
Optimization of linear disorder predictors yields tight association between crystallographic disorder and hydrophobicity

NATHAN B. HOLLADAY,^{1,2} LISA N. KINCH,³ AND NICK V. GRISHIN^{1,3}

¹Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA

²Medical Scientist Training Program, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA

³Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA

(RECEIVED May 4, 2007; FINAL REVISION July 2, 2007; ACCEPTED July 3, 2007)

Abstract

X-ray crystallographic protein structures often contain disordered regions that are observed as missing electron density. Diffraction data may give little or no direct evidence as to the specific nature of disordered regions. We have developed a weighted window-based disorder predictor optimized using crystallographic data. Performance of a predictor is strongly influenced by chain termini. Optimized score adjustment values for amino- and carboxy-terminal positions demonstrate a simple, monotonic relationship between disorder and residue distance from termini. This optimized disorder predictor performs similarly to DISOPRED2 on crystallographically disordered regions. Data-optimized residue disorder propensities show strong linear correlation with experimentally determined amino acid transfer energies between water and hydrogen-bonding organic solvents, which primarily reflect residue hydrophobicity (exemplified by the Nozaki-Tanford hydrophobicity scale). Disorder propensities do not correlate as well with transfer energies between water and apolar solvents, which primarily reflect a different hydrophobic property: residue hydrophilicity (also reflected by the Kyte-Doolittle hydrophobicity scale). Our results suggest that while hydrophobic side-chain interactions are primarily involved in determining stability of the folded conformation, hydrogen bonding, and similar polar interactions are primarily involved in conformational and interaction specificity.

Keywords: X-ray crystallography; protein disorder; protein structure; hydrophobicity; hydrophilicity; simulated annealing; predictor optimization

Reprint requests to: Nick V. Grishin, Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA; e-mail: grishin@chop.swmed.edu; fax: (214) 648-9099.

Abbreviations: LOR, log odds ratio; ROC, receiver operating characteristic; SCOP, Structural Classification of Proteins (<http://scop.berkeley.edu>); SVM, support vector machine; opt seq, optimized amino acid residue disorder scale for simple sequence predictors; opt prof, optimized amino acid residue disorder scale for profile predictors; LOR seq, log odds ratios for simple sequences; LOR prof, log odds ratios for profiles.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.072980107>.

For over fifty years, scientists have discussed how fundamental physical properties of amino acids might relate to protein structure (Waugh 1954; Kauzmann 1959). Yet significant gaps remain in our understanding of how the simple physical properties of amino acids dictate the complex structural characteristics of proteins (Chandler 2005). Protein disorder, or the lack of consistent folded structure, represents one such characteristic.

To help evaluate disorder in proteins, a number of predictors have been developed (Galzitskaya et al. 2006a,b; Han et al. 2006; Vullo et al. 2006; Wang and Donald 2006;

Dosztanyi et al. 2007; Hirose et al. 2007; Shimizu et al. 2007; Sickmeier et al. 2007; Sugase et al. 2007). The DisProt Website (Sickmeier et al. 2007), at <http://www.disprot.org>, provides links to several predictors. Some predictors utilize various predetermined residue-type characteristics to make predictions (Linding et al. 2003b; Coeytaux and Poupon 2005; Dosztanyi et al. 2005a,b; Prilusky et al. 2005; Peng et al. 2006), while other predictors use “machine learning” of disorder data to develop complex networks of parameters (Romero et al. 1997; Linding et al. 2003a; Vucetic et al. 2003; Ward et al. 2004; Cheng et al. 2005; Yang et al. 2005) that are not easily related to physical terms (Lise and Jones 2005). Ferron et al. (2006) provide an informative review of disorder predictors.

We have developed data-optimized disorder predictors that use a weighted sliding-window algorithm to calculate position-dependent disorder scores. Using such an approach allows position-dependent predictions that can account for local influence of adjacent residues within the span of the window. Some predictors specifically address disorder-prone chain-terminal regions (Li et al. 1999; Ward et al. 2004). We use a simple method to address predictions at termini, which reveals similar behavior of disorder at each terminus. As opposed to predictors that utilize complex networks of parameters, the optimized parameters from our window-based predictors provide insight into disorder. Thus, optimized disorder parameters are compared with amino acid properties to provide insight into the physical underpinnings of disorder.

Several numerical indices represent physicochemical properties of amino acids measured either experimentally using individual amino acids or statistically using information available in protein structures. Many of these indices have been assembled into an AAindex database (Kawashima and Kanehisa 2000) and have been classified into hierarchical clusters of similar properties (Tomii and Kanehisa 1996). Numerical indices reflecting various properties have been used in predicting disorder, such as coil propensity (Linding et al. 2003b), hydrophathy (Prilusky et al. 2005), and interaction energy (Dosztanyi et al. 2005b). Property scales may also be derived by training predictors on disorder data (Weathers et al. 2004).

Among properties associated with disorder, “hydrophobicity” is commonly mentioned (Uversky et al. 2000; Williams et al. 2001; Linding et al. 2003a; Dosztanyi et al. 2005b; Peng et al. 2006). However, the term *hydrophobicity* has been loosely applied to widely varying scales (Creighton 1993, 2002), such as those of Kyte and Doolittle (1982) and Nozaki and Tanford (1971). These scales display only marginal correlation. Thus, the implications of associations between disorder and hydrophobicity are unclear. Our amino acid disorder propensities demonstrate a strong linear correlation with experimentally derived transfer energies between polar

organic and aqueous phases, while showing a weaker correlation with transfer energies between apolar organic solvent and water (Radzicka and Wolfenden 1988). We therefore discriminate between two distinct amino acid properties that describe side-chain affinities to organic solvents.

The first property is characterized by the transfer energies between water and polar organics, for example, alcohols. In this case, both solvents offer hydrogen-bonding interactions, but only the organic phase offers hydrophobic protection. Assuming that hydrogen-bonding capacity is fully realized in both phases, partitioning is driven by the hydrophobic effect, and we term this property “hydrophobicity.” The second property refers to transfer energies between water and nonpolar organics, for example, cyclohexane. In this case, hydrogen bonds can be formed in water but not in cyclohexane. Thus, the dominant contribution to partitioning is hydrogen-bond formation, and we term this property “hydrophilicity.” In our terminology, hydrophobicity and hydrophilicity are not two polar opposites but represent two different, although not fully uncorrelated, amino acid properties. The two properties are distinguishable by the type of environment that surrounds amino acids in partitioning experiments, with the former describing partitioning between polar solvent and water, and the latter playing a dominant role in partitioning between nonpolar solvent and water. We discuss how the particular pattern in our data-optimized parameters helps clarify the relationship between this more precisely defined property of hydrophobicity and disorder.

Results and Discussion

Predictor performance

Figure 1 compares performance of our predictors (including simple sequence and profile-based predictors, with and without tail adjustments) with DISOPRED2 (Ward et al. 2004), a support vector machine (SVM)/neural network-based predictor of disorder also developed using crystallographic data that uses PSI-BLAST-generated sequence alignment profiles (Altschul et al. 1997) in its prediction. Some recent papers have compared disorder predictor performance with probability excess measurements (Yang et al. 2005; Esnouf et al. 2006; Su et al. 2006). Probability excess values typically depend on the cutoff score used in making a binary decision, and the probability excess reflects a single point on the receiver operating characteristic (ROC) curve. We use traditional full ROC curves for performance comparison.

