doi:10.1016/j.jmb.2005.01.071 J. Mol. Biol. (2005) 347, 827–839

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins

Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon*

Institute of Enzymology
Biological Research Center
Hungarian Academy of Sciences
1518 Budapest, PO Box 7
Hungary

The structural stability of a protein requires a large number of interresidue interactions. The energetic contribution of these can be approximated by low-resolution force fields extracted from known structures, based on observed amino acid pairing frequencies. The summation of such energies, however, cannot be carried out for proteins whose structure is not known or for intrinsically unstructured proteins. To overcome these limitations, we present a novel method for estimating the total pairwise interaction energy, based on a quadratic form in the amino acid composition of the protein. This approach is validated by the good correlation of the estimated and actual energies of proteins of known structure and by a clear separation of folded and disordered proteins in the energy space it defines. As the novel algorithm has not been trained on unstructured proteins, it substantiates the concept of protein disorder, i.e. that the inability to form a well-defined 3D structure is an intrinsic property of many proteins and protein domains. This property is encoded in their sequence, because their biased amino acid composition does not allow sufficient stabilizing interactions to form. By limiting the calculation to a predefined sequential neighborhood, the algorithm was turned into a position-specific scoring scheme that characterizes the tendency of a given amino acid to fall into an ordered or disordered region. This application we term IUPred and compare its performance with three generally accepted predictors, PONDR VL3H, DISOPRED2 and GlobPlot on a database of disordered proteins.

Keywords: intrinsically unstructured proteins; prediction of disordered proteins; low-resolution force fields; interresidue interactions; foldability

**Corresponding author*

Introduction

Intrinsically unstructured/disordered proteins/domains (IUPs), such as p21,[1] the N-terminal domain of p53[2] or the transactivator domain of CREB,[3] exist in a largely disordered structural state, yet they carry out basic cellular functions.[4–7] Their existence defies the classical structure–function paradigm, founded on the tenet that a well-defined 3D structure is the prerequisite of protein function. The importance of protein disorder, nevertheless, is underlined by its prevalence in various proteomes[8,9] and by its correlation with basic functional modes, such as signal transduction and transcriptional regulation.[9,10]

The identification of IUPs thus far proceeded by collecting scattered data obtained with a range of experimental techniques. As a result, available datasets are rather limited in size and are heterogeneous in terms of experimental conditions, techniques and interpretation of data. They also lack consistency, due to the absence of clear conceptual and operational definition(s) of structural disorder. All these result in false positive and false negative classifications, i.e. the inclusion of ordered segments in disorder databases and the exclusion (and inclusion in ordered reference databases) of disordered proteins/segments. Furthermore, the databases are also biased due to the overrepresentation of a few experimental techniques, such as X-ray crystallography, NMR

Abbreviation used: IUP, unstructured/disordered proteins/domains.

E-mail address of the corresponding author: simon@enzim.hu

and CD. As each technique probes different aspects of protein structure, they do not necessarily correctly identify disorder. For example, loopy proteins, which have no repetitive secondary structure,[11] would appear disordered by CD but ordered by the other techniques. With NMR, disorder often is concluded from poor signal dispersion, which does not distinguish between random coils and molten globules of high potential to fold in the presence of a partner. In X-ray crystallography, crystal packing may enforce certain disordered regions to become ordered, and disordered binding segments are often crystallized in complex with their partner and are classified ordered despite their lack of structure in isolation. In addition, wobbly domains would appear disordered, despite their intrinsic structural order. In consequence, predictors trained on these datasets for assessing disorder[5,9] reflect these uncertainties.

The basis of predicting protein disorder is the difference in sequence characteristics between folded and disordered proteins. Typically, IUPs exhibit a strong bias in their amino acid composition and even a reduced alphabet is able to recognize them at the level of complete sequences.[12] Other results indicate, however, that there are differences in sequence properties among different types of disordered proteins.[13] Various factors have been suggested to be important in terms of protein disorder, including flexibility, aromatic content,[14] secondary structure preferences[15] and various scales associated with hydrophobicity.[14,16] Beside low mean hydrophobicity, high net charge was also suggested to contribute to disorder.[17] All these different analyses, though, hint that the amino acid composition of IUPs results in their inability to fold due to the depletion of typically buried amino acid residues and enrichment of typically exposed amino acid residues,[5] which implies that globular proteins have specific sequences with the potential to form a sufficiently large number of favorable interactions, whereas IUPs do not. Here, we attempt to put this inference on a quantitative footing by taking an energetics point of view. On this ground, the sequences encoding for globular proteins and IUPs can be distinguished.

For globular proteins, the contribution of interresidue interactions to total energy is often approximated by low-resolution force fields, or statistical potentials, energy-like quantities derived from globular proteins based on the observed amino acid pairing frequencies.[18,19] In deriving the actual potentials, different principles have been applied.[18,20–23] The resulting empirical energy functions are well suited to assess the quality of structural models[24] and have been used for fold recognition or threading,[25,26] but also in docking,[27] *ab initio* folding,[28] or predicting protein stability.[29] Their success in a wide range of applications suggests the existence of a common set of interactions, simultaneously favored in all native, as opposed to alternate, structures.

Our current formulation derives from the general view that the primary structure of a globular protein determines its native conformation, and therefore its energy, which corresponds to the global minimum in conformational space. This energy represents the lowest level attainable by the sequence at the optimum of interresidue interactions. In this work, we introduce a novel approach to predict this optimum energy independently of a presumed structure. By applying this principle to a predefined sequential neighborhood of a particular amino acid in a sequence, this approach can be turned into a position-specific scoring scheme for disorder, termed IUPred. As IUPred has not been trained on potentially erroneous data, its unbiased assessment of the structural status of an unknown sequence/segment is of confirmatory value.

Theory

Estimation of the pairwise energy from amino acid composition

The pairwise energy of a protein in its native state is the function of its conformation as well as its amino acid sequence. The total energy can be calculated by taking all contacts in the protein, and weighting them by the corresponding interaction energy. In our model, the energy depends only on amino acid types, as specified by a 20 by 20 interaction matrix, **M** (see Table 1). The pairwise energy content can be written as:

$$E = \sum_{ij=1}^{20} M_{ij} C_{ij}$$

where M_{ij} is the interaction energy between amino acid types i and j, and C_{ij} is the number of interactions between residues of types i and j in the given conformation.

We approximate E/L, the total energy per amino acid, by means of the protein's amino acid composition. Without considering the actual conformation, we rely on statistics collected from a database of globular proteins. The rationale behind this approach is that the energy contribution of a residue depends not only on its amino acid type, but also on its potential partners in the sequence. We assume that if the sequence contains more amino acid residues that can form favorable contacts with the given residue, its expected energy contribution is more favorable. The simplest formula which describes this relationship is a quadratic expression in the amino acid composition.

