
Thermodynamic environments in proteins: Fundamental determinants of fold specificity

JAMES O. WRABL,^{1,2} SCOTT A. LARSON,¹ AND VINCENT J. HILSER¹

¹Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, Texas 77555-1055, USA

²Present address: Department of Biochemistry/HHMI, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA

(RECEIVED January 24, 2002; FINAL REVISION May 15, 2002; ACCEPTED May 16, 2002)

Abstract

To investigate the relationship between an amino acid sequence and its corresponding protein fold, a database of thermodynamic stability information was assembled as a function of residue type from 81 nonhomologous proteins. This information was obtained using the COREX algorithm, which computes an ensemble-based description of the native state of proteins. Dissection of the COREX stability constant into its fundamental energetic components resulted in 12 thermodynamic environments describing the tertiary architecture of protein folds. Because of the observation that residue types partitioned unequally between these environments, it was hypothesized that thermodynamic environments contained energetic information that connected sequence to fold. To test the significance of this hypothesis, the thermodynamic stability information was incorporated into a three-dimensional-to-one-dimensional scoring matrix, and simple fold recognition experiments were performed in a manner such that information about the fold target was never included in the scoring. For 60 out of 81 fold targets, the correct sequence for the target scored in the top 5% of 3858 decoy sequences, with Z-scores ranging from 1.76 to 12.23. Furthermore, a scoring matrix assembled from the residues of 40 nonhomologous all- α proteins was used to thread sequences against 12 nonhomologous all- β protein targets. In 10 of 12 cases, sequences known to adopt the native all- β structure scored in the top 5% of 3858 decoy sequences, with Z-scores ranging from 1.99 to 7.94. These results indicate that energetic information encoded by thermodynamic environments represents a fundamental property of proteins that underlies classifications based on secondary structure.

Keywords: Native state ensemble; threading and fold recognition; protein structure prediction; residue thermodynamics; protein stability; secondary structure

Supplemental material: See www.proteinscience.org.

Although the precise mechanisms that amino acid sequences use to arrive at their native conformations are still under investigation, the fact that many proteins fold spontaneously into unique structures indicates that given the proper solvent context, all of the thermodynamic information necessary to define the final tertiary structure is con-

tained in the primary sequence (Anfinsen 1973). For this reason, it should be possible to predict the fold of a protein from its sequence alone. Indeed, information from sequence analysis (Koonin et al. 2000), secondary structure prediction (Jones et al. 1999), threading (Panchenko et al. 2000), and even ab initio methods (Bonneau et al. 2001) can be used in many cases to correctly assign folds to sequences.

It is hypothesized that ensemble-based structural energetics, as calculated by the COREX algorithm (Hilser and Freire 1996), can be used as a tool for improved characterization of the enthalpic and entropic relationships between amino acid sequences and the folds they adopt. The COREX algorithm models the native state of a protein as a statistical

Reprint requests to: V.J. Hilser, Department of Human Biological Chemistry and Genetics, 5.162 Medical Research Building, University of Texas Medical Branch, Galveston, TX 77555-1055; e-mail: vince@hbcg.utmb.edu; fax: (409) 747-6816.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.0203202>.

thermodynamic ensemble of partially unfolded conformational microstates. For each microstate i in the ensemble, the Gibbs free energy is calculated from a previously described parameterization based on surface area and conformational entropy terms, as described in Materials and Methods (Baldwin 1986; Lee et al. 1994; Xie and Freire, 1994; Gomez et al. 1995; D'Aquino et al. 1996; Habermann and Murphy 1996). By determining whether each residue is folded or unfolded in each microstate, a residue-specific energetic description of each protein is calculated. The validity of the data produced by the COREX algorithm has been assessed through comparisons of predicted protection factors with protection factors obtained from native-state hydrogen exchange experiments (Hilser and Freire 1996; Hilser et al. 1998). The agreement between calculated and experimental protection factors shows that the calculated native-state ensemble provides a reasonable representation of the actual native-state ensemble.

One statement of the protein folding process asserts that an amino acid sequence does not merely fold into a structure, but folds into an ensemble approximated by a single time-averaged conformation, such as an X-ray crystal structure. Therefore, in addition to providing insight into theoretical understanding of protein structure and energetics, information derived from the native-state ensemble may have practical application in fold recognition and structure prediction. As a first step in this direction, previous work from this laboratory (Wrabl et al. 2001) has shown that when a database of protein targets is subjected to a prediction scheme in which each structure is divided into three thermodynamic environments defined by ensemble-based structural energetics (i.e., low stability, medium stability, and high stability), the resulting fold-recognition capability is similar to a prediction scheme based on secondary structure propensities (i.e., helix, sheet, and other).

In the present work, it is shown that more precise thermodynamic environments can be defined based on the relative contributions of solvation and conformational entropy to the stability of the different microstates in the ensemble. The explicit incorporation of enthalpic and entropic terms in this formalism greatly improves the fold-prediction capability. More importantly, it is observed that a database of energetic information constructed from α -helical proteins (i.e., little or no β -sheet) is sufficient to assign sequences to folds for all- β proteins (i.e., little or no α -helix). This indicates that information derived from COREX ensemble-based structural energetics represents a fundamental descriptor of proteins that transcends secondary structure classifications.

Results

Calculation of the native state ensemble using COREX

The COREX algorithm is a statistical thermodynamic model in which a native protein is depicted as an ensemble

of states rather than as a single static structure (Hilser and Freire 1996). The high-resolution nuclear magnetic resonance or X-ray structure file for a protein functions as a template, and an ensemble of partially unfolded microstates is generated by folding and unfolding different regions of the protein in all possible combinations (Hilser and Freire 1996, Wrabl et al. 2001). Under equilibrium conditions, the probability of any given conformational microstate, i , is given by

$$P_i = \frac{K_i}{\sum_{i=1}^{N_{states}} K_i} = \frac{K_i}{Q} \quad (1)$$

where $K_i = [\exp(-\Delta G_i/RT)]$ is the statistical weight of each microstate, and the summation in the denominator is the partition function, Q , for the system. The Gibbs free energy for each microstate, ΔG_i , relative to the fully-folded reference state is calculated from surface area- and conformational entropy-based parameterizations described previously (Hilser and Freire 1996), and illustrated in Materials and Methods. In other words, the ΔG_i , of each microstate arises from differences in solvation of apolar and polar surface area, and from differences in conformational entropy between each microstate and the reference state. Therefore, dividing the free energy into its component terms gives

$$\Delta G_i = \Delta G_{apolar,i} + \Delta G_{polar,i} + \Delta G_{confS,i} \quad (2)$$

As Equation 2 indicates, different values for the component contributions can provide similar magnitudes for ΔG_i . In other words, different microstates can have similar stabilities but different thermodynamic mechanisms for achieving that stability. This is an important concept that becomes useful when defining stability environments within protein structures.