As with DISOPRED (Jones and Ward 2003; Ward et al. 2004), use of profiles enhances overall performance compared with using simple sequence (Fig. 1A). However, the

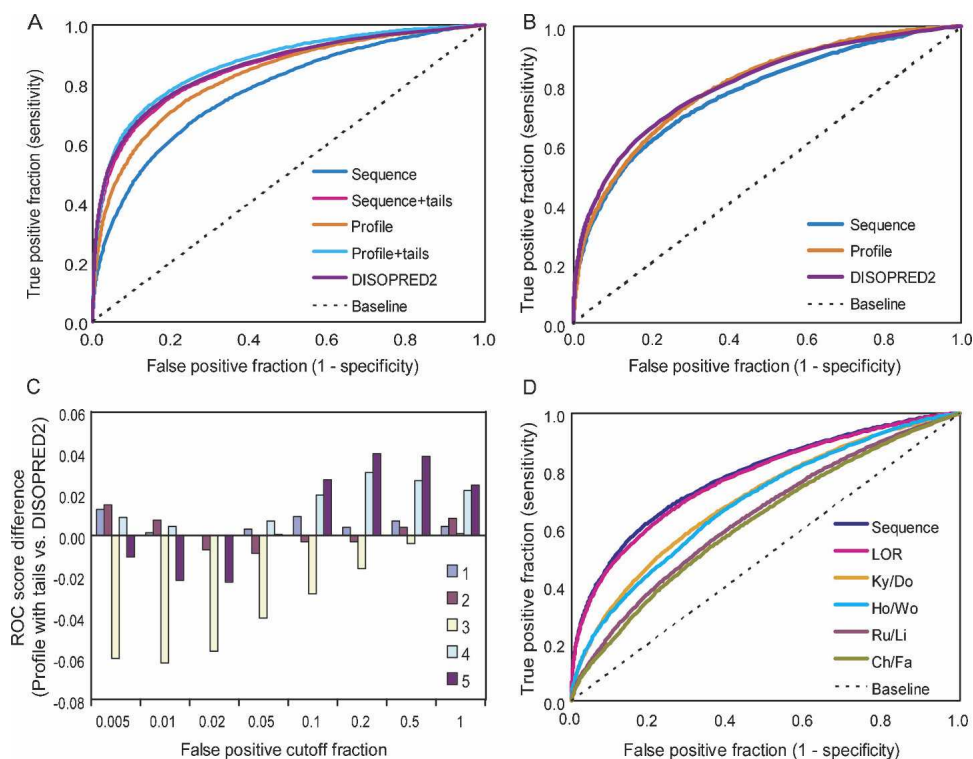


Figure 1. Performance comparison. DISOPRED2 serves as a reference. (A) Average ROC curves, including terminal residues in performance analysis. (B) “Average” ROC curves, excluding 30 terminal residues in performance analysis. (C) Differences between performance of profile with tail adjustments predictor and DISOPRED2 on individual testing data subsets. ROC scores at different cutoffs obtained by subtracting DISOPRED2’s ROC score from the profile with tail adjustments predictor’s ROC score (see Methods for explanation of ROC scores). (D) Average ROC curve for optimized simple sequence predictor (Sequence) compared with performance substituting various pre-existing scales for disorder propensities (see abbreviation footnote and body of text for further details). Areas under ROC curves: Sequence, 0.779; LOR, 0.771; Ky/Do, 0.697; Ho/Wo, 0.689; Ru/Li, 0.637; Ch/Fa, 0.621. (Area under curve for random predictions = 0.5.)

improvement primarily occurs in the prediction of disorder in chain terminal regions, which make up more than half of the disorder in data sets (see Methods). When performance on these terminal residues is excluded from analysis, the simple sequence-based predictor performs comparably with the profile predictor and DISOPRED2 (Fig. 1B). Furthermore, when terminal residues *are* included in performance analysis, our treatment of sequence ends substantially narrows the performance gaps between the simple predictor, the profile predictor, and DISOPRED2 (Fig. 1A).

Although the predictors with tail adjustments are similar in overall performance to DISOPRED2, variations in performance on cross-validation data subsets reveal that predictors behave differently (Fig. 1C). Data subset 3 appears to be substantially affected by the Structural Classification of Proteins (SCOP) family “RNA-polymerase beta-prime” (ID = 64490), which contains sizeable disordered regions. Two of the chains contain long, imperfect repeats of the sequence, “PSTPSYS,” a pattern that may be detected somehow by DISOPRED2.

Optimized parameters

Optimized predictor parameters (between 55 and 119, depending on the predictor) include sliding window weights, amino acid disorder propensities, and tail adjustment values (Fig. 2). Parameters optimized on different cross-validation data sets showed high consistency (data not shown). Window weight parameters form a bell-shaped curve, showing the relative degree to which nearer and farther positions are appropriately taken into account in predicting crystallographic disorder. Positions on the C-terminal side weighted slightly more than their N-terminal counterparts for both profile and simple sequence predictors (Fig. 2A). The reason for this mild asymmetry is unclear.

Disorder values for the sequence- and profile-based predictors follow different patterns (Fig. 2B), but for both predictors, tryptophan is the most order-associated residue type, and serine is the most disorder-associated standard residue type. In the simple sequence predictors, threonine and alanine, with disorder values close to 0, have approximately average ordering propensity; W, C, F,

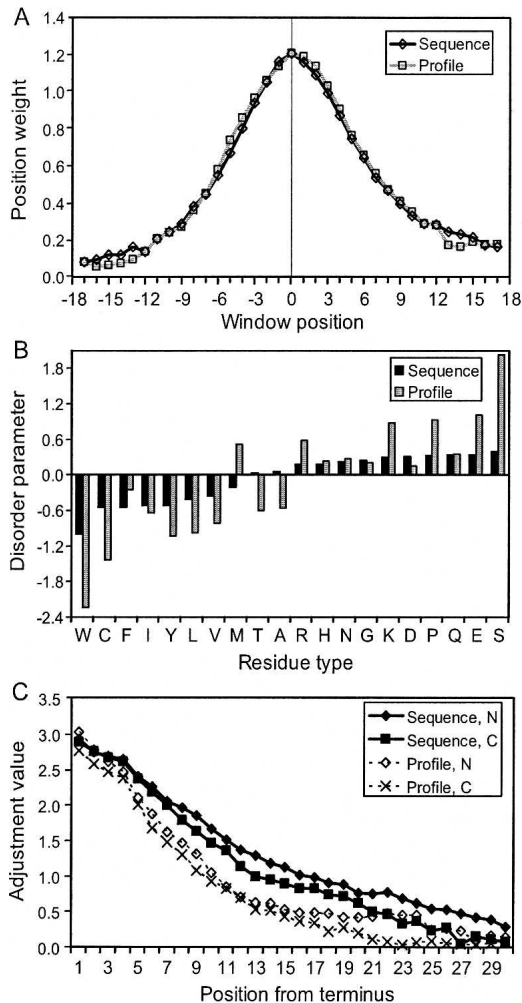


Figure 2. Predictor parameters. (A) Optimized weights are depicted for each position in the sliding window as open diamonds. The calculated disorder score is assigned to the residue position (0) indicated by a vertical line. (B) Simple sequence predictor disorder values are in black bars, and profile predictor values are in gray bars. Negative values indicate more order than average, and positive values indicate more disorder than average. LORs are based on residue frequency in disorder data sets. (C) Tail adjustment values, which are added to disorder scores at positions near amino and carboxy termini when the tail adjustment option is used.

