystallography or protein proteolysis. This is a very roundabout way to collect a plethora of phenomena as exercised, in DisProt (Vucetic *et al.*, 2005). However, many unstructured regions are neither covered by DisProt, nor by existing prediction methods (Oldfield *et al.*, 2005a).

1.2 Unstructured regions are important for biomedicine

Proteins with unstructured regions are increasingly implicated in important functional activities. Due to their intrinsic adaptability, they participate in many regulatory processes, such as the transcription and translation machineries, signal transduction pathways and macromolecular transport by the nuclear pore complex (Devos *et al.*, 2006; Dunker *et al.*, 2005; Iakoucheva *et al.*, 2002; Romero *et al.*, 1998). Protein regions involved in alternative splicing and transcription are also often unstructured (Liu *et al.*, 2006; Romero *et al.*, 2006). Defective proteins in regulatory processes leading to uncontrolled proliferation of cells have been repeatedly associated with cancer (Hanahan and Weinberg, 2000). Unstructured regions appear to play a critical role in initiating malignant tumors. For instance, translocation of genes that fuse the unstructured N-terminus of CBP (CREB binding protein) with the MLL (mixed lineage leukemia) gene is associated with leukemia (Yang, 2004). A large-scale analysis revealed this to be a frequent theme: cancer-related proteins are significantly enriched in unstructured regions (Iakoucheva *et al.*, 2002). Partially unfolded intermediates can trigger or promote diseases, e.g. some human cataracts are associated with the aggregation of partially unfolded intermediates of H₂D-Cryst (Flaugh *et al.*, 2005). The aggregation of proteins with

*To whom correspondence should be addressed.

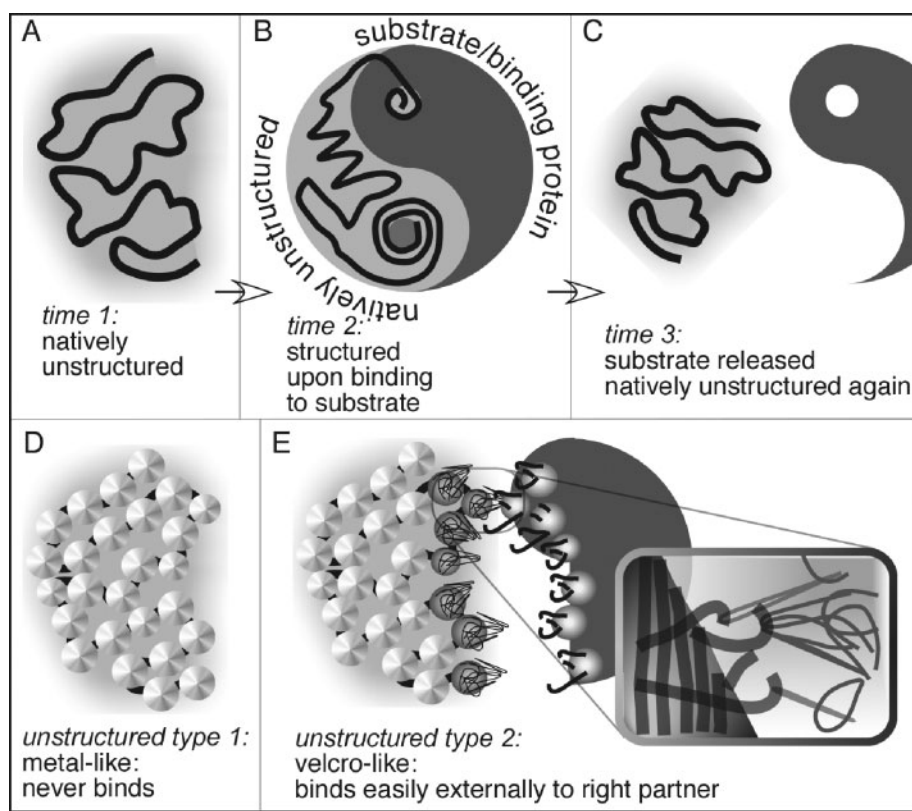


Fig. 1. Unstructured regions may become structured upon binding. A sensitive mechanism of regulation for many cellular processes is facilitated through the existence of proteins with natively unstructured regions. (A) Unstructured regions are often only unstructured in isolation and are therefore, inactive in their native state. (B) Through binding a substrate (which can also be unstructured), many unstructured regions may become ordered/folded, thereby activating a particular function. (C) The interaction between unstructured region and substrate is often reversible, and the complex may dissociate upon small changes in the environment. (D) Some methods that predict unstructured regions resemble the identification of metal-like regions that have low preferences to bind to anything. (E) In contrast, some unstructured regions will have high binding affinity but only for binding externally. In the image, we picture this type as Velcro-like, i.e. the reversible connection between hooks (French: velours) and loops (French: crouchets). In this image the unstructured regions are the loops.

unstructured regions has also been associated with neurodegenerative diseases, e.g. Huntington's disease is directly linked to the aggregation of polyglutamine from CBP (Nucifora *et al.*, 2001).