Let N_i denote the number of amino acid residues of type i in the sequence and $n_i = N_i/L$ its frequency. The energy per amino acid is approximated by:

$$\frac{E_{\text{estimated}}}{L} = \sum_{ij}^{20} n_i P_{ij} n_j$$

where **P** is the energy predictor matrix, which tells

Table 1. M matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−0.20	−0.44	0.16	0.26	−0.46	−0.26	0.50	−0.57	0.10	−0.36	−0.22	0.07	0.14	0.01	0.20	−0.09	−0.05	−0.42	0.05	−0.50
C	−0.44	−2.99	0.21	0.19	−0.88	−0.34	−1.11	−0.36	−0.09	−0.53	−0.43	−0.52	−0.14	−0.43	−0.24	0.13	−0.22	−0.62	0.24	−0.79
D	0.16	0.21	0.17	0.55	0.38	0.35	−0.23	0.44	−0.39	0.28	0.35	−0.02	1.03	0.49	−0.37	0.19	−0.12	0.69	0.04	0.43
E	0.26	0.19	0.55	0.60	0.55	0.65	0.18	0.37	−0.47	0.33	0.29	0.01	0.69	0.04	−0.52	0.18	0.37	0.39	0.03	0.17
F	−0.46	−0.88	0.38	0.55	−0.94	0.17	−0.40	−0.88	0.01	−1.08	−0.78	0.22	0.20	0.26	−0.19	−0.22	0.02	−1.15	−0.60	−0.88
G	−0.26	−0.34	0.35	0.65	0.17	−0.12	0.18	0.24	0.19	0.24	0.02	−0.04	0.60	0.46	0.50	0.28	0.28	0.27	0.51	−0.35
H	0.50	−1.11	−0.23	0.18	−0.40	0.18	0.42	−0.00	0.79	−0.24	−0.07	0.20	0.25	0.69	0.24	0.21	0.11	0.16	−0.85	−0.26
I	−0.57	−0.36	0.44	0.37	−0.88	0.24	−0.00	−1.16	0.15	−1.25	−0.58	−0.09	0.36	−0.08	0.14	0.32	−0.27	−1.06	−0.68	−0.85
K	0.10	−0.09	−0.39	−0.47	0.01	0.19	0.79	0.15	0.42	0.13	0.48	0.26	0.50	0.15	0.53	0.10	−0.19	0.10	0.10	0.04
L	−0.36	−0.53	0.28	0.33	−1.08	0.24	−0.24	−1.25	0.13	−1.10	−0.50	0.21	0.42	−0.01	−0.07	0.17	0.07	−0.97	−0.95	−0.63
M	−0.22	−0.43	0.35	0.29	−0.78	0.02	−0.07	−0.58	0.48	−0.50	−0.74	0.32	0.01	0.26	0.15	0.48	0.16	−0.73	−0.56	−1.02
N	0.07	−0.52	−0.02	0.01	0.22	−0.04	0.20	−0.09	0.26	0.21	0.32	0.14	0.27	0.37	0.13	0.15	0.10	0.40	−0.12	0.32
P	0.14	−0.14	1.03	0.69	0.20	0.60	0.25	0.36	0.50	0.42	0.01	0.27	0.27	1.02	0.47	0.54	0.88	−0.02	−0.37	−0.12
Q	0.01	−0.43	0.49	0.04	0.26	0.46	0.69	−0.08	0.15	−0.01	0.26	0.37	1.02	−0.12	0.24	0.29	0.04	−0.11	0.18	0.11
R	0.20	−0.24	−0.37	−0.52	−0.19	0.50	0.24	0.14	0.53	−0.07	0.15	0.13	0.47	0.24	0.17	0.27	0.45	0.01	−0.73	0.01
S	−0.09	0.13	0.19	0.18	−0.22	0.28	0.21	0.32	0.10	0.17	0.48	0.15	0.54	0.29	0.27	−0.06	0.08	0.12	−0.22	−0.14
T	−0.05	−0.22	−0.12	0.37	0.02	0.28	0.11	−0.27	−0.19	0.07	0.16	0.10	0.88	0.04	0.45	0.08	−0.03	−0.01	0.11	−0.32
V	−0.42	−0.62	0.69	0.39	−1.15	0.27	0.16	−1.06	0.10	−0.97	−0.73	0.40	−0.02	−0.11	0.01	0.12	−0.01	−0.89	−0.56	−0.71
W	0.05	0.24	0.04	0.03	−0.60	0.51	−0.85	−0.68	0.10	−0.95	−0.56	−0.12	−0.37	0.18	−0.73	−0.22	0.11	−0.56	−0.05	−1.41
Y	−0.50	−0.79	0.43	0.17	−0.88	−0.35	−0.26	−0.85	0.04	−0.63	−1.02	0.32	−0.12	0.11	0.01	−0.14	−0.32	−0.71	−1.41	−0.76

Contact potential derived from 785 proteins using the approach of Thomas & Dill.[20]

how the energy of amino acid i depends on the jth element of the amino acid composition vector. The parameters P_{ij}, applicable for all proteins, are determined by least-squares fitting using globular proteins. The fitting was carried out by treating each amino acid type and the corresponding row in matrix **P** separately, to ensure that the energetic contribution is well approximated for all amino acid types. Using the additivity of the energy of pairwise interactions, we dissect the total energy of the kth protein into amino acid specific contributions $E^k = \Sigma e_i^k$, where e_i^k is the energy of all amino acid residues type i interacting with all other amino acid residues in the sequence. The e_i^k depends on the number of contacts this residue makes with other amino acid residues of type j in the sequence.

$$e_i^k(\text{calculated}) = \sum_{j=1}^{20} M_{ij} C_{ij}^k$$

This quantity is approximated by the expression:

$$e_i^k(\text{estimated}) = N_i^k \sum_{j=1}^{20} P_{ij} n_j^k$$

The parameters of the corresponding row of matrix **P** are obtained by minimizing the function

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2$$

Letting $\partial Z_i / \partial P_{ij} = 0$ for all P_{ij} leads to a set of linear equations which are solved for each amino acid type by using the GSL scientific library. Only the symmetrical part of the matrix is considered, as the anti-symmetrical part is cancelled out in quadratic forms. The resulting **P** is given in Table 2.

Results

Comparison of estimated and calculated energies for globular proteins

The validity of the energy predictor matrix was checked by comparing the energies calculated from amino acid interactions of proteins with a known structure to the energies estimated from their amino acid compositions. The fitting was carried out using 674 proteins from the Glob_list (for the definition of this and other databases, see Materials and Methods), omitting those with high cysteine content (above 9%) as they had unusually favorable energy because of cystine pairs. The calculated energy is given in an arbitrary energy unit [aeu], with more negative values indicating more favorable interactions. Figure 1 shows that there is a clear linear relationship between calculated and estimated energies. The goodness of fit can be characterized by a correlation coefficient and the $r2$ value: $r2 = 1 - SS_{\text{reg}}/SS_{\text{tot}}$, where SS_{tot} and SS_{reg} are the sums of the squares of distances from the mean of the calculated energies, and of estimated and