Residue-specific thermodynamic properties

Compiling residue-specific thermodynamic information for proteins and applying it to fold recognition is complicated by at least two factors. First, protein folding is highly cooperative. As a consequence, there has been no thermodynamically rigorous relationship established between experimentally measured thermodynamic quantities, such as the effect of point mutations on stability, and the contribution of an individual amino acid to the stability of a fold. Second, it is not clear how thermodynamic information can be used to augment existing fold-recognition algorithms. For example, how can the entropy of a target structure be incorporated into a threading algorithm?

Fortunately, the ensemble-based formalism embodied by the COREX algorithm provides a vehicle for deriving resi-

due-specific information that implicitly reflects the properties of the ensemble as a whole. Previously, it has been shown that the probabilities calculated from Equation 1 can be used to derive an important statistical descriptor of the equilibrium (Hilser and Freire 1996). Defined as the residue stability constant, κ_j , this quantity is the ratio of the summed probability of all states in the ensemble in which a particular residue, j , is in a folded conformation ($\sum P_{f,j}$) to the summed probability of all states in which residue j is in an unfolded (i.e., nonfolded) conformation ($\sum P_{nf,j}$):

$$\kappa_{f,j} = \frac{\sum P_{f,j}}{\sum P_{nf,j}} \quad (3)$$

Equation 3, in turn, can be used to define a residue-specific free energy of folding for the protein:

$$\Delta G_{f,j} = -RT \ln \kappa_{f,j} \quad (4A)$$

which can be expanded to give:

$$\Delta G_{f,j} = RT \ln Q_{nf,j} - RT \ln Q_{f,j} \quad (4B)$$

where $Q_{nf,j}$ and $Q_{f,j}$ are the subpartition functions for states in which residue j is unfolded and folded, respectively. As indicated by the functional form of Equation 4B, the residue-specific free energy provides the difference in energy between the subensembles in which each residue is folded and unfolded. In other words, the residue stability constant does not provide the contribution of each amino acid to the stability of a protein. Rather, it provides the relative stability of that region of the protein, implicitly considering the contribution of all amino acids in the protein toward the observed stability at that position.

As shown in Figure 1 and shown previously (Hilser and Freire 1996), the stability constants (Equation 3) provide a residue-specific description of the regional differences in stability within a protein structure. The importance of this quantity from the point of view of fold recognition is twofold. First, as mentioned above, the stability constant can be compared directly to protection factors obtained from native-state hydrogen exchange experiments, thus providing an experimentally verifiable residue-specific description of the ensemble. Second, as amino acids are nonrandomly distributed across high, medium, and low stability environments, the stability constant as a function of residue position provides a convenient one-dimensional representation of the three-dimensional structure. It has been established that such a description contains significant structure-encoding information (Wrabl et al. 2001).

To investigate the possibility of identifying additional thermodynamic determinants of fold specificity beyond the three stability classes previously identified and shown in

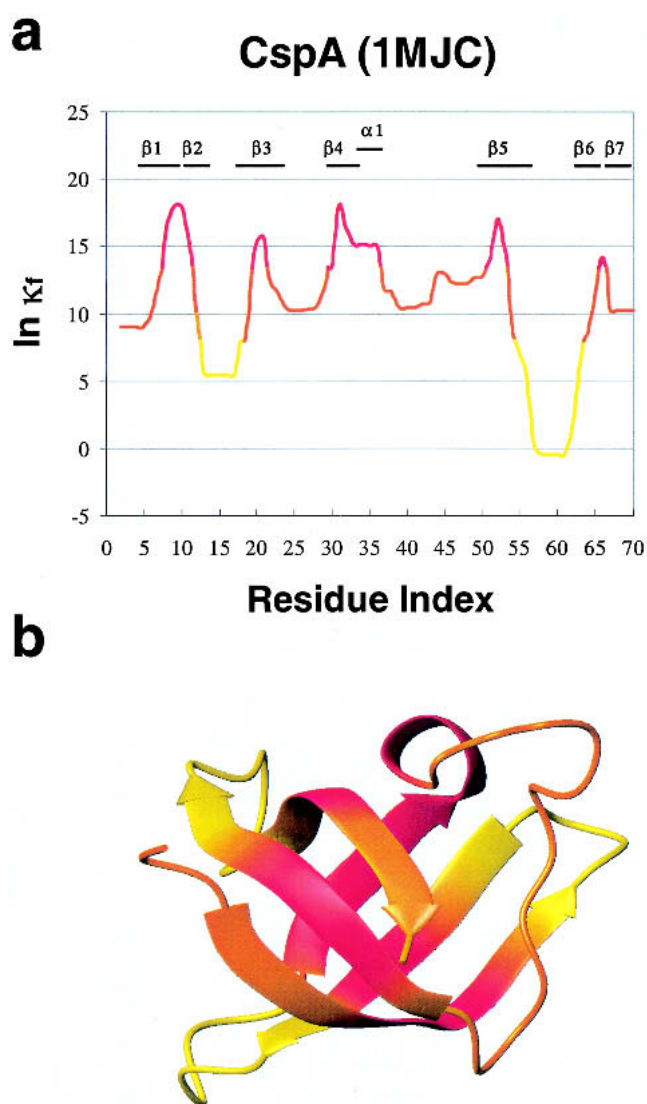


Fig. 1. Results of a COREX calculation for the bacterial cold-shock protein cspA (Protein Data Bank 1mjc). (a) Plot of calculated thermodynamic stability, $\ln \kappa_{f,j}$ (Equation 3), as a function of residue number for cspA. The simulated temperature was 25°C. Regions of relatively high, medium, and low stability, as defined in Equations 19 through 21, are shown in blue, green, and red, respectively. Secondary structure elements, as defined by the program DSSP (Kabsch and Sander 1983) are labeled. (b) The relative calculated stabilities of each residue in the 1mjc crystal structure. Note that a given secondary structural element is predicted to have regions of varying stability, and that the most stable regions of the molecule are often, but not necessarily, within the hydrophobic core.