1.3 Many phenomena, many approaches to prediction

Many concepts are pursued to predict unstructured regions (Fig. S1, Supplementary Material). Some methods utilize machine-learning algorithms to discriminate between residues that appear to be regularly structured and residues that are invisible in the electron density maps from X-ray structures (Cheng *et al.*, 2005; Linding *et al.*, 2003a; Romero *et al.*, 1998; Ward *et al.*, 2004; Yang *et al.*, 2005). Other predictors target the difference in amino acid propensities between unstructured and well-ordered regions. Disordered regions tend to have high net charge, low hydrophobicity (Uversky *et al.*, 2000) and high loop content (Linding *et al.*, 2003b). Methods that implement this idea usually identify long and biologically relevant unstructured regions (Prilusky *et al.*, 2005) and usually miss short unstructured regions (Jin and Dunbrack, 2005). Unstructured regions appear to have lower contact densities [Equation (1)]

than well-structured regions and can therefore be identified by average contact propensities (Dosztanyi *et al.*, 2005a, 2005b; Garbuzynskiy *et al.*, 2004). Such methods use average amino acid contact propensity scores (derived from regular structures) with or without pairwise interaction energy matrices. SCRPRED method uses neural networks to identify clusters of internally contacting residues. This method was found to be useful in identifying residues in long unstructured regions for a few proteins (Dosztanyi *et al.*, 1997; Orosz *et al.*, 2004).

Some methods combine more than one approach. For instance, a neural network trained on residues missing in electron density maps and on residues with high B-factor loops (Linding *et al.*, 2003a). Another example is a meta-method that uses two different prediction methods, each optimized on unstructured regions of different lengths (Peng *et al.*, 2006). The combined methods typically outperform individual approaches on average.

NORS are long regions with no regular secondary structure, i.e. ≥ 70 sequence-consecutive residues depleted of predicted helices and strands that are relatively exposed to the solvent (Fig. S1D, Supplementary Material). Most NORS regions

are unstructured, but many unstructured regions are not NORS (Liu and Rost, 2003; Liu *et al.*, 2002). NORSnet is a new method that succeeded in the distinction between natively unstructured and well-structured loops (Schlessinger *et al.*, 2007). On the one hand, different methods capture different aspects from the plethora of unstructured regions. On the other hand, very few methods capture specific aspects and all methods still miss many long unstructured regions identified experimentally (G.T. Montelione, unpublished data).

Here, we focused on one particular aspect of unstructured regions, namely their intrinsic low contact density [Equation (1)]. It is this intrinsic ‘flexibility’ that makes unstructured regions so adaptable. Not surprisingly then, predictions of unstructured regions based on average contact propensities perform very well (Dosztanyi *et al.*, 2005b; Garbuzynskiy *et al.*, 2004). We hypothesized that a method based on protein-specific internal contact predictions could specifically identify unstructured regions relevant for protein interactions. To test this assumption, we developed a novel method, *Ucon* (prediction of natively unstructured regions through contacts), that identified unstructured regions in DisProt (Vucetic *et al.*, 2005). In our analysis, *Ucon* appeared more accurate than commonly used methods by many measures, and it identified different proteins with unstructured regions.

2 METHODS

2.1 Unstructured proteins

The terminology ‘unstructured proteins’ can be misleading because we do not always have experimental data to establish an entire protein as natively unstructured. Typically, the literature refers to intrinsically disordered or natively unstructured proteins as those that have at least a certain number of residues in unstructured regions. Here, we distinguished between well-structured proteins and proteins with at

least one stretch of ≥ 30 consecutive residues in a predicted unstructured region.

2.2 Contact density

The contact density for a given window centered around residue j was defined as:

$$\rho_j(w) = \frac{1}{(2w+1)} \sum_{i=1}^N \sum_{k=i-w}^{i+w} \delta_{ik}, \text{ with } \delta_{ik} = \begin{cases} 1, & \text{if } ik \text{ in contact} \\ 0, & \text{else} \end{cases} \quad (1)$$

where w stood for a window of w sequence-consecutive residues; these could correspond to a short, local segment in a protein or to the entire protein (with $w = N/2$ and N being the number of residues in a protein). δ_{ik} indicated the spatial relation (contact/not) between two residues. Our method did not use any particular threshold for the definition of a contact, however, the underlying prediction method PROFcon was trained on the standard applied for CASP (Grana *et al.*, 2005), i.e. (i, k) were considered to be in contact if their C-betas were closer than 8 \AA (0.8 nm).

2.3 Basic concept of prediction method

We calculated the contribution of each residue i to the energy (e_i^p) according to the following formula:

$$e_i^p = \frac{1}{2L+1} \sum_{k=i-L}^{i+L} c_{ik}^p \cdot M_{ik} \quad (2)$$

where c_{ik}^p is the contact propensity for the pair (i, k) as predicted by PROFcon and M_{ik} is the interaction energy between the two residues estimated from a particular pairwise interaction potential (below). L is a free parameter that was optimized (below). We calculated e_i^p for each residue in a chain P and smoothed the values by a moving average of length S (below). Unstructured regions were then identified as peaks in this profile (Fig. 2). In practice, we defined a threshold T above which a residue was labeled as unstructured.