Table 2. P energy predictor matrix

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	−1.65	−2.83	1.16	1.80	−3.73	−0.41	1.90	−3.69	0.49	−3.01	−2.08	0.66	1.54	1.20	0.98	−0.08	0.46	−2.31	0.32	−4.62
C	−2.83	−39.58	−0.82	−0.53	−3.07	−2.96	−4.98	0.34	−1.38	−2.15	1.43	−4.18	−2.13	−2.91	−0.41	−2.33	−1.84	−0.16	4.26	−4.46
D	1.16	−0.82	0.84	1.97	−0.92	0.88	−1.07	0.68	−1.93	0.23	0.61	0.32	3.31	2.67	−2.02	0.91	−0.65	0.94	−0.71	0.90
E	1.80	−0.53	1.97	1.45	0.94	1.31	0.61	1.30	−2.51	1.14	2.53	0.20	1.44	0.10	−3.13	0.81	1.54	0.12	−1.07	1.29
F	−3.73	−3.07	−0.92	0.94	−11.25	0.35	−3.57	−5.88	−0.82	−8.59	−5.34	0.73	0.32	0.77	−0.40	−2.22	0.11	−7.05	−7.09	−8.80
G	−0.41	−2.96	0.88	1.31	0.35	−0.20	1.09	−0.65	−0.16	−0.55	−0.52	−0.32	2.25	1.11	0.84	0.71	0.59	−0.38	1.69	−1.90
H	1.90	−4.98	−1.07	0.61	−3.57	1.09	1.97	−0.71	2.89	−0.86	−0.75	1.84	0.35	2.64	2.05	0.82	−0.01	0.27	−7.58	−3.20
I	−3.69	0.34	0.68	−1.93	−5.88	−0.65	−0.71	−6.74	−0.01	−9.01	−3.62	−0.07	0.12	−0.18	0.19	−0.15	0.63	−6.54	−3.78	−5.26
K	0.49	−1.38	−1.93	−2.51	−0.82	−0.16	2.89	−0.01	1.24	0.49	1.61	1.12	0.51	0.43	2.34	0.19	−1.11	0.19	0.02	−1.19
L	−3.01	−2.15	0.23	1.14	−8.59	−0.55	−0.86	−9.01	0.49	−6.37	−2.88	0.97	1.81	−0.58	0.60	−0.41	0.72	−5.43	−8.31	−4.90
M	−2.08	1.43	0.61	2.53	−5.34	−0.52	−0.75	−3.62	1.61	−2.88	−6.49	0.21	0.75	1.90	2.09	1.39	0.63	−2.59	−6.88	−9.73
N	0.66	−4.18	0.32	0.20	0.73	−0.32	1.84	−0.07	1.12	0.97	0.21	0.61	1.15	1.28	1.08	0.29	0.46	0.93	−0.74	0.93
P	1.54	−2.13	3.31	1.44	0.32	2.25	0.35	0.12	0.51	1.81	0.75	1.15	−0.42	2.97	1.06	1.12	1.65	0.38	−2.06	−2.09
Q	1.20	−2.91	2.67	0.10	0.77	1.11	−0.58	1.90	1.28	2.97	−1.54	0.91	0.85	−0.07	−1.91	−0.76	0.01	
R	0.98	−0.41	−2.02	−3.13	−0.40	0.84	2.05	0.19	2.34	−0.60	2.09	1.08	1.06	0.91	0.21	0.95	0.98	0.08	−5.89	0.36
S	−0.08	−2.33	0.91	0.81	−2.22	0.71	0.82	−0.15	0.19	−0.41	1.39	0.29	1.12	0.85	0.95	−0.48	−0.06	0.13	−3.03	−0.82
T	0.46	−1.84	−0.65	1.54	0.11	0.59	−0.01	0.63	−1.11	0.72	0.63	0.46	1.65	−0.07	0.98	−0.06	−0.96	1.14	−0.65	−0.37
V	−2.31	−0.16	0.94	0.12	−7.05	−0.38	0.27	−6.54	0.19	−5.43	−2.59	0.93	0.38	−1.91	0.08	0.13	1.14	−4.82	−2.13	−3.59
W	0.32	4.26	−0.71	−1.07	−7.09	1.69	−7.58	−3.78	0.02	−8.31	−6.88	−0.74	−2.06	−0.76	−5.89	−3.03	−0.65	−2.13	−1.73	−12.39
Y	−4.62	−4.46	0.90	1.29	−8.80	−1.90	−3.20	−5.26	−1.19	−4.90	−9.73	0.93	−2.09	0.01	0.36	−0.82	−0.37	−3.59	−12.39	−2.68

The pairwise energy per amino acid is estimated as a quadratic form in the amino acid composition vector using the elements of this matrix.

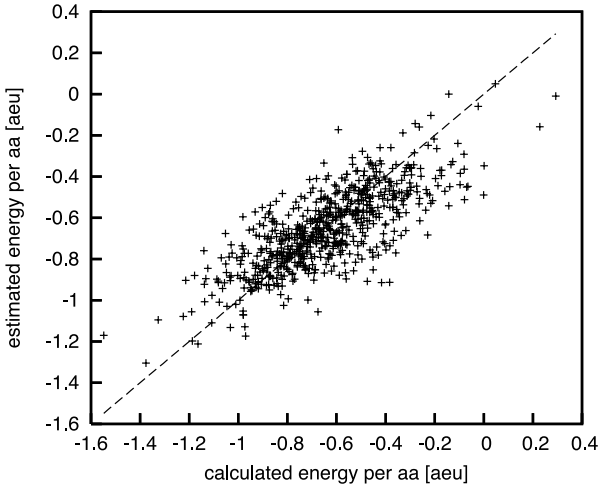

Figure 1. Correlation of estimated and calculated total interaction energies of globular proteins. The total pairwise interaction energy of 674 globular proteins from Glob_list (omitting proteins with high cysteine content), was estimated from their amino acid compositions by a method based on a quadratic formula in the amino acid composition and are shown as a function of the actual energies calculated from their known 3D structures. The energies are in arbitrary energy units, as defined in the text. The broken line represents perfect agreement between the estimated and calculated energy values.

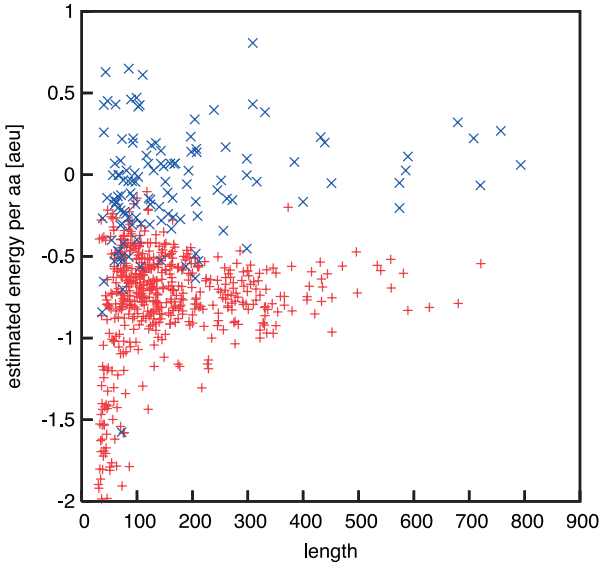

Figure 2. Estimated pairwise interaction energies of globular proteins and IUPs. The total pairwise interaction energy of 559 globular proteins in Filt_Glob_list (red +) and 129 disordered proteins in IUP_list (blue ×) was estimated from their amino acid composition and plotted as a function of their length. Values more negative represent more stabilization due to pairwise amino acid interactions. The average pairwise interaction energy of globular proteins and IUPs are −0.81 and −0.07 [aeu], respectively.

calculated energies, respectively. The value of $r2$ can be between 0, where the average is used as an estimate, and 1, which is the ideal case. It describes how well the variance in the original data is explained by the fitted model using the least-squares approximation. In our case, the $r2$ value was 0.58, and the correlation coefficient was 0.76. Both values indicate a reasonable level of agreement between the estimated and calculated energies.