I, Y, L, V, and M display less than average disorder propensity; and R, H, N, G, K, D, P, Q, E, and S display higher than average disorder propensity.

Simple sequence predictor disorder propensities (Table 1) correlate strongly with statistical residue disorder propensities (log odds ratios; $R^2 = 0.946$; Table 2), but a similar relationship is not seen for profiles (Table 2), indicating that statistical derivation of parameters does not always produce the same results as optimization. Of note, the calculation of log odds ratios (LORs) closely resembles that of free energy ($\Delta G = \ln K$), and LORs are additive (behave linearly).

Tail adjustments are optimized in conjunction with composition-based disorder scoring and with the exclusion of polyhistidine tag influence (see Methods), thus removing confounding effects that might result from sequence bias at termini. The tail adjustments thus produced demonstrate that disorder approximately follows a simple, monotonic function of distance from the terminus that is equal for N- and C-termini (Fig. 2C), which may be the result of decreased constraint at chain termini. Tail adjustment values for the first three positions are negatively displaced because of the exclusion of disordered stretches shorter than four residues from optimization performance measures (see Methods).

Disorder propensity and hydrophobicity

Table 1 lists disorder propensities for the simple sequence predictor. Other groups have found similar scales in association with disorder (Williams et al. 2001; Linding et al. 2003a; Weathers et al. 2004). Substituting optimized parameters with rationally selected parameter scales in the simple disorder predictor allows comparison of different indices in predicting crystallographically disordered regions (Fig. 1D), including some indices utilized by available disorder predictors (Linding et al. 2003b; Prilusky et al. 2005; Peng et al. 2006). When substituted

Table 1. Simple sequence predictor disorder values for common residue types, selenomethionine (sM), and N-terminal methionine (nM)

Residue type	Optimized disorder value
W	-1.00739
C	-0.540732
F	-0.540414
I	-0.514274
Y	-0.513589
L	-0.418184
sM	-0.373603
V	-0.358167
M	-0.216377
T	0.033232
A	0.0642762
R	0.176914
H	0.18568
N	0.221683
G	0.241088
K	0.300523
D	0.313504
P	0.32731
Q	0.336406
E	0.33729
S	0.400289
nM	0/0.502239 ^a

^a0 for predictor without tail adjustments, 0.502239 for predictor with tail adjustments. All other types of residues receive a disorder value of 0.

Table 2. Correlations (R^2) for selected scales

	Opt seq	Opt prof	LOR seq	LOR prof	DisProt	No/Ta	Ho/Wo	oct/wat	chx/wat	chx/oct	Ky/Do	95% bur	14 Å cont	Z bur	MPS	P sheet	P helix	Ch/Fa coil	Ru/Li coil
Opt seq ^a	1	0.725	0.946	0.924	0.747	0.977	0.649	0.788	0.496	0.359	0.424	0.304	0.840	0.924	0.543	0.639	0.001	0.208	0.326
Opt prof ^b	0.646	1	0.739	0.692	0.531	0.571	0.450	0.504	0.327	0.267	0.241	0.206	0.528	0.620	0.266	0.529	0.002	0.152	0.194
LOR seq ^c	0.969	0.643	1	0.955	0.868	0.952	0.554	0.770	0.407	0.294	0.397	0.271	0.850	0.897	0.458	0.599	0.000	0.186	0.289
LOR prof ^d	0.975	0.570	0.955	1	0.874	0.957	0.963	0.960	0.593	0.410	0.314	0.292	0.909	0.940	0.611	0.547	0.093	0.295	0.487
DisProt ^e	0.793	0.348	0.878	0.809	1	0.825	0.491	0.690	0.324	0.224	0.349	0.260	0.785	0.718	0.410	0.526	0.001	0.135	0.225
No/Ta ^f	0.982	0.559	0.953	0.957	0.850	1	1	0.939	0.326	0.123	0.060	0.032	0.761	0.918	0.416	0.481	0.056	0.179	0.442
Ho/Wo ^g	0.977	0.574	0.925	0.963	0.754	1	1	0.834	0.694	0.562	0.485	0.482	0.675	0.688	0.737	0.544	0.015	0.128	0.110
oct/wat ^h	0.941	0.458	0.927	0.960	0.837	0.962	0.945	1	0.702	0.577	0.472	0.472	0.772	0.835	0.725	0.594	0.000	0.248	0.215
chx/wat ⁱ	0.491	0.262	0.477	0.593	0.357	0.312	0.477	0.574	1	0.983	0.803	0.790	0.588	0.565	0.823	0.503	0.000	0.265	0.162
chx/oct ^j	0.314	0.175	0.303	0.410	0.205	0.08	0.300	0.384	0.964	1	0.807	0.853	0.450	0.446	0.758	0.414	0.002	0.258	0.094
Ky/Do ^k	0.225	0.066	0.247	0.314	0.207	0.024	0.206	0.331	0.818	0.846	1	0.759	0.633	0.562	0.760	0.495	0.006	0.277	0.201
95% bur ^l	0.219	0.061	0.225	0.292	0.193	0.009	0.196	0.301	0.809	0.849	0.903	1	0.439	0.416	0.708	0.317	0.001	0.192	0.045
14 Å cont ^m	0.830	0.400	0.856	0.909	0.774	0.775	0.807	0.905	0.764	0.599	0.580	0.521	1	0.895	0.729	0.654	0.014	0.320	0.375
Z bur ⁿ	0.914	0.492	0.895	0.940	0.728	0.914	0.932	0.952	0.590	0.409	0.382	0.315	0.907	1	0.654	0.626	0.003	0.263	0.324
MPS ^o	0.510	0.159	0.507	0.611	0.476	0.390	0.537	0.655	0.819	0.726	0.746	0.741	0.774	0.669	1	0.459	0.026	0.310	0.231
P sheet ^p	0.562	0.435	0.583	0.547	0.421	0.455	0.515	0.510	0.387	0.292	0.353	0.253	0.602	0.606	0.318	1	0	0.234	0.393
P helix ^q	0.042	0.015	0.054	0.093	0.046	0.066	0.046	0.087	0.133	0.124	0.171	0.071	0.153	0.115	0.173	0.006	1	0.679	0.468
Ch/Fa coil ^r	0.211	0.148	0.232	0.295	0.152	0.187	0.201	0.270	0.355	0.319	0.395	0.223	0.407	0.348	0.339	0.189	0.844	1	0.712
Ru/Li coil ^s	0.427	0.276	0.445	0.487	0.320	0.427	0.412	0.494	0.378	0.286	0.387	0.180	0.565	0.591	0.372	0.498	0.475	0.771	1

Above the diagonal, calculations include all amino acid types. Below the diagonal, calculations exclude ionic residues (D, E, H, K, R), cyteine, and proline.

R^2 values above 0.7 are highlighted in gray.

^aopt seq: optimized amino acid residue disorder scale for simple sequence predictors.

^bopt prof: optimized amino residue disorder scale for profile predictors.

^cLOR seq: log odds ratio for simple sequences (see Methods).

^dLOR prof: log odds ratio for profiles (see Methods).

^eDisProt: the database of disordered proteins (Sickmeier et al. 2007).

^fNo/Ta: Nozaki-Tanford hydrophobicity scale (Nozaki and Tanford 1971).