Figure 1, a database of 81 nonhomologous proteins was analyzed using the COREX algorithm (Table 1). Figure 1 shows COREX results for the cold-shock protein CspA (PDB 1mjc), in which the structure is color coded according to the different stability environments. As noted previously, high-stability regions are generally found in the core of the protein, and low stability regions are found mostly in the loops. However, what are not revealed from analysis of the

Table 1. Proteins used in the COREX thermodynamic database

Number	PDB ID	Length	SCOP Class ^a	SCOP Fold ^a
1.	1a1iA	85	small proteins	classic zinc finger C2H2
2.	1a6s	87	all α	retroviral matrix protein
3.	1a8o	70	all α	acyl carrier protein like
4.	1aa3	63	$\alpha + \beta$	anti LPS factor/RecA domain
5.	1aba	87	α/β	thioredoxin fold
6.	1adr	76	all α	λ repressor like DNA-binding domain
7.	1aiw	62	all β	WW domain-like
8.	1an4A	65	all α	helix loop helix DNA-binding domain
9.	1aoiB	83	all α	histone fold
10.	1avyC	68	coiled coil proteins	parallel coiled coil
11.	1b9gA	57	small proteins	insulin-like
12.	1bdd	60	all α	bacterial Ig/albumin-binding domain
13.	1bdo	80	all β	barrel sandwich hybrid
14.	1bf4A	63	$\alpha + \beta$	IL8 like
15.	1bg8A	76	all α	protein HNS-dependent expression A
16.	1bo9A	73	all α	annexin
17.	1c1yB	77	α/β	P-loop containing NTP hydrolases
18.	1cc5	83	all α	cytochrome C
19.	1chc	68	small proteins	RING finger domain C3HC4
20.	1ctf	68	$\alpha + \beta$	ribosomal protein L7/12 C-fragment
21.	1cyo	88	$\alpha + \beta$	cytochrome b5
22.	1d3bB	81	all β	Sm motif, SNRNP
23.	1doqA	69	all α	SAM domain-like
24.	1dt4A	73	$\alpha + \beta$	KH domain
25.	1egwA	71	$\alpha + \beta$	SRF-like
26.	1eo0A	77	all α	N cbl-like
27.	1fgp	70	all β	N domains of minor coat protein g3p
28.	1gdc	72	small proteins	glutacorticoid receptor DNA binding
29.	1hcrA	52	all α	DNA/RNA binding 3 helical bundle
30.	1hdj	77	all α	long α hairpin
31.	1hoe	74	all β	α amylase inhibitor tendamistat
32.	1hp8	68	all α	p8-MTCPI
33.	1iieA	75	all α	MHC class II extoplasmic trimerization
34.	1iro	53	small proteins	rubredoxin-like
35.	1isuA	62	small proteins	HiPIP
36.	1kdxA	81	all α	Kix domain of CBP
37.	1kjs	74	all α	anaphylotoxins (complement system)
38.	1kveA	63	$\alpha + \beta$	yeast killer toxins
39.	1kwaA	88	all β	PDZ domain-like
40.	1mho	88	all α	EF hand-like
41.	1mjc	69	all β	OB fold
42.	1mknA	59	small proteins	midkine
43.	1mof	53	peptides	MoMLV p15fragment (residues 409–426)
44.	1mwpA	96	$\alpha + \beta$	SRCR-like
45.	1nhm	79	all α	HMG box

Table 1. Continued

Number	PDB ID	Length	SCOP Class	SCOP Fold
46.	1nkl	78	all α	saposin
47.	1npsA	88	all β	crystallins proteins yeast killer toxin
48.	1nre	81	all α	α_2 macroglobulin protein (RAP)
49.	1ntcA	91	all α	FIS-like
50.	1nxb	62	small proteins	snake toxin-like
51.	1opd	85	$\alpha + \beta$	histidine containing phosphocarrier proteins
52.	1otfA	59	$\alpha + \beta$	tautomerase/MIF
53.	1pcfA	66	$\alpha + \beta$	transcriptional coactivator PC4 C-domain
54.	1pgb	56	$\alpha + \beta$	α -grasp (ubiquitin-like)
55.	1plc	99	all β	cupredoxins
56.	1ptf	87	$\alpha + \beta$	PHr proteins
57.	1ptq	50	small proteins	proteins kinase (cys2, phorbol binding)
58.	1ptx	64	small proteins	knottins
59.	1qa4A	56	peptides	HIV-1 Nef protein fragments
60.	1qgwB	67	$\alpha + \beta$	Nonglobular $\alpha \beta$ subunits of globular
61.	1qqvA	67	all α	thermostable subdomain from chicken villin
62.	1r1bA	56	all α	SIS/NS1 RNA-binding domain
63.	1rop	56	all α	ROP-like
64.	1rzl	91	all α	bifunctional inhibitor/lipid transfer protein
65.	1shg	57	all β	SH3-like barrel
66.	1sknP	74	all α	binding domain of skn-1
67.	1svfB	62	coiled coil proteins	stalk segment of viral fusion proteins
68.	1tbaA	67	all α	TAFII 230 nTBP-binding fragment
69.	1tgsI	56	small proteins	ovomucoid/PCI-1-like inhibitors
70.	1trlA	62	all α	thermolysin-like metalloproteases C-domain
71.	1ugiD	82	$\alpha + \beta$	cystatin-like
72.	1utg	70	all α	uteroglobin-like
73.	1vcc	77	$\alpha + \beta$	DNA topoisomerase I domain
74.	2abd	86	all α	acyl CoA binding protein-like
75.	2bopA	85	$\alpha + \beta$	ferredoxin-like
76.	2ci2I	65	$\alpha + \beta$	CI2 family of serine protease inhibitors
77.	2knt	58	small proteins	BPTI-like
78.	2spgA	66	all β	β clip
79.	3eipA	84	$\alpha + \beta$	FKBP-like
80.	3ncmA	92	all β	immunoglobulin-like β sandwich
81.	5hpgA	84	small proteins	kringle-like

^a The structural classification for determining extent of homology as found in the SCOP database (Murzin et al. 1995).

stability constants alone are the underlying thermodynamic determinants for the stability of the different regions. For instance, the stability constants for loop residues 13–18 and 55–63 are all classified as low stability. Are the enthalpic and entropic contributions to the stability similar for both loops? How much of the solvation contribution is owing to apolar and polar surface? Of the total stability, what fraction is owing to conformational entropy as opposed to solvation?