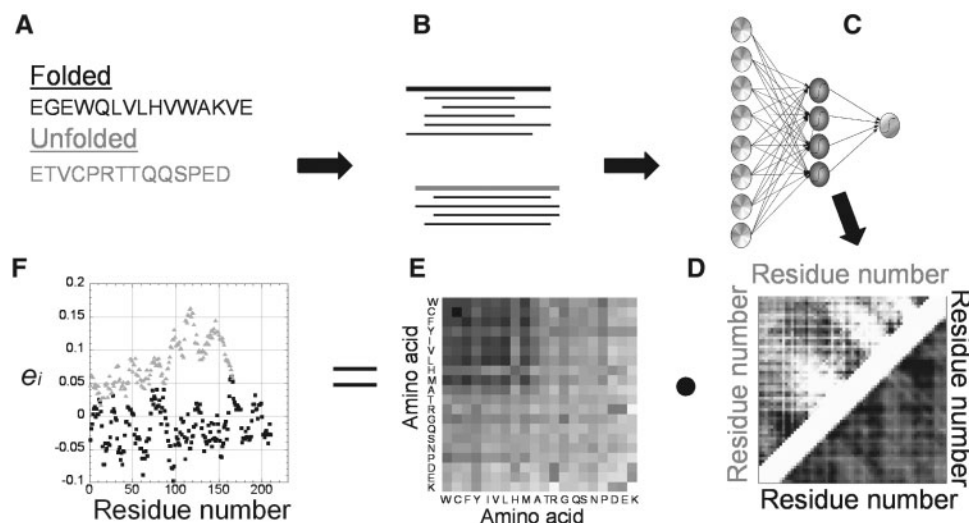


Fig. 2. Schematic representation of the prediction method. We run each protein against the database with PSI-BLAST in order to create position-specific profiles (A, B). The profiles, along with other sequence-derived information such as predicted secondary structure and solvent accessibility, constitute the input to the neural network-based contact prediction method, PROFcon (C) that predicts 2D contact maps (D). Each dot in the 2D-map represents a predicted residue–residue interaction. The darker the dot, the higher the probability of the corresponding two residues is to interact. Next, we multiply the 2D-maps with energy-like statistical potential defined by Jernigan and Miyazawa (E) to derive a position-specific score creating a profile for each sequence (F).

2.4 Predicting the contact matrix

Contact propensities were predicted by PROFcon (Punta and Rost, 2005a), one of the best methods at CASP6 (Grana *et al.*, 2005). Upon submission of the protein sequence, PROFcon (using features such as evolutionary profiles, predicted secondary structure and solvent accessibility) returns for each residue i a list of predicted contact propensities c_{ik} ($k=1, N$; N : protein length and $|k-i|>5$) between 0 and 1. We used these propensities to predict a contact matrix for each protein (Figure 2A–D).

2.5 Pairwise interaction potential

Statistical pairwise contact potentials are knowledge-based quantities that represent the interactions between amino acid types. Many groups obtain energy-like values through Boltzmann relations of observed pairwise contact frequencies. Over 30 such potentials exist; many appear similar (Pokarowski *et al.*, 2005) to at least one of the Miyazawa–Jernigan potentials (Miyazawa and Jernigan, 1999). We therefore used the latter, and also tested a potential (Thomas and Dill, 1996) that was used to predict unstructured regions (Dosztanyi *et al.*, 2005b).

2.6 Data sets

As positives, we used proteins with unstructured regions from DisProt (version 3.0) (Vucetic *et al.*, 2005). This set included proteins that have experimentally verified, biologically relevant unstructured regions. We optimized our method on a subset of DisProt that included proteins with unstructured regions with ≥ 30 residues. As negatives (well-structured), we chose PDB proteins that were used to test PROFcon, i.e. none of the proteins had been used for its development (Punta and Rost, 2005a). We excluded structures with backbone atoms missing from the middle of the chain and structures that had unstructured regions in the termini (>15 residues within C- or N-term). Since PROFcon currently flunks on proteins that are longer than 550 residues, we discarded these proteins from our sets.

2.7 Testing/cross-validation

UniqueProt (Mika and Rost, 2003) generated sequence-unique subsets: the maximal similarity between any protein used for training and testing was an HSP-value <10 (Rost, 1999; Sander and Schneider, 1991), e.g. $<31\%$ pairwise sequence identity for >250 aligned residues. Alignments were generated by three iteration PSI-BLAST (Altschul *et al.*, 1997) versus UniProt with a protocol established earlier (Przybylski and Rost, 2002). Our final cross-validation set included 174 proteins with at least one unstructured region longer than 30 (174 was a subset of 243 proteins with unstructured region of any length), and 223 structured proteins.

In order to optimize the free parameters (T , S , L), we performed a 5-fold cross-validation: we randomly divided our data into five sets (no protein pair was sequence-similar between training and test set). We chose the parameters that maximized the area under the ROC curve (AUC) on 4/5 of the data, and tested on the remaining 1/5 (*test set*). We rotated five times and averaged over the five test sets (Table S1, Supplementary Material).

Of all methods that we assessed only FoldIndex gives a per-protein prediction (i.e. *decides* whether or not a protein is predicted to be *unstructured*). Therefore, we had to introduce a criterion to assess RONN, DISOPRED2, IUPred and NORSnet. We did this in the same way in which we optimized our method: first, we labeled as positives (*proteins with unstructured regions*) all proteins for which the method identified at least 30 consecutive residues as unstructured (*other methods*, Table S1, Supplementary Material, Fig. 3A) at a given threshold. Second, we identified the minimal number of consecutive

residues predicted to be unstructured that optimized the AUC for each method for training (4/5 of data) and recorded the results for testing (1/5 of data). As some methods (RONN, NORSnet) performed better without optimization, we provided both values (see Table S1, Supplementary Material).