^gHo/Wo: Hopp-Woods scale (Levitt 1976; Hopp and Woods 1981).

^hoct/wat: "octanol" (polar organic solvent) to water transfer energies (Guy 1985; Radzicka and Wolfenden 1988).

ⁱchx/wat: cyclohexane to water transfer energies (Radzicka and Wolfenden 1988).

^jchx/oct: cyclohexane to "octanol" transfer energies (Radzicka and Wolfenden 1988) (calculated as difference between chx/wat and oct/wat scales).

^kKy/Do: Kyte-Doolittle scale (Kyte and Doolittle 1982) derived from the Levitt scale (Levitt 1976).

^l95% bur: free energies calculated (Radzicka and Wolfenden 1988) from Chothia's statistics on whether residues are buried or not (Chothia 1976).

^m14 Å cont: 14 Å contact number (Nishikawa and Ooi 1986).

ⁿZ bur: Zhou-Zhou burability scale (Zhou and Zhou 2004).

^oMPS: Punta-Maritan non-X-ray diffraction/NMR experiment transmembrane scale (Punta and Maritan 2003).

^pP sheet: Palau sheet propensity scale (Palau et al. 1982).

^qP helix: Palau helix propensity scale (Palau et al. 1982).

^rCh/Fa: Chou-Fasman coil propensity scale (Chou and Fasman 1974).

^sRu/Li: Russell-Linding coil propensity scale (Linding et al. 2003b).

for optimized disorder propensities, predetermined coil propensity scales (Ru/Li: Russell-Linding propensity scale) (Linding et al. 2003b) (Ch/Fa: Chou-Fasman coil propensity scale) (Chou and Fasman 1974) do not perform as well in our sliding-window predictor at discriminating crystallographically disordered regions (Fig. 1D, purple- and olive-colored ROC curves). The Kyte-Doolittle scale (Ky/Do) (Kyte and Doolittle 1982) and the Hopp-Woods scale (Ho/Wo; derived from the Levitt scale) (Levitt 1976; Hopp and Woods 1981) yield better, but still suboptimal, performance (Fig. 1D, orange- and cyan-colored ROC curves). Predictions using LORs approach those of the optimized disorder propensities. In agreement with this observation, calculated LORs display strong correlation to optimized disorder propensities ($R^2 = 0.946$, Table 2).

Several structure-derived scales show strong correlations to disorder propensity (Table 2). A residue “burability” scale (Z bur) (Zhou and Zhou 2004) displays the highest correlation to disorder (Z bur, $R^2 = 0.924$), while a residue interactivity scale (Bastolla et al. 2005) is ranked second ($R^2 = 0.909$). Other high-ranked scales measure contacts within a 14 Å sphere (Nishikawa and Ooi 1986) (14 Å cont, $R^2 = 0.840$) and nonbonded interactions for residues well separated in sequence (Oobatake and Ooi 1977) ($R^2 = 0.837$). Another amino acid stability scale (Vihinen et al. 1994), which was optimized on crystal structure temperature factor data using a sliding-window averaging technique, displays good correlation ($R^2 = 0.865$). Finally, structure-based hydrophobicity scales calculated for different protein classes (Cid et al. 1992) correlate with our optimized disorder values ($R^2 = 0.839$ for α/β , $R^2 = 0.834$ for β , and $R^2 = 0.831$ for all averaged). Overall, these scales suggest a relationship between order and the degree to which a residue tends to come into contact with other residues in existing protein structures (Williams et al. 2001; Dosztanyi et al. 2005b). Of note, our disorder propensities also correlate reasonably well with statistical disorder propensities calculated from the DisProt database (Sickmeier et al. 2007), with $R^2 = 0.747$ for all residues.

Among experimentally derived scales, disorder values correlate with polar organic solvent to water side-chain transfer energies (oct/wat; Fig. 3A, $R^2 = 0.788$) (Guy 1985; Radzicka and Wolfenden 1988). Because of satisfied hydrogen-bonding potential in both partitions, these transfer energies reflect the strength of the hydrophobic effect. In contrast, cyclohexane to water transfer energies (chx/wat) (Radzicka and Wolfenden 1988) show a weaker relationship to disorder propensities (Fig. 3B, $R^2 = 0.496$). Because of an absence of hydrogen-bonding potential in one partition (chx), such a scale tends to reflect the preference of an amino acid to form hydrogen bonds with water or to “like” water (hydrophilicity). The

transfer energy scales show that the relative hydrophobicity of an amino acid residue depends on the nonaqueous reference solvent, and they reflect the general diversity of scales that can be considered to be hydrophobic.

Scales comparable to the oct/wat transfer energies display similar tight correlations to disorder. For example, the incomplete hydrophobicity scale of Nozaki and Tanford (1971) (No/Ta), which reflects calculated side-chain free energies of transfer primarily between ethanol and water, is strongly correlated with disorder (Fig. 3C, $R^2 = 0.977$, missing 8 residues). The No/Ta scale excludes several side-chain types that might be expected to display special behavior, such as ionic residues, proline, and cysteine. However, a more complete (and inverted) version of this scale (Ho/Wo) used to predict protein epitopes also exhibits a linear correlation (Fig. 3E, $R^2 = 0.649$). In fact, this scale shows a clear linear relationship with disorder values for uncharged residues ($R^2 = 0.910$, excluding R, K, D, and E). Other scales reflective of the chx/wat transfer energies, such as Kyte-Doolittle (Ky/Do) hydrophobicities for predicting transmembrane segments, correlate weakly with our disorder parameters (Fig. 3D, $R^2 = 0.424$).

As previously described (Linding et al. 2003a), optimized disorder values are *not* well correlated with coil propensity: The strongest association found for this scale type being the Russell-Linding (Ru/Li) scale (Linding et al. 2003b) ($R^2 = 0.326$, Table 2). One might hypothesize that disorder is related to backbone flexibility and that glycine and perhaps small, nonbulky residues such as alanine promote disorder through their decreased steric hindrance of backbone motion. However, backbone flexibility does not appear to be a significant causal factor. Proline is conformationally limited yet is relatively disorder promoting. Glycine, on the other hand, is highly flexible but does not deviate substantially from the hydrophobicity trend.

Ionic residues do not consistently follow a linear relationship in hydrophobicity/disorder plots (e.g., Fig. 3A,E). Such residues pose specific problems for effectively measuring hydrophobicity in experimental settings. Partitioning experiments may not yield accurate hydrophobicities for ionic residues: Water has a stronger dielectric constant than polar organic solvents and is better able to accommodate charge; additionally, transfer energies depend on the solution pH (which affects solute charge) and on any special adjustments (Radzicka and Wolfenden 1988) made in calculations. In contrast to relatively long-range ionic interactions with surrounding solvent, hydrogen bonding and other polar solvent-solute interactions primarily involve close contacts. It is reasonable to conclude that aqueous and *polar* organic phases each offer interactions of approximately equivalent strength to the hydrogen-bonding moieties on amino acid side chains and that the predominant difference between solvation energies in aqueous and