To address the residue-specific enthalpic and entropic determinants, we consider the following development. The ΔG_i for each microstate i in the ensemble is composed of solvation (*sol*) and conformational (*conf*) entropy terms as described by Equation 2. Rewriting Equation 2 in terms of the enthalpic and entropic components gives

$$\Delta G_i = \Delta H_{i,\text{sol}} - T(\Delta S_{i,\text{sol}} + \Delta S_{i,\text{conf}}) \quad (5)$$

Each of the solvation terms in Equation 5 can be further expanded into contributions based on apolar and polar surface area:

$$\begin{aligned} \Delta G_i = & (\Delta H_{i,\text{sol},\text{apolar}} + \Delta H_{i,\text{sol},\text{polar}}) \\ & - T(\Delta S_{i,\text{sol},\text{apolar}} + \Delta S_{i,\text{sol},\text{polar}}) \\ & - T(\Delta S_{i,\text{conf}}) \end{aligned} \quad (6)$$

However, the identical values for the apolar and polar areas of each microstate are used for the respective terms in the enthalpy and entropy calculations. Therefore, the absolute values for the enthalpy and entropy terms for a given area type are simply related by constants k_1 (for apolar area) and k_2 (for polar area), yielding the following expression:

$$\begin{aligned} \Delta G_i = & (\Delta H_{i,\text{sol},\text{apolar}} + \Delta H_{i,\text{sol},\text{polar}}) \\ & - T(k_1 \Delta H_{i,\text{sol},\text{apolar}} + k_2 \Delta H_{i,\text{sol},\text{polar}}) \\ & - T(\Delta S_{i,\text{conf}}) \end{aligned} \quad (7)$$

Grouping area types together and simplifying gives

$$\begin{aligned} \Delta G_i = & [(\Delta H_{i,\text{sol},\text{apolar}}) * (1 - T * k_1)] \\ & + [(\Delta H_{i,\text{sol},\text{polar}}) * (1 - T * k_2)] \\ & - T(\Delta S_{i,\text{conf}}) \end{aligned} \quad (8)$$

Equation 8 reveals that for a given free energy and conformational entropy, the relative contribution of polar and apolar surface to the solvation free energy can be ascertained simply from the ratio of polar to apolar enthalpy for each state.

To arrive at a residue-specific contribution of polar and apolar solvation, we need only consider that for a given thermodynamic parameter (i.e., enthalpy or entropy), an average excess quantity can be defined, which represents the population-weighted contribution of all states in the ensemble. For instance, the average excess enthalpy and entropy can be defined as follows:

$$\langle \Delta H \rangle = \sum_{i=1}^{N_{\text{states}}} P_i \cdot \Delta H_i = \sum_{i=1}^{N_{\text{states}}} \frac{K_i \cdot \Delta H_i}{Q} \quad (9A)$$

$$\langle \Delta S \rangle = \sum_{i=1}^{N_{\text{states}}} P_i \cdot \Delta S_i = \sum_{i=1}^{N_{\text{states}}} \frac{K_i \cdot \Delta S_i}{Q} \quad (9B)$$

Following from Equations 9A and 9B, residue-specific descriptors of the polar and apolar enthalpy were defined accordingly. The polar component of the enthalpy was defined as the difference between the average excess polar enthalpy from the subensemble in which residue j is folded ($\langle \Delta H_{\text{pol},f,j} \rangle$) and the average excess polar enthalpy from the sub-ensemble in which residue j is unfolded ($\langle \Delta H_{\text{pol},nf,j} \rangle$):

$$\Delta H_{\text{pol},j} = \langle \Delta H_{\text{pol},f,j} \rangle - \langle \Delta H_{\text{pol},nf,j} \rangle \quad (10)$$

where

$$\langle \Delta H_{\text{pol},f,j} \rangle = \sum_{i=1}^{N_{j,\text{folded}}} \left(\frac{(\Delta H_{\text{pol},f,i} \cdot e^{-\Delta G_i/RT})}{Q_{f,j}} \right) \quad (11)$$

$$\langle \Delta H_{\text{pol},nf,j} \rangle = \sum_{i=1}^{N_{j,\text{not folded}}} \left(\frac{(\Delta H_{\text{pol},nf,i} \cdot e^{-\Delta G_i/RT})}{Q_{nf,j}} \right) \quad (12)$$

It is important to note that the summations in Equations 11 and 12 are only over the subensembles in which residue j is folded and unfolded, respectively, and the parameters $Q_{f,j}$ and $Q_{nf,j}$ are the subpartition functions for those subensembles. By identical reasoning, the residue-specific apolar component to the enthalpy of residue j and the residue-specific conformational entropy component of residue j are defined as

$$\Delta H_{\text{apol},j} = \langle \Delta H_{\text{apol},f,j} \rangle - \langle \Delta H_{\text{apol},nf,j} \rangle \quad (13)$$

$$\Delta S_{\text{conf},j} = \langle \Delta S_{\text{conf},f,j} \rangle - \langle \Delta S_{\text{conf},nf,j} \rangle \quad (14)$$

As in the case with the residue stability constant (Equation 3), the expressions for the residue-specific $\Delta H_{\text{apol},j}$, $\Delta H_{\text{pol},j}$, and $\Delta S_{\text{conf},j}$ do not provide the contributions of residue j to the respective overall thermodynamic properties. Instead, Equations 10, 13, and 14 reflect the average thermodynamic environments of that residue, accounting implicitly for the contribution of all the amino acids over all the states in the ensemble.