Pairwise energy content for globular proteins and IUPs

Figure 2 shows the estimated energies for globular proteins and IUPs as a function of their length. For the globular proteins of Filt_Glob_list the average energy is −0.81 [aeu]. The estimated energies of IUPs in IUP_list are less favorable, with an average of −0.07 [aeu]. The separation between the two sets becomes more pronounced for longer sequences, while there is some overlap for shorter sequences. Based on the P-value of 2.2×10^{-16} obtained using the Wilcoxon rank sum test, we can reject the hypothesis that the two sets of energies are from the same distributions. Overall, the difference substantiates our assumption that the pairwise energy content is less favorable for IUPs than for globular proteins.

A corollary to this separation is that the estimated energy content may also distinguish partially ordered IUPs, i.e. molten globules and pre-molten globules, from fully disordered (coil-like) proteins, because the former are expected to have more energetically favorable interactions. To test this assumption, 55 coil-like and 52 pre-molten globule-like proteins have been taken from Table 1 in the work done by Uversky,[30] and their total energy content has been estimated. These datasets, which partially overlap with IUP_list, show a 0.3 [aeu] separation in the average energy content (data not shown). Thus, our approach is able to assess the energetic consequence of the subtle structural differences between fully and partially disordered proteins.

Decomposition

So far, the quadratic form for the estimated energy has been given in the natural coordinate system, each axis corresponding to one amino acid. Now we rotate the coordinate system into the one defined by the eigenvectors of the **P** matrix, in which the expression for the estimated energy is reduced to the diagonal form:

$$E(\text{estimated}) = \lambda_1 p_1^2 + \lambda_2 p_2^2 \cdots + \lambda_{20} p_{20}^2$$

Here λ_i is the ith eigenvalue corresponding to the $\underline{v}^{\mathrm{i}}$ eigenvector, and p_i is the corresponding coordinate of the amino acid composition vector ($\underline{n}$) in the new coordinate system, calculated as a scalar

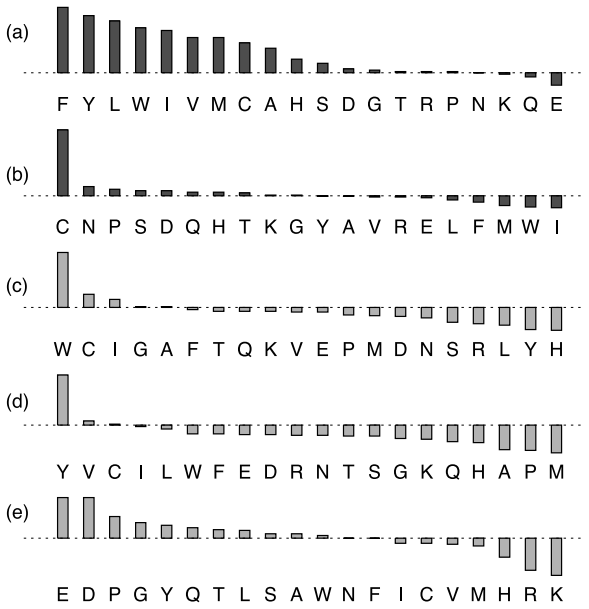

Figure 3. Decomposition of the energy predictor matrix to eigenvalues representing stabilizing and destabilizing interactions. The energy predictor matrix **P**, was decomposed into (a) and (b) negative and (c)–(e) positive eigenvectors. Their corresponding eigenvalues specifying the weights in the energy function are: (a) −52, (b) −40, (c) 24, (d) 13 and (e) 10 (cf. Decomposition). These vectors represent stabilizing and destabilizing contributions to the total pairwise energy content, and can be rationalized in terms of simple physical principles, such as (a) hydrophobicity, (b) cysteine abundance, (d) structure-breaking amino acid residues and (e) net charge of the protein.

product, $p_i = \underline{n}\underline{v}^i$. Since p_i^2 is non-negative, terms corresponding to positive/negative eigenvalues give a positive/negative contribution to the estimated energy. Some of the individual eigenvectors can be directly interpreted in terms of physicochemical factors and linked to stabilization or destabilization depending on its sign. Figure 3 shows the two largest negative (stabilizing) and three largest positive (destabilizing) eigenvectors. We find hydrophobicity (Figure 3(a)) and cysteine content (Figure 3(b)) as dominant factors in stabilization. The vector with the highest eigenvalue is closest to the Sweet–Eisenberg empirical hydrophobicity scale (correlation coefficient: 0.94) among more than 400 different amino acid propensities collected in the AAIndex database.[31] Interestingly, the same scale among hydrophobicities was found to be the best for discriminating structured proteins from IUPs in a systematic search among 265 amino acid properties.[16] Of particular relevance to our assessment of the determinants of protein disorder, the Sweet–Eisenberg scale is based on amino acid replaceability, which correlates with the tendency of side-chains to be buried or exposed in protein crystal structures.[32] As for the destabilizing factors, there is no obvious interpretation of the factor with the largest positive eigenvalue (Figure 3(c)), whereas the next two suggest that the abundance of structure-breaking amino acid residues like Pro, Asn, and Gly (Figure 3(d)) or high net charge (Figure 3(e)) leads to destabilization. It should be made clear, though, that these features are not incorporated in the predictor, but they can be extracted due to their importance in supporting energetically favorable/unfavorable structural states. The intriguing point is that they show good correlation with features used in previous approaches, which are knowledge-based in terms of protein disorder.[14,16,17]

Besides the angles between the vectors, all p_i values depend also on the norm of the amino acid composition vector $\left(\mathrm{NORM} = \sum_{i=1}^{20} n_i^2 = \sum_{i=1}^{20} p_i^2\right)$. This norm takes its minimum value (0.223) for sequences with equal amino acid frequencies, and largest (1.0) when the sequence is composed of a single amino acid. For globular proteins, it varies between 0.23 and 0.39. There are 24 sequences with their norm above 0.35 in the IUP sets, including 16 out of 17 sequences that have at least 50% of residues predicted to have low complexity by the SEG program.[33] Thus, the norm is also a measure of the complexity of the sequence, incorporated into the estimated energy as a scaling factor. According to our model, a low complexity sequence is not necessarily disordered, if its amino acid composition is dominated by stabilizing factors, allowing the formation of favorable contacts. On the other hand, sequences with the least favorable energy are of low complexity as well, as a result of the dominance of one or a few amino acid types that have unfavorable interaction energies for each relation.