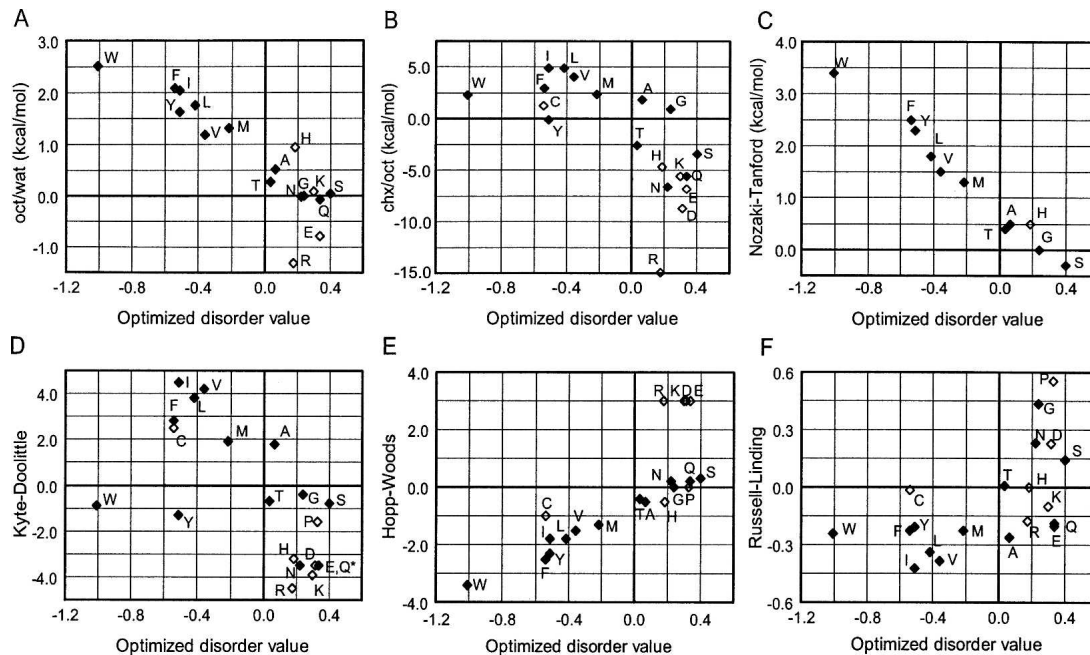


Figure 3. Comparison of optimized disorder values with selected amino acid indices. Numerical values reported for each amino acid (labeled according to residue type) in various scales. (A) Octanol to water transfer energies (Guy 1985; Radzicka and Wolfenden 1988). (B) Cyclohexane to water transfer energies (Radzicka and Wolfenden 1988). (C) Nozaki-Tanford (Nozaki and Tanford 1971) side-chain transfer energies; glycine, as the reference amino acid, is given a value of 0. (D) Kyte-Doolittle hydrophathy (Kyte and Doolittle 1982). (E) Hopp-Woods scale (Hopp and Woods 1983) (derived, in part, from Nozaki-Tanford scale). (F) Russell-Linding coil propensities (Linding et al. 2003b) are plotted against optimized disorder values. Ionic residues, Cys and Pro, are marked by open diamonds.

polar organic solution for *nonionic* amino acid side chains comes specifically from the hydrophobic effect.

Proline and cysteine also pose as special cases in considering hydrophobicity/disorder distributions. Cysteine and proline are not included in the oct/wat transfer energies. Wimley et al. (1996) offer values for all 20 standard amino acid types based on experiments using AcWL-X-LL pentapeptide constructs, but intramolecular interactions with large, hydrophobic side chains present a confounding factor in these experiments. The Levitt scale (Levitt 1976) (on which the Hopp-Woods scale is based) (Hopp and Woods 1981) provides interpolated hydrophobicities for nonionic residues not found in the Nozaki-Tanford scale (Nozaki and Tanford 1971). In the Levitt scale, the only nonionic residues that substantially deviate from the disorder/hydrophobicity trend are proline and cysteine. Proline has a higher disorder propensity than expected from its hydrophobicity, and cysteine has a lower than expected disorder propensity. Proline lacks the primary amine found in all other residues because of its cyclic side-chain binding the backbone nitrogen. This distinction makes proline less polar and restricts rotation about the ϕ torsion angle in protein structures. The rigid proline backbone often acts as a secondary structure disruptor for α -helices (Richardson 1981) and as a β -turn promoter (Chou and Fasman 1974), preventing it from

occurring in the middle of secondary structures. The sulfhydryl group of cysteine can form stabilizing disulfide bonds with other cysteine residues and forms relatively weak hydrogen bonds with water.

A hydrophobic spectrum reflects protein structure characteristics

Amino acid scales often represent a difference between two states. For example, disorder propensities compare the incidences of amino acid residue types in the *disordered state* versus the *ordered state*. The hydrophobic effect and hydrogen-bonding interactions make separate contributions to energetic differences between two states. Certain scales (such as disorder propensities) that quantify the difference between two structural states may be broken down into energetic components that are specifically attributable to the hydrophobic effect (“hydrophobic component”) and/or specifically attributable to polar, hydrogen bonding, or ionic interactions (“hydrophilic component”).

Whereas polar organic phases (such as wet octanol, ethanol, or methanol) and water largely differ because of the hydrophobic effect (for *nonionic* residues, as explained above), cyclohexane and octanol differ primarily in polar and hydrogen-bonding interactions. (Subtracting

oct/wat energies from chx/wat energies produces the chx/oct scale [cyclohexane to octanol transfer energies] (Radzicka and Wolfenden 1988.) Oct/wat hydrophobicities and chx/oct hydrophilicities correlate in part because of polar moieties that both reduce hydrophobicity and increase hydrophilicity. Despite some similarity, however, the oct/wat and chx/oct scales show that residue hydrophobicities and hydrophilicities follow distinct patterns (and that the absolute variance in hydrophilicities is larger than that of hydrophobicities).

Correlations with the oct/wat and chx/oct scale were calculated for various structural propensity scales. Excluding ionic residues, C and P, the oct/wat scale contains neither C nor P (Guy 1985; Radzicka and Wolfenden 1988). The curved line in Figure 4A represents correlations for all perfect, positive linear combinations of oct/wat and chx/oct. The proximity of a scale to this line reflects how well the property can be explained by a combination of the hydrophobic and hydrophilic patterns. Different positions along the spectrum are associated with different protein structural aspects: the hydrophobic end with crystallographic disorder and residue “buriability” (Zhou and Zhou 2004) (opt dis and Z bur, respectively, Fig. 4A), the middle with transmembrane helix potential (MPS: Punta-Maritan non-X-ray diffraction/NMR experiment transmembrane scale; Fig. 4A) (Punta and Maritan 2003), and the hydrophilic end with deep residue burial (Chothia 1976; Radzicka and Wolfenden 1988) (95% bur: free energies calculated from Chothia’s statistics [Chothia 1976] on whether residues are buried or not; Fig. 4A). Kyte and Doolittle (1982) used water-vapor transfer energies and residue burial results of Chothia (1976) as well as some manual adjustment to develop their well-known hydropathy scale, which falls at the hydrophilic end of the spectrum. For predictions utilizing hydrophilicity (e.g., transmembrane predictions), the experimental chx/oct scale may offer improvement over Ky/Do.