Residue-specific thermodynamic environments

Using Equations 3, 10, 13, and 14, thermodynamic environments were empirically defined so as to systematically account for the different contributions of solvation and conformational entropy to the overall stability constant of each residue. As shown in Figure 2, three thermodynamic dimen-

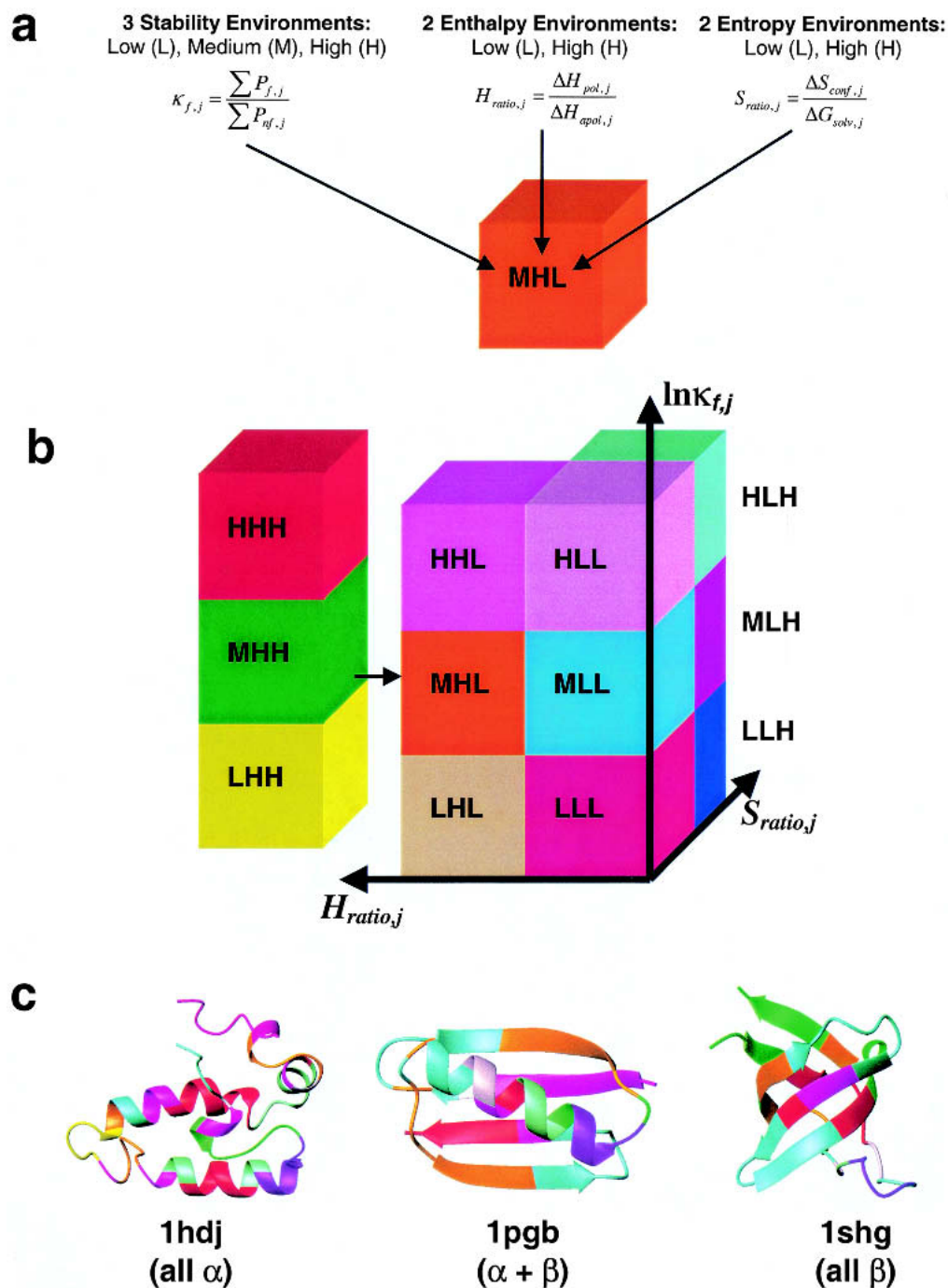


Fig. 2. Description of protein structure in terms of thermodynamic environments. (a) Thermodynamic environment classification scheme used in this work. Three quantities derived from the output of the COREX algorithm—stability ($\kappa_{f,j}$), enthalpy ratio ($H_{ratio,j}$), and entropy ratio ($S_{ratio,j}$)—describe the thermodynamic environment of each residue. Calculation of these quantities is described in Materials and Methods. (b) The 12 thermodynamic environments defined by this classification scheme are shown in a schematic describing protein energetic phase space. Each colored cube represents a region dominated by certain stability, enthalpy, and entropy characteristics. Every residue position in the protein structures used in this work lies somewhere within this phase space. (c) Examples of the distribution of thermodynamic environments of b in three proteins with varying types and amounts of secondary structure. Note that single secondary structure elements do not show unique thermodynamic environments.

sions are considered: stability ($\kappa_{f,j}$), enthalpy ($H_{ratio,j}$), and entropy ($S_{ratio,j}$). The first dimension uses the stability constant classification (Fig. 1) defined by Equation 3. As the particular value for the stability constant can arise from conformational entropy or solvent-related phenomena, a second dimension is used that provides the ratio of the conformational entropy to the total solvation free energy:

$$S_{ratio,j} = \frac{\Delta S_{conf,j}}{\Delta G_{solv,j}} \quad (15)$$

where $\Delta G_{solv,j}$ is the total residue-specific solvation component calculated similar to Equations 10 through 14. Fi-

nally, as the total solvation component can arise from polar or apolar contributions, a third dimension is incorporated that provides the ratio of polar to apolar enthalpy described by Equations 10 and 13:

$$H_{ratio,j} = \frac{\Delta H_{pol,j}}{\Delta H_{apol,j}} \quad (16)$$

The residues making up the 81 proteins analyzed in this study partitioned nonrandomly within the three-dimensional thermodynamic space. The nonrandom distribution of residues resulted in an empirical partitioning of the residue-specific data into 12 thermodynamic categories by dividing