The estimated pairwise energy predicts protein disorder

Based on the significant separation between the estimated pairwise energies of globular and experimentally verified intrinsically unstructured proteins, this approach can be turned into a method to predict protein disorder. For this purpose it is more appropriate to consider the local sequential neighborhood only, since many proteins are not fully ordered or disordered. Thus, the original matrix **P**, derived at the level of global sequences, was recalculated by treating each position separately, and taking into account only its predefined neighborhood in sequence. The energy and amino acid composition for each position was calculated only by considering interaction partners 2–100 residues apart. The choice of this range represents a trade-off between the intention of covering most structured domains, but separating distinct domains in multi-domain proteins. This procedure yields an estimated energy at position p of type i:

$$e_i^p = \sum_{j=1}^{20} P_{ij}^p n_j^p$$

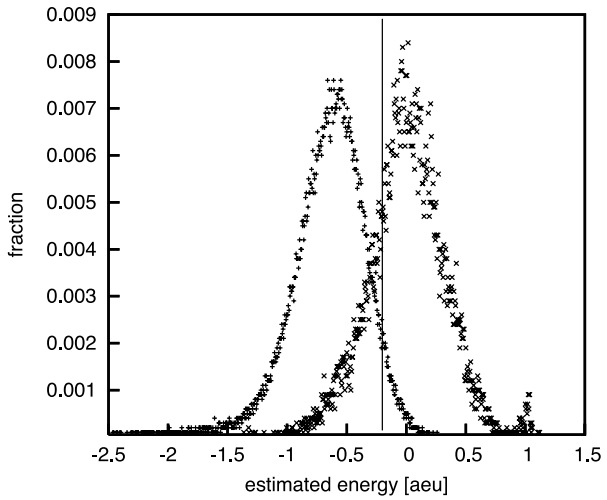

Figure 4. Estimated position-specific pairwise energies of globular proteins and IUPs. The distribution of estimated position-specific pairwise energy scores is shown, calculated by considering the amino acid composition limited to a sequential neighborhood of ± 100 residues and smoothed over 21 residues. The application, termed IUPred, was applied to the Filt_Glob_list (+) and the IUP_list (×), and the frequency of residues was plotted against their local energy content. A threshold of -0.2 [aeu] provides the best separation of individual positions between these two structural classes.

where $\boldsymbol{P}^{p}$ is the position-specific energy predictor matrix. The position-specific estimations of energies were averaged over a window of 21 residues. This method for the prediction of protein disorder is termed IUPred.

By using IUPred, the distribution of scores for globular proteins and IUPs is as shown in Figure 4. The clear separation between the two sets is also apparent at the level of individual positions. From the distribution of globular proteins we determined a threshold where 5% of their positions were predicted as being disordered, similar to the prediction made by Ward *et al.*[9] This value was -0.2 [aeu]: positions with energy content above this cutoff value were predicted to be disordered, whereas positions below were considered as being ordered. Using this limit, 76% of positions of IUPs were predicted to be disordered. These deviations from complete order or disorder are fully acceptable, due to potentially disordered regions in globular proteins in solution observed to be ordered in the solid state, and the existence of significant residual structure in many IUPs.[6,7,34,35]

To see how the choice of the empirical force field affects the predictive power of IUPred, various 20 by 20 **M** scoring matrices were tested. For each **M** matrix the corresponding **P** matrix was derived as desribed in Theory and used for the prediction. We set the threshold to give 5% false positive predictions on the Filt_Glob_list, and calculated the sensitivity of the method as the percentage of predicted disorder on the IUP_list for each interaction matrix (Table 3). The approach of Thomas & Dill[20] yields matrices superior to others, with the much larger dataset bringing about an improvement of almost 3%. The matrix used by Tobi *et al.*[21] performed comparably to the original one of Thomas & Dill in predicting disorder, but the other two showed much less ability to discriminate order from disorder.

Cross-validation of the method

In order to test the ability of our method to generalize on previously unseen data, we carried out a tenfold cross-validation. Glob_list was divided into ten random subsets. One was put aside, and proteins from the remaining nine were used to calculate matrices **M** and **P**, and the cutoff value. This procedure was repeated ten times, and the goodness of fit and the amount of disorder were predicted for the proteins not used in training. It is worth noting that no cross-validation is required for IUPs, as these proteins were not included in any way in the training process.

Over the ten sets, the average of the correlation coefficient was 0.783 ± 0.006 and the $r2$ value was 0.600 ± 0.070, compared with the values obtained for the full set, 0.786 and 0.604, respectively. Both values indicate a similar goodness of fit for globular proteins, independently of whether they were included in the training set. The amount of predicted disorder varied between 3.4% and 6.9%, with the average of $4.96(\pm 0.97)$% for the training sets, compared to 5.0% for the full set.

Table 3. Comparison of different scoring matrices

Interaction matrix	Number of training proteins	Predicted disorder on IUP set (%) (true positives)
Thomas–Dill extended training set	785	75.95
Thomas–Dill[20]	37	73.25
Tobi *et al.*[21]	572	73.09
Mirny–Shakhnovich[22]	104	64.63
Miyazawa–Jernigan[23]	251/1661	63.64

The performance of different interaction matrices in predicting disorder and the number of proteins used to derive them. The Miyazawa–Jernigan matrix was trained on 1661 proteins including homologs, effectively representing 251 families.

Table 4. Performance of disorder prediction methods

Method	True positive rate		False positive rate	
	All positions (%)	Normalized positions (%)	All positions (%)	Normalized positions (%)
IUPred	76.33	67.91	5.33	5.54
PONDR VL3H	66.29	60.74	5.02	7.84
DISOPRED2	63.39	49.08	5.02	6.87
GlobPlot	32.97	30.42	18.07	19.72

Comparison of IUPred, PONDR VL3H, DISOPRED2 and GlobPlot on IUP_list and Filt_Glob_list. The true positive rate was calculated as the percentage of residues predicted as disordered on the IUP_list (sensitivity), while setting the false positive rate (percentage of predicted disordered residues on the Filt_Glob_set), also called specificity, to 5%, or the closest possible value (in the case of GlobPlot). These values are given averaged over all positions, and normalized by the length of the protein. This normalization weights each fragment/protein equally, independently of its length. Predictions by PONDR VL3H were collected from the server at http://www.ist.temple.edu/disprot/predictor.php using the default parameter (window size=1), while DISOPRED2 was downloaded from http://bioinf.cs.ucl.ac.uk/disopred/ and run locally. GlobPlot was also run locally, but with the web server's parameters and taking the CASP-like output (http://GlobPlot.embl.de/).

Comparison of different methods of disorder prediction

Database of disordered proteins

We compared IUPred to three widely used methods for predicting disorder, which differ not just methodologically but also conceptually due to different definitions of disorder. GlobPlot is a simple propensity-based approach evaluating the tendency of residues to be in a regular secondary structure. PONDR VL3H[36] was trained to distinguish experimentally verified disordered proteins from globular proteins by various machine learning approaches. In developing DISOPRED,[9] the definition of disorder was restrained to regions missing from X-ray structures and a support vector machine was trained to specifically recognize these. In contrast, IUPred assigns order/disorder status to residues on the basis of their ability to form favorable pairwise contacts.