The hydropathic spectrum suggests that statistics discerning between fully (or almost fully) buried residues and residues that are more surface exposed reflect a strong influence from *hydrophilic* tendencies (95% bur vs. chx/wat, $R^2 = 0.790$, Table 2) (Radzicka and Wolfenden 1988) not the *hydrophobic* tendencies (95% bur vs. oct/wat, $R^2 = 0.472$, Table 2). This preference is illustrated in the position of the 95% burial scale near the extreme hydrophilic portion of the spectrum (95% bur, Fig. 4B). This position suggests that the environment in the extreme interior of a protein resembles cyclohexane, while the surface environment resembles an amphipathic solvent. Williams et al. (2001) found another scale related to surface proximity (14 Å cont) to be the best among several scales in discriminating between disorder and order. Indeed, comparing this scale with our disorder

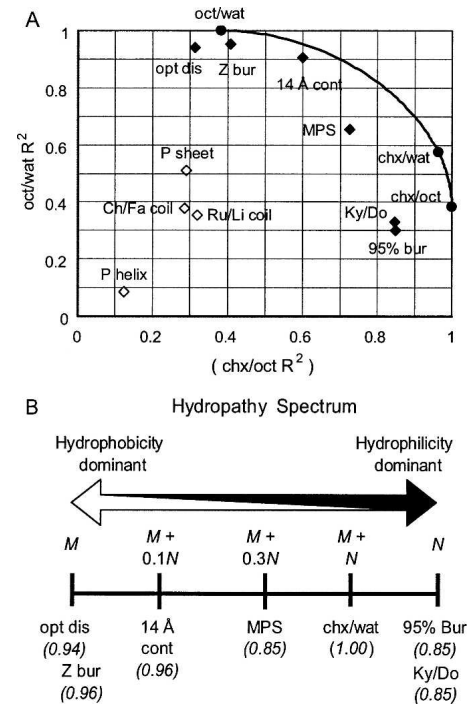


Figure 4. Hydropathy spectrum. (See abbreviation footnote and body of text for explanation of abbreviations.) (A) The relative degree of hydrophobic and hydrophilic components of various scales are approximated by calculating R^2 values (calculated excluding C, P, H, R, K, D, E) against octanol/water (oct/wat) partitioning energies (Guy 1985; Radzicka and Wolfenden 1988) on the Y-axis and cyclohexane/octanol (chx/oct) partitioning energies (Radzicka and Wolfenden 1988) on the X-axis, respectively. The curved line represents correlations for perfect, positive linear combinations of the oct/wat and chx/oct scales. Values for scales that can be largely explained by hydropathy fall near the curve and are represented by black diamonds. Values for the remaining scales are represented by open diamonds. (B) Locations of different scales along the hydropathy spectrum are represented by a linear combination of the oct/wat scale (M) and the chx/oct scale (N) that produces a near-maximum correlation with the scale, thus estimating the relative degrees to which the hydrophilic and hydrophobic components are present, respectively (the general magnitude of N is greater than that of M). Below the name of each scale, in italics, is the strength of association (R^2) of that scale with the respective linear combination of M and N noted above it.

propensities yields a good correlation ($R^2 = 0.840$, Table 2), and the contact scale falls relatively close to our disorder propensities on the hydropathic spectrum (14 Å cont, Fig. 4). The contact number scale somewhat resembles the residue buriability scale, which falls the closest to disorder on the hydropathic spectrum scale (Z bur, Fig. 4).

Propensities to form secondary structures such as α -helices (P helix: Palau helix propensity scale) (Fig. 4A; Palau et al. 1982), β -strands (P sheet: Palau sheet propensity scale) (Fig. 4A; Palau et al. 1982), or coils (Ch/Fa and Ru/Li, Fig. 4A) do not appear to be explained well by hydropathy. Williams et al. (2001) found β -strand propensity

to be somewhat useful in predicting disorder. However, our optimized disorder parameters do not display as strong a correlation with this characteristic ($R^2 = 0.639$, Table 2) as they do with hydrophobicity or residue burial. Disorder and strand propensity seem to be indirectly linked by a common association with hydrophobicity ($R^2 = 0.594$ for P sheet and oct/wat, Table 2), which might explain this relative degree of correlation. Positioning of β -strand propensity and α -helix propensity scales near coil propensity scales on the hydropathy spectrum (Fig. 4A) suggests that the capacity for a residue to form secondary structure is substantially affected by nonhydropathic influences. For example, side-chain branching tends to dictate presence in strands (Street and Mayo 1999; Pal and Chakrabarti 2000). Thus, secondary structure is more influenced by the specific size and shape of residues, while disorder is mainly influenced by the degree to which a residue tends to avoid interaction with water.

Rose et al. (1985) showed that average areas buried for different side chains in folded structures are associated with Nozaki-Tanford hydrophobicities, a simplified interpretation being that hydrophobic parts of side chains tend to be protected and hydrophilic parts exposed to water during folding. This interpretation helps explain the correlation of disorder propensities with various structurally derived scales (Oobatake and Ooi 1977; Nishikawa and Ooi 1986; Zhou and Zhou 2004; Bastolla et al. 2005). The tight disorder/surface burial/hydrophobicity correlations relate disorder propensity to both the structural property of surface burial and the physicochemical property of hydrophobicity. This relationship suggests that residues partition from an ordered state at the protein surface (octanol like) to a disordered state in the surrounding solution (water like). With relatively equal hydrogen-bonding potentials in the two states, a residue can transition into solution when a decrease in the entropy of water (due to exposed hydrophobic surface) (Butler 1937) is overcome with an increase in entropy from becoming disordered. In this case, a residue's hydrophobicity drives partitioning between surface (ordered) and solution (disordered).

Conclusions

Our sequence and profile disorder predictors with tail adjustments perform comparably with DISOPRED2. Although other predictors may yield better predictions in certain circumstances, our sequence-based predictor offers a simple, well-optimized predictor that avoids unknown bias toward special cases and may be useful as a step in a larger bioinformatic sequence analysis. It is useful to know how predictors perform on both internal and chain-terminal sequence regions (Fig. 1).

The disorder that commonly occurs at chain termini (see Table 3) is explained by lack of chain constraint (cysteine, on the other hand, increases constraint through disulfide bonding). Disorder tendency monotonically decreases as position moves away from the terminus in a manner that is essentially independent of whether the amino or carboxyl terminus is involved (Fig. 2C).

Prior research has discussed patterns in disordered protein sequences (Vucetic et al. 2003; Lise and Jones 2005). Others have characterized amino acids in subsets (Li et al. 2000; Weathers et al. 2004; Ferron et al. 2006; Han et al. 2006; Su et al. 2006), for example, disorder-promoting, order-promoting, disorder-neutral. However, our data suggest that much of the variance in disorder behavior among different types of amino acids is reduced quantitatively, at a more fundamental level, by the hydrophobic effect (Butler 1937) alone, and that the transfer of a residue from an amphipathic environment to an aqueous environment (oct/wat) mimics the transfer of a residue from a position offering both hydrophobic and, if needed, hydrogen-bonding interactions (e.g., the protein surface) to a more solvated state (order/disorder transition). Residues with less hydrophobic character have a greater tendency to partition away from the protein and into the surrounding solvent.

In contrast with the hydrophobic effect, our data do not demonstrate a strong direct relationship with hydrogen-bonding interactions, secondary structure, or intrinsic backbone flexibility. Scales that are moderately associated with hydrophobicity (such as the Kyte-Doolittle scale) (Kyte and Doolittle 1982) may be used to predict disorder, but their discriminatory power is not always optimal (Fig. 1D). Other sequence determinants of disorder include cysteine, which forms disulfide bonds, and proline, with a backbone structure that enforces an extended conformation. The degree to which independent positive or negative charges affect disorder is not known, although our data suggest it is relatively small.