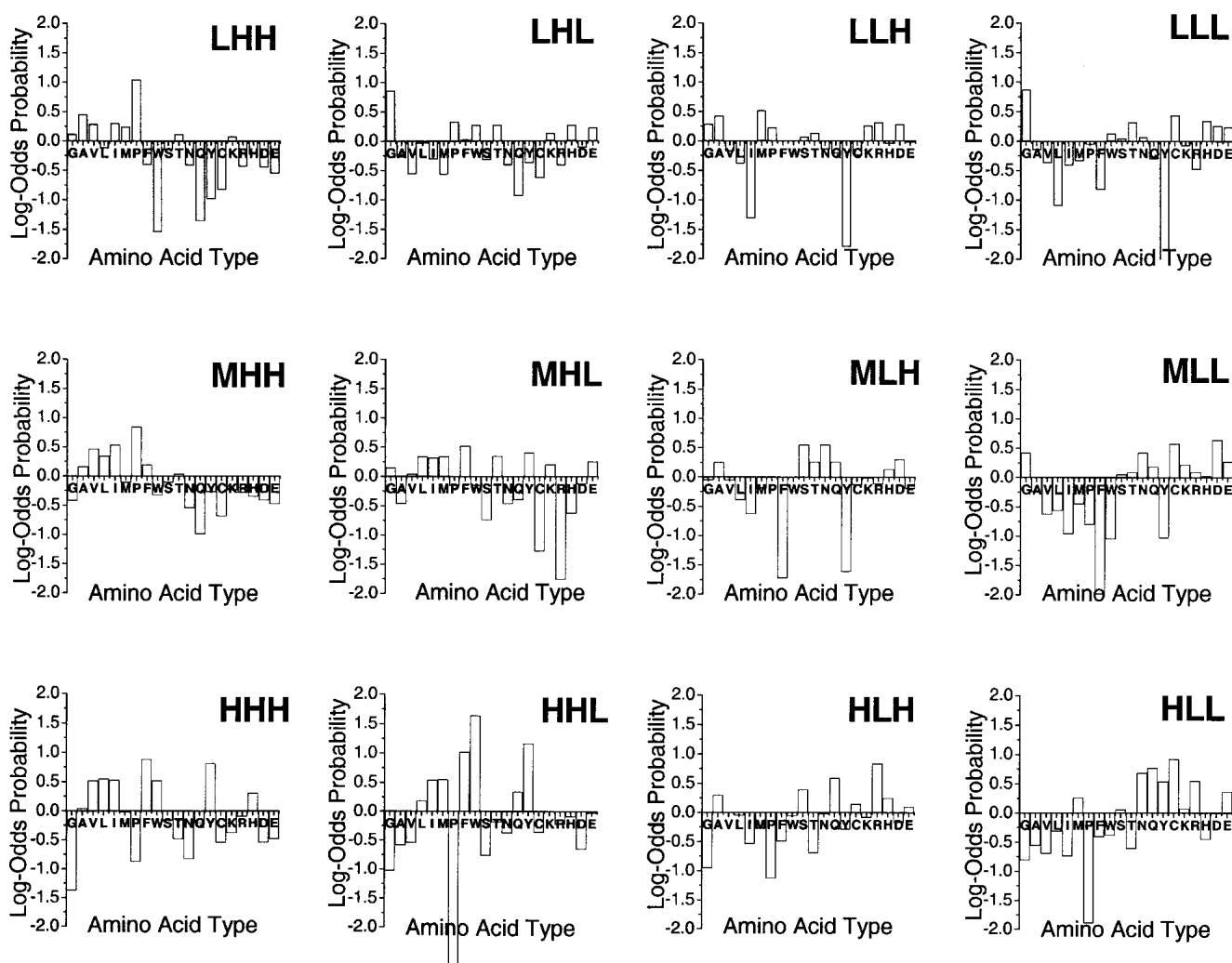


Fig. 3. Three-dimensional-to-one-dimensional scores relating amino acid types to 12 protein structural thermodynamic environments. The scores were calculated from normalized probabilities (log-odds ratios) of observing amino acid types in thermodynamic environments calculated from protein structures using the COREX algorithm, as described in the text. The 12 thermodynamic environments were classified empirically, as described in Materials and Methods. The three-letter abbreviation in each panel represents the stability, enthalpic, and entropic descriptor of the thermodynamic environment. For example, MLH represents a protein thermodynamic environment of medium stability, low polar/apolar enthalpy ratio, and high conformational entropy/Gibbs' solvation energy ratio.

Table 2. Statistics of amino acid type as a function of the 12 thermodynamic environments

	LHH	LHL	LLH	LLL	MHH	MHL	MLH	MLL	HHH	HHL	HLH	HLL	Total
ALA	48	19	26	30	74	24	55	44	53	15	56	22	466
ARG	12	10	14	13	29	4	23	37	28	13	58	40	281
ASN	11	9	8	20	20	13	40	46	12	10	22	41	252
ASP	14	16	16	32	30	24	41	75	21	10	25	27	331
CYS	5	5	5	20	12	4	14	37	11	7	18	36	174
GLN	4	5	7	13	12	13	28	34	21	19	38	42	236
GLU	17	30	16	42	38	47	34	70	30	25	44	53	446
GLY	32	55	21	77	39	41	38	79	12	9	15	16	434
HIS	6	8	4	12	11	5	12	14	17	6	13	6	114
ILE	27	12	3	15	70	34	15	14	56	30	16	12	304
LEU	28	25	12	12	92	55	30	33	91	33	41	29	481
LYS	36	31	24	35	55	51	46	76	38	29	42	45	508
MET	11	4	8	7	16	15	10	10	14	13	9	14	131
PHE	8	10	0	6	30	25	3	3	48	29	10	10	182
PRO	45	18	11	17	76	17	22	13	11	1	7	3	241
SER	19	13	13	26	41	13	53	42	31	9	44	29	333
THR	23	22	13	32	44	36	37	41	21	15	14	14	312
TRP	1	5	0	6	7	5	0	3	13	21	6	4	71
TYR	4	6	1	1	17	20	3	7	40	30	11	23	163
VAL	34	12	12	20	84	33	34	25	71	13	35	16	389
Total	385	315	214	436	797	479	538	703	639	337	524	482	5849

the stability data into three categories, the enthalpy data into two categories, and the entropy data into two categories (Fig. 2).

Amino acid propensities for thermodynamic environments

Each of the 5849 residues in the database was empirically binned into one of 12 thermodynamic environments, as described in Materials and Methods. For brevity, these environments are denoted by a shorthand notation; for example, the thermodynamic environment described in terms of low $\kappa_{r,j}$, low $H_{ratio,j}$, and high $S_{ratio,j}$ is given as "LLH." Statistics for amino acid type as a function of each of the thermodynamic environments were tabulated (Table 2), and the log-odds probability for an amino acid type to be in each thermodynamic environment was calculated. The resulting histograms (Fig. 3) revealed a nonrandom distribution of the amino acids within the thermodynamic environments. For example, hydrophobic residues such as Ile, Phe, and Val were observed with lower frequency in the MLL environment, whereas polar and charged amino acids such as Asp, Gln, and Lys were observed with higher frequency in this environment. These distributions could not always be rationalized on the basis of side-chain chemical properties, however, as the basic amino acids Arg and Lys showed very different propensities to occur in the MHL environment. This latter observation is a reflection of the fact that ensemble-derived energetics include averaged tertiary enthalpic and entropic information that is not encoded by individual side-chain properties alone.