To make a realistic comparison of these methods (Table 4), their cutoff values were set so that they yielded the same percentage (5%) of false positive predictions (predicted disordered when in fact ordered) on Filt_Glob_list. The agreement between pairs of predictions were also calculated: two predictions were said to agree when both predicted order or disorder for a given position, and the numbers of agreements were normalized by the total number of positions (Table 5). As GlobPlot was not intended as a per position prediction method, it was included as a simple control to evaluate the performance of a propensity-based approach. Although the performance of the other methods, IUPred, PONDR VL3H and DISOPRED2, is comparable, there are some clear differences among them. IUPred predicted the largest amount of disorder, followed by PONDR VL3H; DISOPRED2 tended to predict the most order at the same level of false prediction rate. These differences are also apparent in the ROC curve, giving the false positive rate against true positive rate for these three methods (Figure 5). Except for very low level of false prediction rate, IUPred achieves the highest true positive rate. Intriguingly, in pairwise comparisons the three methods are very similar, each of them agreeing with the other two on about three-quarters of positions (Table 5).

The goal of this comparison was to assess the performance of IUPred in terms of predicting long disordered regions; however, it cannot be regarded as a complete benchmarking. The test set for disorder was rather small (only 129 proteins), and there could be a significant amount of local order included in this set that the various methods would treat differently. DISOPRED2 was specifically designed to predict short disordered regions in the context of globally ordered proteins, and its performance is expected to be higher on these datasets. Furthermore, some parameters (e.g. window size) could also influence the performance of the methods (VL3H, GlobPlot). Despite these limitations, the results clearly show that IUPred

Table 5. Similarity between disorder prediction methods

Method	Agreement (%)			
	IUPred	PONDR VL3H	DISOPRED2	GlobPlot
IUPred	100	91.61	91.97	80.76
PONDR VL3H	76.60	100	92.26	79.24
DISOPRED2	77.19	77.05	100	79.91
GlobPlot	48.07	47.84	51.31	100

The similarity between pairs of methods was calculated as the number of agreements over all positions in IUP_list (lower triangle) and in Filt_Glob_list (upper triangle). Predictions were collected as given in the legend to Table 4.

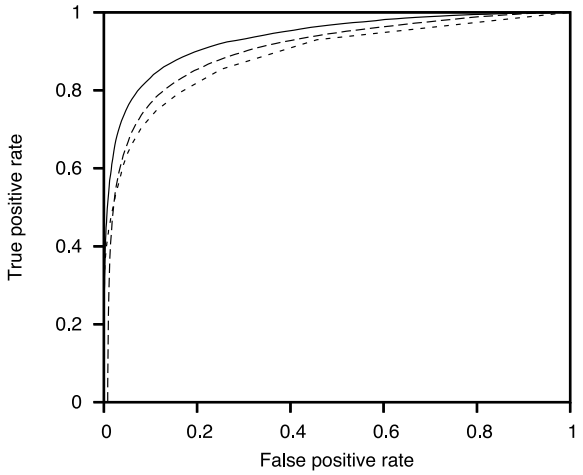

Figure 5. ROC curve for IUPred, PONDR VL3H, and DISOPRED2. Receiver operator characteristic (ROC) curve for IUPred (continuous), PONDR VL3H (broken) and DISOPRED2 (short broken). The true positive rate was calculated as the percentage of residues predicted as disordered on the IUP_list (sensitivity), the false positive rate is the percentage of predicted disordered residues on the Filt_Glob_set, also called specificity.

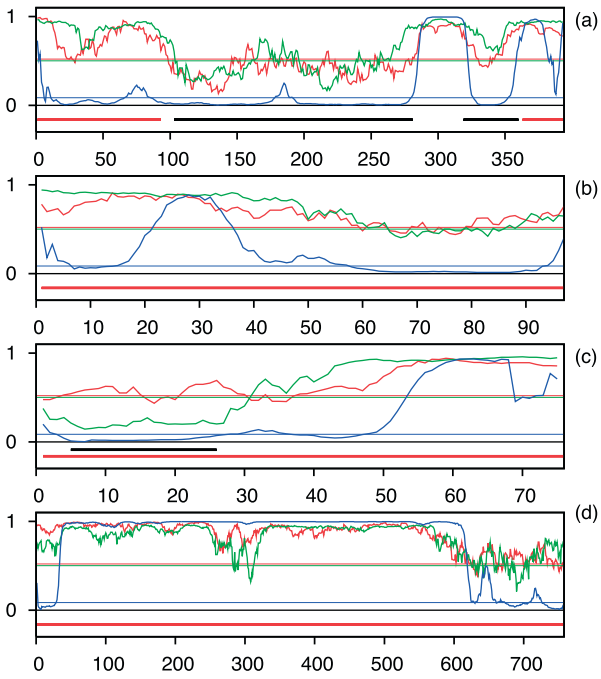

Figure 6. Comparison of IUPred with two other predictors of disorder. IUPred scoring (red) is compared with PONDR VL3H (green) and DISOPRED2 (blue) for (a) p53, (b) FlgM, (c) PKI-alpha and (d) MAP tau. The energy values of IUPred were normalized to fall between [0,1]. Thin horizontal lines of the appropriate color represent threshold values, above which the score is characteristic of disorder (0.5 for IUPred and PONDR and 0.086 for DISOPRED2, the default values in the latter two cases). Below the scores, the region experimentally shown to be disordered (thick red line) or structured in itself or in the presence of a binding partner (thick black line) is indicated.

is a competent predictor of protein disorder. This is achieved by considering only globular proteins during the training, without using any information on intrinsically unstructured proteins.

Examples of individual proteins

As a further means of comparison we present the analysis of four representative proteins, with the prediction output of the position-specific predictors IUPred, PONDR VL3H and DISOPRED2 (Figure 6). The proteins were selected because a good deal of structural information is available on the extent and mode of their disorder. p53 (Figure 6(a)) is a tumor suppressor transcription factor, the structural disorder of which has been convincingly demonstrated for the N-terminal (1–93) and C-terminal (363–393) domains.[2,37] FlgM, or anti-sigma-28 factor (Figure 6(b)), is one of the first proteins to be identified as intrinsically unstructured along its entire length.[38] Its C-terminal half, with residues 60–73 and 83–90 in particular, show some α-helical preference in solution,[39] possibly relevant to the physiological function of this protein.[40] PKI-alpha (Figure 6(c)), a heat-resistant inhibitor of cAMP-dependent protein kinase, is also disordered along its entire length[41] with its inhibitory segment (1–13) and nuclear export signal (35–47) tending to adopt an α-helical structure in the unbound state.[42] Microtubule-associated protein tau (Figure 6(d)) belongs to a family of heat-stable MAPs (also including MAP2 and MAP4), which are disordered along their entire length and bind microtubules *via* a C-terminal microtubule-binding domain.[43]

The plots by the three methods agree reasonably well, with some differences. DISOPRED2 tends to predict more order than either IUPred or PONDR VL3H, even at places where experimental evidence is for a largely disordered state, such as the N-terminal domain of p53 and PKI-alpha or the C-terminal region of FlgM. Interestingly, these regions show some tendency to be transiently ordered, as stated above. This local preference for order is probably captured by IUPred and PONDR, as witnessed by the value of their disorder score approaching or even crossing the threshold. The noted tendency of the C-terminal region of tau to be ordered is also worthy of note in light of its interaction with microtubules; this might occur *via* preformed structural elements, as demonstrated to be a general feature of IUPs.[35] These examples further illustrate the similarities and differences among the three prediction methods, with IUPred predicting the most disorder for fully or largely disordered proteins, and DISOPRED2 predicting the least. In addition, by looking at these examples, it is advisable to treat regions of disagreement among the predictors with caution and consider them as potential recognition sites.