Similar patterns of disorder have been found by other researchers (Linding et al. 2003b; Weathers et al. 2004), which include noncrystallographic data sets (Williams et al. 2001). Thus, a hydrophobicity-based model of disorder is probably applicable beyond crystallographic disorder. Our optimized parameters also correlate ($R^2 = 0.747$; see Table 2) with a curated disorder data set provided by DisProt that also includes disordered regions from noncrystallographic data sets. However, distinct patterns may occur in long disordered regions (Peng et al. 2006). In fully unfolded chains, intrinsic conformation (or lack thereof) may become a more important factor in addition to stability, which may explain, beyond window size, why FoldIndex, which uses Kyte-Doolittle hydropathies, performs substantially better on fully disordered sequences than on crystallographic disorder (Esnouf et al. 2006).

Table 3. Data set statistics

Data set	No. of SCOP families	No. of domains	No. of terminal residues excluded	No. of missing (“disordered”) residues	No. of nonmissing (“ordered”) residues
Simple	1912	28128	0	183,902	5,563,922
			18	85,490	4,957,018
			30	66,460	4,473,062
Profile	1773	23386	0	157,195	4,496,369
			18	71,792	3,958,638
			30	55,243	3,531,262

disorder propensities were then held constant and tail adjustment values were optimized for the sequence- and profile-based predictors. For the sequence-based predictor with tail adjustments, a separate disorder propensity was optimized for amino-terminal methionine.

Data set

Optimization and testing data include X-ray crystallographic data from domains in 1912 families in the first fivefold classes of the SCOP version 1.67 (Andreeva et al. 2004) (all α , all β , α/β , $\alpha + \beta$, and multidomain) (Table 3). Structures dated before 2000 or with a resolution >3.0 Å were excluded. Residues found in SEQRES entries with missing C – α carbon coordinates (or occupancy 0) are considered disordered, while remaining residues are considered ordered. Residues in disordered stretches less than four residues long were excluded. Residues between domains were assigned to a nearby domain. Domains with certain issues (e.g., mismatches between SEQRES sequence and structural sequence, or atom occupancies out of the 0–1 range) were excluded.

In optimizing the disorder propensities for sequence and profile-based predictors, residues at termini (18 from the end of the chain or polyhistidine sequence, if present) were excluded from ROC score calculation to avoid spurious effects from compositional bias at sequence ends (notably for methionine and histidine). Terminal residues were included in the ROC score calculation when optimizing tail adjustment parameters.

Different subsets of the data were periodically selected for measuring parameter performance during optimization. To reduce overrepresentation of certain families or individual structures, a data subset included one protein from each family with more than six representatives; families with five or less representatives were represented in proportion to family size (i.e., one-member families represented $\sim 1/5$ of the time; two-member families $\sim 2/5$ of the time, etc.).

Profiles

The same PSI-BLAST results were used both in developing and testing our profile-based predictors and in testing DISOPRED2. Profiles were built using default values (three iterations, E-value cutoff 0.001). Final alignments included segments from up to 1000 different sequences. Modified COMPASS code (Sadreyev and Grishin 2003) was used to generate final profiles. Use of position-specific independent counts reduced overrepresentation of closely related sequences (Sunyaev et al. 1999). Pseudo-counts (Tatusov et al. 1994) were generated using the BLOSUM62 (Henikoff and Henikoff 1992)

matrix. At each position in a profile, the pseudo-count values for the set of all standard residue types were normalized to an exact sum of 1, producing fractional weights that estimate the degree to which each residue-type characterizes that position in the alignment.

Cross validation

To estimate the accuracy of our disorder predictions and to evaluate the consistency of our method, a five-way cross validation was performed. Crystallographic data was split into five subsets. Four out of the five data subsets form a “training set” for optimization of predictor parameters, and the remaining data subset forms a “testing set” for assessment of performance for each of five optimizations. The five sets of optimized parameters were averaged to give final values. Domains were randomly assigned into the five data subsets by SCOP family (see Table 4) so that highly similar sequences could not be shared by concomitant training and testing sets.

Analysis

Our optimized amino acid residue disorder propensities were compared with several amino acid property scales. ROC curves measure performance for the optimized predictors, for DISOPRED2, and with various indices substituted for optimized disorder parameters in the simple sequence predictor. For each different disorder predictor, ROC curves are calculated for each of the five data subsets (see preceding section on cross validation); then corresponding points from the five resulting ROC curves are averaged to produce final ROC curves. In the case of our optimized predictors, optimized parameters are averaged from the results of five optimization runs using five training sets (see section on cross validation); for each of these optimization runs, a ROC curve is calculated using the corresponding testing data set, which was excluded during optimization. As during optimization, certain residues were excluded from ROC curve calculations and other performance analyses (see Data set section).

Additionally, squared correlation coefficients (R^2) were calculated for our disorder propensities and 516 published amino acid property scales found in the AAindex database (Kawashima and Kanehisa 2000), quantifying the degree of covariance between our optimized parameters and each of these scales.

Table 4. Number of SCOP families and individual domains represented in each cross validation data subset

	Data subset	No. SCOP families	No. domains
Sequence predictor data	1	382	5629
	2	382	6034
	3	382	5191
	4	383	5709
	5	383	5565
Profile predictor data	1	359	4645
	2	354	5129
	3	357	4275
	4	347	4615
	5	356	4722

Relationships between scales derived from structural data and scales derived from partitioning experiments were examined.

Our optimization-derived disorder propensities were compared with disorder propensities derived statistically (as LORs) using samplings of data from the data sets used for optimization, balanced in an equivalent fashion. LORs behave in a linear fashion and provide a good statistical measure of relative propensity. The disorder versus order LOR for amino acid type i is calculated as follows:

$$LOR_i = \ln \left(\frac{\left(\frac{P(i|diso)}{1 - P(i|diso)} \right)}{\left(\frac{P(i|ord)}{1 - P(i|ord)} \right)} \right)$$

where $P(i | diso)$ denotes the probability of the occurrence of residue type i in the disordered state and $P(i | ord)$ denotes the probability of residue type i in the ordered state. LORs are also calculated in a similar manner for profiles using the profile frequencies of each amino acid type at ordered and disordered positions. LORs were calculated from DisProt (Sickmeier et al. 2007) by considering noted residues as disordered and all other residues as ordered. N-terminal methionines were treated as a separate residue category.

Acknowledgment

This work was supported by a NIH grant (GM67165) to N.V.G.