Fold-recognition experiments based on amino acid propensities for thermodynamic environments

To validate that the thermodynamic environments provided an accurate description of the tertiary structure of a protein,

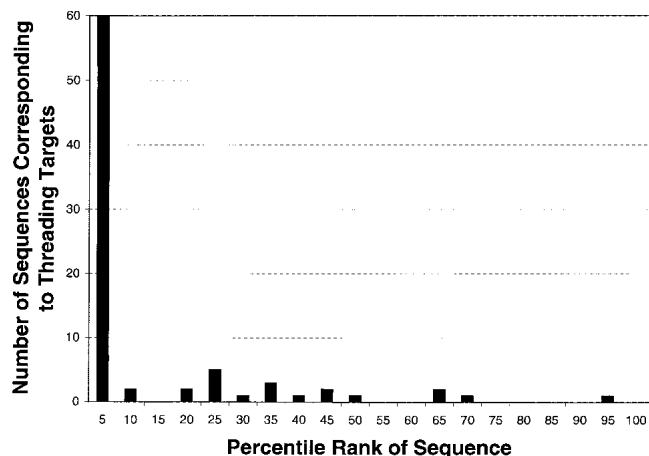


Fig. 4. Fold-recognition results for 81 protein targets using a scoring matrix composed of thermodynamic information from protein structures. The horizontal axis represents the percentile ranking of the score against the target structure for the sequence corresponding to the target structure. Low percentiles (high scores) indicate relatively more success in matching a sequence to its target structure. For example, the sequence corresponding to the target cold-shock protein (Protein Data Bank 1mjc) received the 157th highest score of 3858 sequences against the cold-shock protein thermodynamic profile. This result placed the sequence for the cold-shock protein in the fifth percentile bin in Fig. 4. When aligned with their respective thermodynamic profiles, the majority (44/81) of sequences scored better than 99% of the 3858 sequences in the database.

the three-dimensional profiling method of Eisenberg and coworkers (Gribskov et al. 1987; Bowie et al. 1991) was used. The general strategy of the three-dimensional profiling method is to describe the tertiary structure of a protein as a one-dimensional string of residue-specific “environmental classes” (Bowie et al. 1991). Based on the statistical preferences of each of the 20 amino acids for different environmental classes, a sequence alignment algorithm optimally aligns amino acid sequences to the environmental string. Instead of defining environmental classes in terms of static structure properties such as secondary structure or solvent accessibility, they have been defined here in terms of the COREX thermodynamic environments described above.

Simple fold-recognition experiments were performed based on amino acid distributions within the 12 thermodynamic environments. The 81 amino acid sequences coding for the native structures used in the database (in addition to 3777 decoy sequences) were each threaded against the 81 target thermodynamic environment profiles. The decoy sequences were obtained from the Protein Data Bank (Berman et al. 2000) and were inclusive for all sequences coding for experimentally solved structures that range from 35 to 100 residues in length. Nearly three fourths (60/81) of the correct sequences scored in the top fifth percentile when threaded against their corresponding thermodynamic environment profile (Fig. 4), and the Z-scores (the number of standard deviations a particular sequence scored above the mean score of all chains when length is taken into account) for these successful threadings ranged from 1.76 to 12.23 (Table 3). It was also observed that sequences belonging to proteins of the same SCOP fold class as a given target also scored well against the thermodynamic profile of the target. Approximately 30% of the 2041 sequences belonging to the same SCOP fold class as one of the 60 targets which scored in the top fifth percentile also scored in the top fifth percentile (data not shown).

Thermodynamic information is more fundamental than secondary structure information

Secondary structure, although useful in the analysis and classification of protein folds, is an easily reported observable that is of questionable utility in explaining the underlying physical chemistry of protein structure. In fact, secondary structure can be viewed as a manifestation of the backbone/side-chain van der Waals' repulsions that divide ϕ/ψ space, modified by the thermodynamic stability afforded by local and tertiary interactions such as hydrogen bonding and the hydrophobic effect (Baldwin and Rose 1999; Srinivasan and Rose 1999). Any reasonable description of the energetics of protein structure must be able to reflect these realities independent of secondary structural propensities of amino acids and the secondary structural classifications of folds.

Although the COREX energy function accounts for specific interactions only in an implicit way, the results of a COREX calculation may provide deeper insight than secondary structure into the structural determinants of protein folds. For example, Figure 2c compares the thermodynamic environment profiles for an all- α protein and an all- β protein threaded over their native folds. Visual inspection of the two color-coded structures reveals that different thermodynamic environments span single types of secondary structure, and that the same thermodynamic environment is found in different types of secondary structural elements.

To investigate the possibility that the thermodynamic environments calculated by COREX represents a more fundamental descriptor of proteins that transcends structural classifications, the threading procedure was repeated on a subset of proteins from the original database, sorted by secondary structure. First, a scoring table was assembled from the 31 proteins in Table 1 that are classified by the SCOP database as being “all- α ” proteins. Second, the 12 “all- β ” proteins from Table 1 were threaded using the scoring table derived solely from the “all- α ” proteins. In other words, amino acid propensities for the thermodynamic environments from only all- α proteins were used to perform fold-recognition experiments on only all- β proteins. For >80% of the targets (10/12), sequences known to adopt the native all- β structures scored in the top 5% of the 3858 decoy sequences, (Fig. 5). This result is a clear demonstration that the ener-