2.8 Performance measures

We measured accuracy/specificity (Acc), coverage/sensitivity (Cov) and false positive (FP) rate by the standard formulas:

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \text{Cov} = \text{TPrate} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \text{FPrate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (3)$$

TP are the true positives (proteins with unstructured regions experimentally observed AND correctly predicted); FP are the false positives (structured regions that are predicted to be unstructured); TN are the true negatives (observed and predicted as well-structured) and FN are the false negatives (observed to be unstructured and predicted to be structured). In analogy, we computed the accuracy and coverage for the negatives, i.e. proteins that do not have regions with 30 or more consecutive residues that are unstructured:

$$\text{Acc}_{\text{neg}} = \frac{\text{TN}}{\text{TN} + \text{FN}}; \text{Cov}_{\text{neg}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

We compiled two other frequently used measures, namely the two-state accuracy Q_2 (percentage of proteins correctly predicted in either of the two groups: proteins with and proteins without unstructured regions) and the arithmetic average over Acc and Acc_{neg} (*Average accuracy*). In contrast to the measures presented in Equation (3) and Equation (4), these values thrive at simultaneously reflecting all aspects of expected performance:

$$\text{Average accuracy} = \frac{\text{Acc} + \text{Acc}_{\text{neg}}}{2}; Q_2 = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{TN}} \quad (5)$$

Receiver–operator curves (ROCs) were constructed by calculating FP and TP rates at different thresholds defining a positive prediction. The curves were then integrated in order to calculate the area under the curve (AUC).

3 RESULTS

3.1 Predicted contacts capture natively unstructured regions

PROFcon predicted protein-specific internal contact maps that captured the low contact density [Equation (1)] of natively unstructured regions. Particular examples suggested that PROFcon alone somehow discriminates well-ordered and unstructured regions (Fig. S2, Supplementary Material). To test this hypothesis, we calculated a contact-based ‘unstructured propensity’ for each residue in our data set (e'_i , [Equation (2)] with $M_{ij}=1$). Parameter optimization under 5-fold cross-validation (see Methods section) performed substantially better than random (Fig. 3A, Table S1, Supplementary Material). Performance was significantly higher, when considering only the strongest PROFcon predictions [Equation (2), $M_{ij}=1$, $c_{ij}>0.5$].

3.2 Statistical potential combined with predicted contacts improves performance

The energetic stability of a particular region ultimately determines whether or not that region is unstructured.

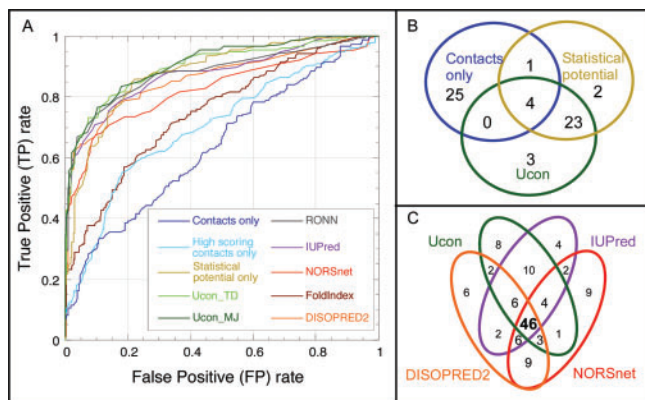


Fig. 3. Assessing the contribution of different sources of information. (A) Comparison between five internal methods [dark blue—PROFcon predictions alone; light blue—high-probability PROFcon predictions only; yellow—Miyazawa–Jernigan potential; green—our final combined method, *Ucon*; *Ucon_MJ* (dark green) and *Ucon_KD* (light green) differed in the potential] and five external methods (gray—RONN; purple—IUPred; red—NORSnet; brown—FoldIndex and orange—DISOPRED2). *Ucon* performed significantly best (Table S1, Supplementary Material). (B) We compared the 30 proteins with the strongest signal for long unstructured regions by the three of methods presented (numbers in circles mutually exclusive). While *Ucon* (green) and the method that used only the pairwise Miyazawa–Jernigan statistical potential (yellow) yielded similar results, the method that utilized only predicted contact propensity (blue) identified very different proteins. (C) We compared the 80 proteins that gave the strongest signal to have long unstructured region by four methods. NORSnet and *Ucon* appeared the most orthogonal identifying 9 and 8, respectively unique proteins that were not identified by any other method. Conversely, the pairs of most overlapping methods were IUPred–*Ucon* and DISOPRED2–NORSnet, respectively sharing 66 and 64 proteins.

Therefore, we weighted each predicted contact [c_{ik} in Equation (2)] with an energy term (M_{ik}) representing the specific contribution of that predicted interaction to stability. This sequence-based score, which combined the intra-chain contacts explicitly predicted by PROFcon [c_{ik} , in Equation (2)] with statistical pairwise potentials [M_{ik} in Equation (2)], clearly discriminated between well-structured and unstructured regions (Fig. S3, red and blue curves, Supplementary Material).

3.3 Final method performed best

When using the combined score to discriminate unstructured from well-structured regions, the results were considerably better than when using predicted contacts alone (Fig. 3A). Furthermore, *Ucon* appeared more accurate than the best state-of-the-art prediction methods for unstructured regions yielding AUC=0.912 (Fig. 3A, Table S1, Supplementary Material).