References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402. doi: 10.1093/nar/25.17.3389.
- Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. 2004. SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**: D226–D229. doi: 10.1093/nar/gkh039.
- Bastolla, U., Porto, M., Roman, H.E., and Vendruscolo, M. 2005. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins* **58**: 22–30.
- Butler, J.A.V. 1937. The energy and entropy of hydration of organic compounds. *Trans. Faraday Soc.* **33**: 229–236.
- Cai, W. and Shao, X. 2002. A fast annealing evolutionary algorithm for global optimization. *J. Comput. Chem.* **23**: 427–435.
- Chandler, D. 2005. Interfaces and the driving force of hydrophobic assembly. *Nature* **437**: 640–647.
- Cheng, J., Sweredoski, M.J., and Baldi, P. 2005. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining Knowledge Discovery* **11**: 213–222.
- Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**: 1–12.
- Chou, P.Y. and Fasman, G.D. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **13**: 211–222.
- Cid, H., Bunster, M., Canales, M., and Gazitua, F. 1992. Hydrophobicity and structural classes in proteins. *Protein Eng.* **5**: 373–375.
- Coeytaux, K. and Poupon, A. 2005. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* **21**: 1891–1900.
- Creighton, T.E. 1993, 2002. *Proteins: Structures and molecular properties*, 2nd ed. W.H. Freeman Company, New York.
- Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. 2005a. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**: 3433–3434.
- Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. 2005b. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**: 827–839.
- Dosztanyi, Z., Sandor, M., Tompa, P., and Simon, I. 2007. Prediction of protein disorder at the domain level. *Curr. Protein Pept. Sci.* **8**: 161–171.
- Esnouf, R.M., Hamer, R., Sussman, J.L., Silman, I., Trudgian, D., Yang, Z.R., and Prilusky, J. 2006. Honing the in silico toolkit for detecting protein disorder. *Acta Crystallogr. D Biol. Crystallogr.* **62**: 1260–1266.
- Ferron, F., Longhi, S., Canard, B., and Karlin, D. 2006. A practical overview of protein disorder prediction methods. *Proteins* **65**: 1–14.
- Galzitskaya, O.V., Garbuzynskiy, S.O., and Lobanov, M.Y. 2006a. FoldUnfold: Web server for the prediction of disordered regions in protein chain. *Bioinformatics* **22**: 2948–2949.
- Galzitskaya, O.V., Garbuzynskiy, S.O., and Lobanov, M.Y. 2006b. Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput. Biol.* **2**: e177. doi: 10.1371/journal.pcbi.0020177.
- Gribskov, M. and Robinson, N.L. 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* **20**: 25–33.
- Guy, H.R. 1985. Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys. J.* **47**: 61–70.
- Han, P., Zhang, X., Norton, R.S., and Feng, Z.P. 2006. Predicting disordered regions in proteins based on decision trees of reduced amino acid composition. *J. Comput. Biol.* **13**: 1723–1734.
- Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915–10919.
- Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., and Noguchi, T. 2007. POODLE-L: A two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* doi: 10.1093/bioinformatics/btm302.
- Hopp, T.P. and Woods, K.R. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.* **78**: 3824–3828.
- Hopp, T.P. and Woods, K.R. 1983. A computer program for predicting protein antigenic determinants. *Mol. Immunol.* **20**: 483–489.
- Jones, D.T. and Ward, J.J. 2003. Prediction of disordered regions in proteins from position specific score matrices. *Proteins (Suppl 6)* **53**: 573–578.
- Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**: 1–63.
- Kawashima, S. and Kanehisa, M. 2000. AAindex: Amino acid index database. *Nucleic Acids Res.* **28**: 374. doi: 10.1093/nar/28.1.374.
- Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**: 105–132.
- Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**: 59–107.
- Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z. 1999. Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform.* **10**: 30–40.
- Li, X., Obradovic, Z., Brown, C.J., Garner, E.C., and Dunker, A.K. 2000. Comparing predictors of disordered protein. *Genome Inform. Ser. Workshop Genome Inform.* **11**: 172–184.
- Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. 2003a. Protein disorder prediction: Implications for structural proteomics. *Structure* **11**: 1453–1459.
- Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**: 3701–3708. doi: 10.1093/nar/gkg5.
- Lise, S. and Jones, D.T. 2005. Sequence patterns associated with disordered regions in proteins. *Proteins* **58**: 144–150.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087–1092.
- Nishikawa, K. and Ooi, T. 1986. Radial locations of amino acid residues in a globular protein: Correlation with the sequence. *J. Biochem.* **100**: 1043–1047.
- Nozaki, Y. and Tanford, C. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **246**: 2211–2217.
- Oobatake, M. and Ooi, T. 1977. An analysis of non-bonded energy of proteins. *J. Theor. Biol.* **67**: 567–584.
- Pal, D. and Chakrabarti, P. 2000. β -sheet propensity and its correlation with parameters based on conformation. *Acta Crystallogr. D Biol. Crystallogr.* **56**: 589–594.
- Palau, J., Argos, P., and Puigdomenech, P. 1982. Protein secondary structure. Studies on the limits of prediction accuracy. *Int. J. Pept. Protein Res.* **19**: 394–401.
- Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**: 208. doi: 10.1186/1471-2105-7-208.

- Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. 2005. FoldIndex: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**: 3435–3438.
- Punta, M. and Maritan, A. 2003. A knowledge-based scale for amino acid membrane propensity. *Proteins* **50**: 114–121.
- Radzicka, A. and Wolfenden, R. 1988. Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27**: 1664–1670.
- Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**: 167–339.
- Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., and Dunker, A.K. 1997. Identifying disordered regions in proteins from amino acid sequence. *Proc. IEEE Int. Conf. Neural Networks* **1**: 90–95.
- Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229**: 834–838.
- Rose, G.D., Fleming, P.J., Banavar, J.R., and Maritan, A. 2006. A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci.* **103**: 16623–16633.
- Sadreyev, R. and Grishin, N. 2003. COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326**: 317–336.
- Shimizu, K., Muraoka, Y., Hirose, S., Tomii, K., and Noguchi, T. 2007. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* **8**: 78. doi: 10.1186/1471-2105-8-78.
- Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., et al. 2007. DisProt: The database of disordered proteins. *Nucleic Acids Res.* **35**: D786–D793. doi: 10.1093/nar/gkl.
- Street, A.G. and Mayo, S.L. 1999. Intrinsic β -sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci.* **96**: 9074–9076.
- Su, C.T., Chen, C.Y., and Ou, Y.Y. 2006. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics* **7**: 319. doi: 10.1186/1471-2105-7-319.
- Sugase, K., Dyson, H.J., and Wright, P.E. 2007. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447**: 1021–1025.
- Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G., and Kuznetsov, E.N. 1999. PSIC: Profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* **12**: 387–394.
- Tatusov, R.L., Altschul, S.F., and Koonin, E.V. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci.* **91**: 12091–12095.
- Tomii, K. and Kanehisa, M. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **9**: 27–36.
- Uversky, V.N., Gillespie, J.R., and Fink, A.L. 2000. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* **41**: 415–427.
- Vihinen, M., Torkkila, E., and Riikonen, P. 1994. Accuracy of protein flexibility predictions. *Proteins* **19**: 141–149.
- Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. 2003. Flavors of protein disorder. *Proteins* **52**: 573–584.
- Vullo, A., Bortolami, O., Pollastri, G., and Tosatto, S.C. 2006. Spritz: A server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.* **34**: W164–W168. doi: 10.1093/nar/gkl166.
- Wang, L. and Donald, B.R. 2006. A data-driven, systematic search algorithm for structure determination of denatured or disordered proteins. *Comput. Syst. Bioinformatics Conf.* **5**: 67–78.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and 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.
- Waugh, D.F. 1954. Protein–protein interactions. *Adv. Protein Chem.* **9**: 325–437.
- Weathers, E.A., Paulaitis, M.E., Woolf, T.B., and Hoh, J.H. 2004. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett.* **576**: 348–352.
- Williams, R.M., Obradovic, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J., and Dunker, A.K. 2001. The protein non-folding problem: Amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.* **6**: 89–100.
- Wimley, W.C., Creamer, T.P., and White, S.H. 1996. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* **35**: 5109–5124.
- Yang, Z.R., Thomson, R., McNeil, P., and Esnouf, R.M. 2005. RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**: 3369–3376.
- Zhou, H. and Zhou, Y. 2004. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* **54**: 315–322.