Discussion

The growing number of examples of IUPs has encouraged us to revisit the issue of the foldability of polypeptide chains. In order to understand the differences between IUPs and folded proteins better, we estimated the pairwise energy content of proteins in their native structural state by a quadratic form involving the amino acid composition vector and the energy predictor matrix. The parameters of the matrix were derived by least-squares fitting using globular proteins of known structure, which also allowed the goodness of the estimation to be tested. The robustness of this approach is quite surprising, considering that every protein structure is an intricate architecture of a multitude of interresidue contacts. We did not attempt to predict the exact pattern of these interactions, i.e. the structure, in detail, rather the compatibility of a given polypeptide with the formation of sufficient favorable interactions, as observed in globular proteins.

The success of our approach underlines some common, fundamental properties of sequences with stable folded structures. The native structure of folded proteins corresponds to a pronounced energy minimum, with no other conformations having comparable energy.[44,45] Ensuring this energetic separation demands the native structure to efficiently use the interactions compatible with the given sequence. As the maximum capacity for each amino acid to participate in these interactions is limited by its chemical nature, the amino acid composition can be related to the total interaction energy pertaining to the most favorable interaction pattern among all residues present in a protein. The effectiveness of our model in estimating the pairwise energy content can be attributed to folded structures being close to the optimal energy level allowed by the amino acid residues in the sequence.

The energy per residue of stably folded proteins falls into a quite narrow range, dominated by favorable interactions; the total pairwise energy estimated by our approach is consistent with this energy range. In contrast, the predicted energy of IUPs is higher. An important conceptual point is that a polypeptide with an amino acid composition compatible with a folded structure does not necessarily have a unique structure. This can be easily demonstrated by considering the random permutation of sequences of folded proteins. Although we predict the same energy for the myriad of sequences compatible with a particular amino acid composition, for most of them we expect no corresponding unique structure. Similarly, we predict globular-like energy for truncated domains or proteins, although these sequences are not likely to fold on their own. Nonetheless, these polypeptides are not IUPs either, since they exhibit some tendency to form contacts. The way around this dilemma in the present approach is that it predicts the optimum of energy, which is not generally achievable by a random sequence. Folded structures, however, are realized by highly evolved sequences, compatible with these energies. IUP sequences are also special, selected by evolution to avoid the formation of favorable contacts in any conformation. The finding that the estimated total interaction energy reproduces the basic difference between structured and disordered proteins basically underlines the concept of protein disorder, i.e. that the lack of a well-defined 3D structure is an intrinsic property of certain evolved proteins.

Our ability to reproduce these special features depends on using the right potentials for approaching the actual interaction energies. The goodness of such extracted potentials is usually tested by their ability to identify the native structure as the lowest energy state among all the proteins in a dataset. The particular approach, proposed by Thomas & Dill,[20] relies on the Boltzmann relation to extract energy-like quantities from amino acid pairing frequencies, but relative to a reference state obtained through an iterative protocol to reflect the predicted ensemble of interactions. This approach aims not only at discriminating the native structure from decoys but also at giving the ratios of the interaction energies correctly. Thus, these potentials are the closest to reproducing the true energies that drive amino acid residues to form, or avoid, contacts. This could explain why the Thomas–Dill matrix outperformed other matrices in estimating the pairwise energy content of disordered proteins.

In the light of this special property of the underlying interaction matrix, we can also interpret the unexpected finding that the average energy level of IUPs is very close to zero, i.e. stabilizing and destabilizing interactions cancel. Although the absolute energy values were arbitrary, this finding is invariant to scaling, thus this energetic neutrality is a genuine property of disordered proteins. At the level of individual proteins, this neutrality may result from an overall lack of long-range interactions, but also from the balance of local organization and long-distance repulsion. For some IUPs, however, the balance appears to be set off towards net stabilization. Indeed, the predicted pairwise energy content of proteins with a molten-globule type of disorder[30] on average is more favorable compared to coil-like disordered proteins. It will be interesting to see how such individual structural features correlate with function.

As seen, our approach provides a realistic approximation of structural interaction energy of proteins, enabling the prediction of intrinsic structural disorder. This idea of the importance of interaction capacity has also been raised recently by work in which the average number of contacts per residue was used as an indicator of disorder.[46] Our quadratic formula combined with the energy predictor matrix captures the energetic aspect of this observation. By limiting the calculation to a predefined sequential neighborhood, it yields a position-specific score characteristic of the tendency of a given amino acid to fall into a structurally ordered or disordered region. This application we

term IUPred and intend to make it publicly available *via* the Internet. The logic of IUPred differs from previous prediction algorithms, which were trained on disordered proteins/segments. As already alluded to, these approaches mostly suffer from inconsistencies in the underlying databases, i.e. the inclusion of sequences of intrinsic order classified as disordered and sequences of intrinsic disorder classified as ordered. As IUPred was not trained on such data, its unbiased assessment of the structural status of an unknown sequence/segment is of confirmatory value. We have tested this conclusion by comparing IUPred with generally accepted predictors, PONDR VL3H, DISOPRED2 and GlobPlot, on disordered databases and by examining predictors on individual proteins. Although predictions were similar with the IUP dataset, IUPred predicted the most disorder and agreed best with experimental data (Table 4).

Decomposition of the energy matrix connects our model to previous attempts to predict IUPs by using simple physico-chemical properties of proteins. Some of the eigenvectors with the largest eigenvalues showed strong correlations with physico-chemical parameters, such as hydrophobicity, cysteine content, structure-breaking properties and net charge (Figure 3). Two of these, hydrophobicity and net charge have been used in the Uversky plot to separate globular proteins and IUPs,[17] and the importance of structure-breaking amino acid residues has also been noted.[5,6] We found that the eigenvector with the highest eigenvalue matches the Sweet–Eisenberg hydrophobicity scale[32] the best, in accordance with a previous analysis of amino acid factors discriminating structured proteins and IUPs.[16] This concurrence vindicates our approach, as it does not rely on prior experimental data on IUPs, still it automatically finds and combines the properties that are important for this task. Our predictor combines these factors in a quadratic function, which distinguishes it from previous, propensity-based linear predictors.[13–15,46] As a result of the higher-order statistics, the contribution of a given amino acid to disorder/order discrimination is context-dependent, i.e. it depends on the amino acid type as well on the amino acid composition of the sequential neighborhood of the given residue. For example, the contribution of Lys would be different if it is surrounded by other positive charges, implying an increase in the probability of unfavourable interactions, than if it is surrounded by negatively charged residues. This is in accordance with high net charge, and not simply the total charge, being one of the key determinants of disorder.[17] This interdependence of residues is manifest in the appearance of flavors of disorder.[13]

In summary, our model estimates the pairwise energy of proteins from their amino acid compositions. This allows us to test sequences for foldability, even in the absence of a structural model. By sequentially limiting the calculation, it serves as a predictor of protein disorder. By applying this scheme for IUPs, we showed that these proteins have a special amino acid composition, which, independently of the actual sequence, does not allow the formation of sufficient favorable contacts expected for folded proteins. Given the heterogeneity and ambiguity of experimental techniques used to demonstrate the lack of structure so far, a key inference from our studies is that IUPs share a common property that distinguishes them from the class of folded proteins.