On the one hand, position-specific contact predictions provided an edge over other approaches. On the other hand, methods based on statistical potentials calculating the per-residue propensity for unstructured regions by simply assuming equal contact probability within a sliding window (Dosztanyi *et al.*, 2005b; Prilusky *et al.*, 2005) performed very well. FoldIndex, for instance, which uses a simple one-body

hydrophobic potential (Kyte and Doolittle, 1982) divided by the net-charge contribution, yielded an AUC of 0.747; IUPred using a pairwise energy potential specifically optimized to capture unstructured regions yielded an AUC=0.878.

Methods that utilize machine-learning algorithms to learn the features of residues that do not have coordinates in the PDB, usually predict regions in DisProt rather accurately (Dosztanyi *et al.*, 2005b; Peng *et al.*, 2006). This surprises since the two data sets (development on PDB, test on DisProt) differ, e.g. in amino acid composition and in the length of unstructured regions (Radivojac *et al.*, 2004). On our set, DISOPRED2 and RONN also performed very well (AUC of 0.868 and 0.887, respectively).

3.4 Specific contact maps identified different proteins

We looked at the 30 correct predictions that gave the strongest signal for long unstructured regions by several prediction methods. In particular, we compared predictions by the method only using statistical contact potentials (potentials-only), by the method based only on predicted PROFcon contact maps (PROFcon-only) and by the merger of the two, namely *Ucon* (see Methods section). PROFcon-only identified different proteins than other methods: 25 of the 30 proteins (83%) were only predicted by using contact maps alone (Fig. 3B).

Next, we compared *Ucon* to three methods based on different concepts (Fig. 3C). Each method identified its list of 80 proteins with strongest signal for *unstructured*. Note that for all these methods the cutoff that resulted in 80 true positives also yielded $\geq 95\%$ accuracy on this set, i.e. the 80 were highly reliable predictions. This analysis revealed several points: first, only 46 of the 80 proteins (<60%) were identified by all methods. This suggested that the methods focused on different aspects of unstructured regions. Second, the smallest overlap between any pair of prediction methods was between NORSnet and *Ucon* (54 proteins). As NORSnet focuses on the identification of natively unstructured loops, this result suggested that *Ucon* identified unstructured regions that were most different from unstructured loops. In contrast, IUPred and *Ucon* had the highest overlap (66 of 80 proteins). Since both methods estimate an energy-related score, for many of the overlapping proteins the energetic contribution might be the main determining factor for the lack of regular structure. For instance, proteins such as *DNA repair protein XPAC*, *Stathmin* and *Nucleoplasmin protein* have long stretches of destabilizing charged residues. Although *Ucon* and IUPred overlapped, they still differed for 14 of the 80 proteins (17.5%), i.e. even these rather similar methods differed importantly. We did not find a statistically significant trend in the DisProt function annotations for the 80 proteins.

3.5 Case study: MAX

MAX is a sequence-specific transcription factor that binds DNA and activates or represses, depending on its binding partner (Patikoglou and Burley, 1997). While *MAX* is unstructured in isolation, it adopts a helix-loop-helix leucine zipper (bHLH-LZ) fold upon binding to DNA and to its target protein (PDB identifier: 1AN2 (Ferre-D’Amare *et al.*, 1993), Fig. S4A, Supplementary Material). Since unstructured regions

that bind DNA are often enriched in charged residues, prediction methods based on statistical potentials (such as IUPred and FoldIndex, Fig. S5, Supplementary Material) easily identify these regions as unstructured. In contrast, methods that are trained on missing residues might fail as this protein contains residues that appear in the PDB (in its bound form). Indeed, DISOPRED2 (Ward *et al.*, 2004), missed most of the unstructured regions in *MAX* (Fig. S6, Supplementary Material). Conversely, PROFcon predicted *MAX* to have few high-probability internal contacts suggesting that the helices do not interact internally and that *MAX* may need to bind to an external target in order to adopt regular structure (Fig. S4, Supplementary Material). Our final method, Ucon, that combined PROFcon output with the statistical potential, indeed, correctly identified this helical region as unstructured (Fig. S4C, Supplementary Material).

MAX becomes helical upon binding DNA; this is representative for many *unstructured*->*well-structured* transitions (Fuxreiter *et al.*, 2004). PROFsec predicted 30% of the DisProt residues as helical (data not shown). Unstructured regions that undergo disorder-order transition have slightly more helix (35–36%), in fact, their helix content resembles that of well-structured regions (Fuxreiter *et al.*, 2004). When we applied our new method that exclusively relied on predicted contacts (PROFcon-only) to create a list of the most likely unstructured residues (at an estimated level of 85% accuracy), we found that 35% of these residues were predicted in helices. This was another indication that our new method Ucon captured ‘structured’ proteins with unstructured regions. In other words, the differential between the two numbers (30/35%) are the candidates for the difference between never-bind (Fig. 1D) and Velcro-like binding (Fig. 1E).

3.6 Case study: unstructured region in protein–protein interaction

Myosin is a primary protein involved in muscle contraction. Its regulatory domain is composed of a long (residues 765–832 in SWISSPROT Identifier MYS_AEQIR) helical stretch named the *Lever arm* (Fig. 4A). Experimental evidence suggests that the Lever arm is natively unstructured (Houdusse *et al.*, 1999; Risal *et al.*, 2004). It has also been identified as a Molecular Recognition Feature (MoRF), i.e. it becomes structured upon binding to its target (Mohan *et al.*, 2006). The Lever arm has relatively few charged residues, and interacts with its two binding partners through hydrophobic interfaces: according to the *Protein–Protein Interaction server* (<http://www.biochem.ucl.ac.uk/bsm/PP/server/>), 65% of the interfaces between the Lever arm and its binding partners are hydrophobic. Since most statistical potential-based methods tend to take buried hydrophobic residue as indicators of *well-structured*, they likely miss this region (Fig. S7, Supplementary Material). In contrast, PROFcon uses information such as predicted secondary structure and solvent accessibility; it considers the fact that the long helix does not have a break (that can lead to packing) and has surface exposed hydrophobic residues. Thus, the PROFcon-based contact-only method [$M_{ij}=1$, Equation (2)], predicted this region is to be unstructured because it will bind externally rather than internally. The Lever arm was exactly the

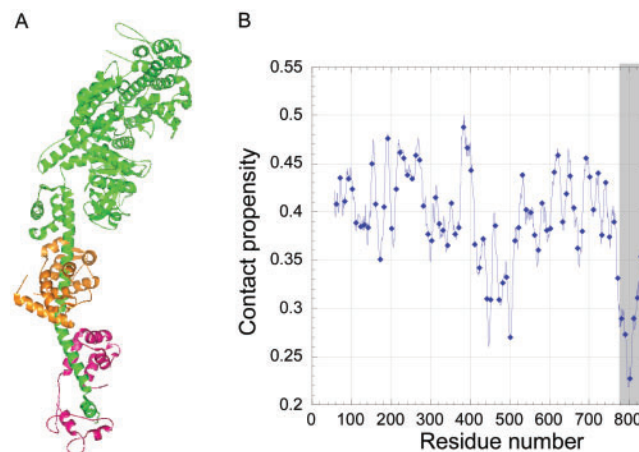


Fig. 4. Low intra chain contact propensities for the lever arm of Myosin striated muscle. PROFcon can capture unstructured regions that bind other proteins. (A) The complex structure of Myosin striated muscle heavy chain (in green) bound to Myosin essential light chain (orange) and Myosin regulatory light chain (magenta) has been determined by Cohen *et al.* [PDB identifier 1SR6 (Risal *et al.*, 2004)]. The interface of the heavy chain with the two other molecules is mediated through a long extended helix (the lever arm) and contains many hydrophobic interactions. (B) PROFcon score has been translated to a 1D score [Equation (2), $M_{ij}=1$] reflecting the propensity of a residue to be internally in contact with other residues. The residues in the lever arm (in gray) have very low contact propensity, thus, due to the fact that they are rather hydrophobic they are likely to bind outside. Note that we omitted the first 55 residues of the protein due to the fact that PROFcon is unable to handle sequences that are extremely long.

type of example that we hoped to identify by the differential analysis of the methods that we introduced here. However, so far, we have not found any other example for which we can verify the prediction.

4 DISCUSSION

4.1 Why do unstructured regions have fewer contacts?

Natively unstructured regions have unusually low contact densities. This appears to be due to several factors. (1) Unstructured regions are deficient in hydrophobic residues, (2) they are enriched in residues such as glycine and proline that break helices or strands, and (3) they have high net charges (Fuxreiter *et al.*, 2004; Radivojac *et al.*, 2004; Tompa, 2005; Uversky *et al.*, 2000; Vucetic *et al.*, 2003). These properties may prevent proteins with unstructured regions from folding independently through mechanisms such as *hydrophobic collapse*, the formation of regular secondary structures (*nucleation*), or their combination [*nucleation condensation* (Fersht and Daggett, 2002)]. The interactions between unstructured regions and their external binding partners often result in the formation of regular secondary structure [mainly helices (Fuxreiter *et al.*, 2004; Oldfield *et al.*, 2005b)] and in the cancellation of destabilizing charges.

However, even in their bound, well-structured form, unstructured regions tend to adopt unusual, relatively *loopy* conformations (Fuxreiter *et al.*, 2004; Mohan *et al.*, 2006).

We have shown (Schlessinger *et al.*, 2007) that the previously observed (Dunker *et al.*, 2005; Mohan *et al.*, 2006; Patil and Nakamura, 2006) abundance of unstructured regions in proteins with many interaction partners (hubs) is more extensive for loopy regions than for the type of unstructured region picked up by IUPred.

Position-specific contacts were overall less successful in identifying unstructured regions than potential-based propensities. Despite the fact that Ucon, the combination of the two, was considerably better than each individual method, we were surprised by the relative advantage of propensities over specific contacts. One reason may be that PROFcon does not predict local contacts and thereby misses the important energetic contributions of helices. Although PROFcon was successfully applied to a similar task (Punta and Rost, 2005b), it could be that PROFcon predictions are just not accurate enough for unstructured regions. Contact density predictions for unstructured regions might be a solution.

4.2 How well do we predict unstructured regions today?

Comparing our cross-validated Ucon to methods that have been developed on largely overlapping data sets likely overestimates the performance for others. Nevertheless, Ucon performed best amongst publicly available methods that were shown to be highly accurate on DisProt. The difference between Ucon and the best runner up not developed in our group may appear small, but it exceeded the levels that distinguished winners from runner-ups at CASP7 (Bordoli *et al.*, 2006), and it did so on much larger sets of positives.