Materials and Methods

Databases

For the purpose of parameter fitting, the September 2001 release of the PDB-select database[47] with $<25\%$ sequence identity cutoff was used. Entries with resolution worse than 2.5 Å, with chain breaks or with C^{α} atoms only, were omitted; the resulting dataset contained 953 protein chains. During the force field optimization, we considered the native structure for non-transmembrane sequences with length between 40 and 350, reducing the number of proteins to 785 (Glob_list), but all structures were used as a skeleton to generate decoys.

In principle, this list could also contain IUPs, e.g. as part of multichain complexes. For the purpose of testing we created a filtered list of globular proteins with the aim to eliminate the potentially dubious cases. A newer release of PDB-select (April 2002) was used, and all entries involving multiple chains, transmembrane segments, or the binding of nucleic acid residues, heme, or metal ions were omitted, resulting in 559 proteins (Filt_Glob_list). The two lists (Glob_list and Filt_Glob_list) are given in the Supplementary Data (Tables S1 and S2).

The IUP dataset (IUP_list) contained 129 proteins and protein segments with experimentally verified disordered status. The complete list is given in the Supplementary Data (Table S3). The total number of residues in this set is 26,794.

Force field optimization

A coarse-grained approach was used to describe the interactions between residues. Amino acid residues were treated as single interaction centers located at their C^{β} atom (virtual C^{β} in the case of Gly). The low-resolution energy of contacts between different amino acid residues, expressed in the form of a 20 by 20 matrix, was calculated from the observed frequencies of amino acid pairs. The interaction matrix was calculated by the iterative algorithm proposed by Thomas & Dill,[20] but on 785 proteins (Glob_list) instead of the original 37. The resulting matrix **M** is given in Table 1.

Acknowledgements

This work was supported by grants T34131 and F043609 from OTKA, Bolyai János fellowships for Zs.D. and P.T., and the International Senior Research Fellowship GR067595 from the Wellcome Trust for P.T. The fruitful discussions with Nicholas

E. Dixon and Tamas Hauer are gratefully acknowledged.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.01.071

References

1. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I. & Wright, P. E. (1996). Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl Acad. Sci. USA*, **93**, 11504–11509.
2. Dawson, R., Muller, L., Dehner, A., Klein, C., Kessler, H. & Buchner, J. (2003). The N-terminal domain of p53 is natively unfolded. *J. Mol. Biol.* **332**, 1131–1141.
3. Radhakrishnan, I., Perez-Alvarado, G. C., Dyson, H. J. & Wright, P. E. (1998). Conformational preferences in the Ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB. *FEBS Letters*, **430**, 317–322.
4. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293**, 321–331.
5. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
6. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
7. Tompa, P. (2003). The functional benefits of protein disorder. *J. Mol. Struct. (Theochem)*, **666-667**, 361–371.
8. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171.
9. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645.
10. Iakoucheva, L., Brown, C., Lawson, J., Obradovic, Z. & Dunker, A. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584.
11. Liu, J., Tan, H. & Rost, B. (2002). Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322**, 53–64.
12. Weathers, E. A., Paulaitis, M. E., Woolf, T. B. & Hoh, J. H. (2004). Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Letters*, **576**, 348–352.
13. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins: Struct. Funct. Genet.* **52**, 573–584.
14. Xie, Q., Arnold, G. E., Romero, P., Obradovic, Z., Garner, E. & Dunker, A. K. (1998). The sequence attribute method for determining relationships between sequence and protein disorder. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 193–200.
15. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucl. Acids Res.* **31**, 3701–3708.
16. Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J. *et al.* (2001). The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.* **6**, 89–100.
17. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
18. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
19. Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.
20. Thomas, P. D. & Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
21. Tobi, D., Shafran, G., Linial, N. & Elber, R. (2000). On the design and analysis of protein folding potentials. *Proteins: Struct. Funct. Genet.* **40**, 71–85.
22. Mirny, L. A. & Shakhnovich, E. I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **264**, 1164–1179.
23. Miyazawa, S. & Jernigan, R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623–644.
24. Melo, F., Sanchez, R. & Sali, A. (2002). Statistical potentials for fold assessment. *Protein Sci.* **11**, 430–448.
25. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–89.
26. Torda, A. E. (1997). Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7**, 200–205.
27. Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* **295**, 337–356.
28. Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* **18**, 338–352.
29. Khatun, J., Khare, S. D. & Dokholyan, N. V. (2004). Can contact potentials reliably predict stability of proteins? *J. Mol. Biol.* **336**, 1223–1238.
30. Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
31. Kawashima, S., Ogata, H. & Kanehisa, M. (1999). AAindex: amino acid index database. *Nucl. Acids Res.* **27**, 368–369.
32. Sweet, R. M. & Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* **171**, 479–488.
33. Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.
34. Uversky, V. N. (2002). What does it mean to be natively unfolded? *Eur. J. Biochem.* **269**, 2–12.
35. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. (2004). Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026.
36. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J. & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53**, 566–572.

37. Bell, S., Klein, C., Muller, L., Hansen, S. & Buchner, J. (2002). p53 contains large unstructured regions in its native state. *J. Mol. Biol.* **322**, 917.
38. Daughdrill, G. W., Chadsey, M. S., Karlinsey, J. E., Hughes, K. T. & Dahlquist, F. W. (1997). The C-terminal half of the anti-sigma factor. FlgM, becomes structured when bound to its target, sigma 28. *Nature Struct. Biol.* **4**, 285–291.
39. Daughdrill, G. W., Hanely, L. J. & Dahlquist, F. W. (1998). The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations. *Biochemistry*, **37**, 1076–1082.
40. Dedmon, M. M., Patel, C. N., Young, G. B. & Pielak, G. J. (2002). FlgM gains structure in living cells. *Proc. Natl Acad. Sci. USA*, **99**, 12681–12684.
41. Hauer, J. A., Taylor, S. S. & Johnson, D. A. (1999). Binding-dependent disorder–order transition in PKI alpha: a fluorescence anisotropy study. *Biochemistry*, **38**, 6774–6780.
42. Hauer, J. A., Barthe, P., Taylor, S. S., Parello, J. & Padilla, A. (1999). Two well-defined motifs in the cAMP-dependent protein kinase inhibitor (PKIalpha) correlate with inhibitory and nuclear export function. *Protein Sci.* **8**, 545–553.
43. Schweers, O., Schonbrunn-Hanebeck, E., Marx, A. & Mandelkow, E. (1994). Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.* **269**, 24290–24297.
44. Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
45. Onuchic, J. N. & Wolynes, P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75.
46. Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. (2004). To be folded or to be unfolded? *Protein Sci.* **13**, 2871–2877.
47. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.

Edited by J. Thornton

(Received 27 October 2004; received in revised form 26 January 2005; accepted 28 January 2005)