DisProt captures only some aspects of unstructured regions. For instance, CASP focuses on a very different aspect, namely residues not visible in electron density maps from X-ray crystallography. This concept, originally introduced by Keith Dunker (Dunker *et al.*, 1998), has many limitations but it clearly is the only completely automated and somehow *objective* definition that creates large data sets. Irrespectively of the additional filter applied to the minimal length of a region that qualifies as disorder, data from otherwise well-structured proteins is dominated by short regions. Ucon will most likely fail for regions shorter than 10 residues, because PROFcon is optimized to predict long-range contacts. Over 80% of the disorder residues considered in CASP7 were in regions with ≤ 10 residues (Bordoli *et al.*, 2006). In contrast, unstructured regions that actually prevent folding into a folded 3D structure in isolation are typically considerably longer than this.

New experimental techniques based on NMR are about to provide the first detailed, objective and large-scale data sets on these types of regions (G.T. Montelione, unpublished data). Preliminary tests (Schlessinger *et al.*, 2007) suggest that even methods for the DisProt type of unstructured regions rather than the CASP-type are significantly less successful when assessed in light of these new data. Clearly, the complete universe of unstructured regions is yet unknown and the regions mapped out today are already extremely varied. We will need different methods for different aspects. Ucon targets the identification of long unstructured regions, and appears to be most successful at this.

The population of experimentally characterized unstructured regions is extremely heterogeneous (Dyson and Wright, 2005). It is therefore not surprising that different methods focus on different *flavors* of unstructured regions (Vucetic *et al.*, 2003). An extreme case is NORSnet that focuses on the identification of unstructured loops (Schlessinger *et al.*, 2007). The second most unique method was Ucon. The observation that the smallest overlap between two prediction methods outputs was between NORSnet and Ucon (54 proteins), coupled with the observation that unstructured regions identified by Ucon have high predicted helix content suggested that Ucon identified unstructured regions that become well-structured upon binding. In summary, Ucon adds three important virtues to the pool of prediction methods: it is highly accurate, quite orthogonal to other methods and it enables some specific interpretation of the meaning of its differences to methods such as NORSnet.

4.3 Even low-accuracy predictions can be useful

Although predictions of inter-residue contacts have been improving over the years, many researchers continue to perceive contact predictions as relatively inaccurate (Grana *et al.*, 2005). However, Ucon is not the first example of a successful application of contact predictions to protein structure and function prediction (Orosz *et al.*, 2004; Ortiz *et al.*, 1999; Pazos *et al.*, 1999; Punta and Rost, 2005b). One of the most interesting aspects of this particular application might be that Ucon could only succeed because PROFcon predicted many specific long-range contacts correctly. Another evidence for the usefulness of contact predictions was that we could correctly identify unstructured regions using contact predictions alone; some of those were not identified by any other method that we tested.

5 CONCLUSIONS

We introduced the combination of two unique approaches to create Ucon, a new method for the prediction of unstructured regions. Ucon compared favorably with methods utilizing either one of these approaches alone for proteins with long (>30 residues) unstructured regions. We remained most surprised by the result that methods based on position-specific and position-independent preferences performed similarly on average. A position-independent method only depends on amino acid composition, i.e. is 'blind' to the specific positions in the sequence (e.g. the sequence AGEREG gives the same preference as does REGGAE). Such a simplification obviously ignores the importance of folding pathways that we know matter greatly. The explanation for the minute difference between these two methods might be that there are a great variety of unstructured regions. Some serve as buffering or filling material. These regions just have to be selected to stay clear off binding to anything. Other unstructured regions, in contrast, have strong binding preferences; however, they are selected not to bind internally but to bind through external transient protein-protein interactions. We provided evidence that our method identified different proteins with unstructured regions than existing methods. Furthermore, we showed that at least for one single example we could specifically identify

regions involved in external protein–protein interactions. Thus, this method might become rather useful for the prediction of protein function, as well as, for more detailed experimental studies of natively unstructured regions.

ACKNOWLEDGEMENTS

Thanks to Lawrence Shapiro, Barry Honig (Columbia) and Mickey Kosloff (Duke) for discussions; to Andrew Kernytsky (Columbia) for comments on the manuscript; to Jinfeng Liu and Guy Yachdav (Columbia) for computer assistance and to Dariusz Przybylski (Columbia) for preliminary information and programs. This work was supported by grants from the National Library of Medicine (NLM, RO1-LM07329), by a grant from the Protein Structure Initiative of the US National Institutes of Health to the Northeast Structural Genomics Consortium (U54-GM074958) and by the grant U54-GM072980 from the NIH. Last, not least, thanks to Keith Dunker (DisProt, Indiana University), and Phil Bourne (PDB, San Diego University), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available. Funding to pay the Open Access publication charges was provided by U54-GM072980 from the US National Institutes of Health.

REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bordoli,L. *et al.* (2006) *Assessment of Disorder Prediction CASP7*. Asilomar, CA, USA.
- Cheng,J. *et al.* (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining Knowl. Discov.*, **11**, 213–222.
- Demchenko,A.P. (2001) Recognition centers in proteins: induced and assisted folding. *J. Mol. Recognit.*, **14**, 42–61.
- Devos,D. *et al.* (2006) Simple fold composition and modular architecture of the nuclear pore complex. *Proc. Natl Acad. Sci. USA*, **103**, 2172–2177.
- Dosztanyi,Z. *et al.* (1997) Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.*, **272**, 597–612.
- Dosztanyi,Z. *et al.* (2005a) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
- Dosztanyi,Z. *et al.* (2005b) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
- Dunker,A.K. *et al.* (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.*, **3**, 473–484.
- Dunker,A.K. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.
- Dunker,A.K. *et al.* (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.*, **272**, 5129–5148.
- Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, **12**, 54–60.
- Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
- Esnouf,R.M. *et al.* (2006) Honing the in silico toolkit for detecting protein disorder. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 1260–1266.
- Ferre-D'Amare,A.R. *et al.* (1993) Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature*, **363**, 38–45.
- Fersht,A.R. and Daggett,V. (2002) Protein folding and unfolding at atomic resolution. *Cell*, **108**, 573–582.
- Fink,A.L. (2005) Natively unfolded proteins. *Curr. Opin. Struct. Biol.*, **15**, 35–41.
- Flaugh,S.L. *et al.* (2005) Interdomain side-chain interactions in human gammaD crystallin influencing folding and stability. *Protein Sci.*, **14**, 2030–2043.
- Fuxreiter,M. *et al.* (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.*, **338**, 1015–1026.
- Garbuzynskiy,S.O. *et al.* (2004) To be folded or to be unfolded? *Protein Sci.*, **13**, 2871–2877.
- Grana,O. *et al.* (2005) CASP6 assessment of contact prediction. *Proteins*, **61** (Suppl. 7), 214–224.
- Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
- Houdusse,A. *et al.* (1999) Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head. *Cell*, **97**, 459–470.
- Iakoucheva,L.M. *et al.* (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.
- Jin,Y. and Dunbrack,R.L., Jr (2005) Assessment of disorder predictions in CASP6. *Proteins*, **61** (Suppl. 7), 167–175.
- Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
- Linding,R. *et al.* (2003a) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
- Linding,R. *et al.* (2003b) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
- Liu,J. and Rost,B. (2003) NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.*, **31**, 3833–3835.
- Liu,J. *et al.* (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
- Liu,J. *et al.* (2006) Intrinsic disorder in transcription factors. *Biochemistry*, **45**, 6873–6888.
- Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.
- Miyazawa,S. and Jernigan,R.L. (1999) Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. *Proteins*, **36**, 347–356.
- Mohan,A. *et al.* (2006) Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.*, **362**, 1043–1059.
- Namba,K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells*, **6**, 1–12.
- Nucifora,F.C., Jr *et al.* (2001) Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity. *Science*, **291**, 2423–2428.
- Oldfield,C.J. *et al.* (2005a) Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.
- Oldfield,C.J. *et al.* (2005b) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.
- Orosz,F. *et al.* (2004) TPPP/p25: from unfolded protein to misfolding disease: prediction and experiments. *Biol. Cell*, **96**, 701–711.
- Ortiz,A.R. *et al.* (1999) Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins: Struct. Funct. Genet.*, **37** (Suppl. 3), 177–185.
- Patikoglou,G. and Burley,S.K. (1997) Eukaryotic transcription factor-DNA complexes. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 289–325.
- Patil,A. and Nakamura,H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.*, **580**, 2041–2045.
- Pazos,F. *et al.* (1999) A platform for integrating threading results with protein family analyses. *Bioinformatics*, **15**, 1062–1063.
- Peng,K. *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
- Pokarowski,P. *et al.* (2005) Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins*, **59**, 49–57.
- Prilusky,J. *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.
- Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins: Struct. Funct. Genet.*, **46**, 195–205.
- Punta,M. and Rost,B. (2005a) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.
- Punta,M. and Rost,B. (2005b) Protein folding rates estimated from contact predictions. *J. Mol. Biol.*, **348**, 507–512.
- Radivojac,P. *et al.* (2004) Protein flexibility and intrinsic disorder. *Protein Sci.*, **13**, 71–80.

- Risal,D. et al. (2004) Myosin subfragment 1 structures reveal a partially bound nucleotide and a complex salt bridge that helps couple nucleotide and actin binding. *Proc. Natl Acad. Sci. USA*, **101**, 8930–8935.
- Romero,P. et al. (1998) Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.*, **3**, 437–448.
- Romero,P. et al. (2004) Natively disordered proteins: functions and predictions. *Appl. Bioinformatics*, **3**, 105–113.
- Romero,P.R. et al. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl Acad. Sci. USA*, **103** (22), 8390–8395.
- Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
- Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
- Schlessinger,A. Liu J, Rost B. (2007) Natively unstructured loops differ from other loops. *PLoS Computat. Biol.*, **3**, e140. doi:10.1371/journal.pcbi.0030140.
- Thomas,P.D. and Dill,K.A. (1996) An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
- Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.
- Uversky,V.N. et al. (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Genet.*, **41**, 415–427.
- Uversky,V.N. et al. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.*, **18**, 343–384.
- Vucetic,S. et al. (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.
- Vucetic,S. et al. (2005) DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137–140.
- Ward,J.J. et al. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
- Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
- Yang,X.J. (2004) The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases. *Nucleic Acids Res.*, **32**, 959–976.
- Yang,Z.R. et al. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.