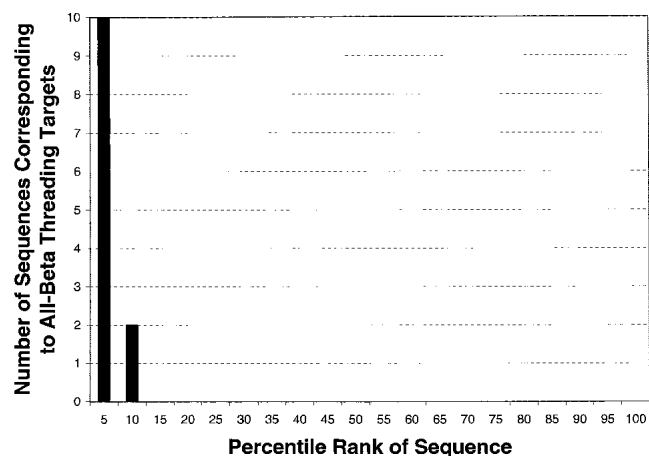


Fig. 5. Fold-recognition results for 12 all- β protein targets using a scoring matrix composed of thermodynamic information from 31 all- α protein structures. The horizontal axis represents the percentile ranking of the score against the target structure for the sequence corresponding to the target structure. Low percentiles (high scores) indicate relatively more success in matching a sequence to its target structure. For example, the sequence corresponding to the all- β target tendamistat (Protein Data Bank 1hoe) received the 26th highest score of 3858 sequences against the tendamistat thermodynamic profile. This result placed the tendamistat sequence in the fifth percentile bin in Fig. 5. All 12 sequences corresponding to β -targets scored better against their respective targets than 90% of the 3858 sequences in the database.

Table 3. Fold-recognition results using a scoring function derived solely from COREX thermodynamic information

Number	Protein Data Bank ID No.	Percentile Rank	Z-Score
1.	1a1iA	0.29	3.49
2.	1a6s	0.67	3.23
3.	1a8o	0.34	3.29
4.	1aa3	3.84	2.08
5.	1aba	0.03	4.10
6.	1adr	0.93	3.71
7.	1aiw	2.36	2.27
8.	1an4A	23.64	0.68
9.	1aoiB	26.31	0.52
10.	1avyC	5.16	1.82
11.	1b9gA	0.18	4.48
12.	1bdd	0.44	5.07
13.	1bdo	0.05	6.25
14.	1bf4A	0.16	4.04
15.	1bg8A	33.23	0.32
16.	1bo9A	0.21	4.06
17.	1c1yB	95.44	-1.46
18.	1cc5	0.13	5.30
19.	1chc	67.88	-0.55
20.	1ctf	32.17	0.22
21.	1cyo	5.47	1.76
22.	1d3bB	0.93	2.70
23.	1doqA	0.03	4.34
24.	1dt4A	0.08	6.83
25.	1egwA	4.33	2.14
26.	1eo0A	0.88	4.01
27.	1fgp	2.13	2.65
28.	1gdc	64.41	-0.45
29.	1hcrA	0.16	4.70
30.	1hdj	1.35	2.80
31.	1hoe	0.13	5.62
32.	1hp8	0.47	4.43
33.	1iieA	0.39	3.28
34.	1iro	0.13	5.40
35.	1isuA	0.54	3.58
36.	1kdxA	0.03	9.34
37.	1kjs	32.40	0.26
38.	1kveA	2.41	2.50
39.	1kwaA	0.29	3.70
40.	1mho	0.39	3.54
41.	1mjc	4.07	1.99
42.	1mknA	3.24	2.33
43.	1mof	65.34	-0.47
44.	1mwpA	24.29	0.56
45.	1nhm	17.26	0.93
46.	1nkl	0.91	3.19
47.	1npsA	0.13	4.36
48.	1nre	24.29	0.54
49.	1ntcA	39.71	0.10
50.	1nxb	0.78	4.10
51.	1opd	4.15	2.09
52.	1otfA	1.09	3.49
53.	1pcfA	40.95	0.17
54.	1pgb	0.13	5.90
55.	1plc	0.13	8.42
56.	1ptf	7.34	1.63
57.	1ptq	9.62	1.33
58.	1ptx	0.47	4.21

Table 3. Continued

Number	Protein Data Bank ID No.	Percentile Rank	Z-Score
59.	1qa4A	45.59	-0.05
60.	1qgwB	2.95	2.25
61.	1qqvA	1.87	2.73
62.	1r1bA	22.76	0.68
63.	1rop	42.48	0.02
64.	1rzi	0.05	6.57
65.	1shg	0.08	6.09
66.	1sknP	0.03	6.28
67.	1svfB	20.14	0.67
68.	1tbaA	1.09	2.68
69.	1tgsI	2.62	2.60
70.	1trlA	23.54	0.53
71.	1ugiD	0.44	7.02
72.	1utg	0.08	5.92
73.	1vcc	0.08	4.48
74.	2abd	0.23	3.96
75.	2bopA	0.03	7.09
76.	2ci2I	5.44	2.06
77.	2knt	0.08	12.23
78.	2spgA	0.39	5.31
79.	3eipA	0.18	5.53
80.	3ncmA	0.44	4.24
81.	5hpgA	0.05	11.02

getic information derived from the COREX calculations is independent of protein secondary structure.

Discussion

A protein structure characterized as a single molecule composed of several rigid secondary structure units is an incomplete picture. Perhaps the most important implication of this work is that much energetic information, essential to a complete understanding of the folding, stability, and function of a protein, is lost when a crystal structure is viewed in a static context. The observation that a COREX thermodynamic profile, without specific sequence or structural information, can be successful in matching an amino acid sequence with its native fold, indicates that thermodynamic environments can provide an approximation of the missing energetic properties of an experimental ensemble.

In assessing the strengths and weaknesses of the approach described here, we note that although not as successful as the original profile search method at matching sequence to structure (Bowie et al. 1991), our thermodynamic descriptors do not encode sequence-specific information. Consequently, sequences of the same SCOP fold class also scored high in the absence of sequence identity, as noted above. Inspection of the probability scores for the different amino acids within each environment (Fig. 3) reveals the origin of this phenomenon. Different amino acids contribute to the

positive and negative scores in each environment, with different amino acids having the same propensity in one environment while having opposite propensities in others (see Tyr and Phe). The result is that the interchangeability of amino acids is environment specific. This observation is intriguing as it may provide “thermodynamic signatures” for specific folds, or even more compelling, it may provide signatures that are shared between proteins with little or no structural similarity.

It is interesting to note that of the 21 target sequences that scored below the fifth percentile, the relative contributions of each amino acid to the scoring was nearly identical to the sequences that scored in the top fifth percentile. Furthermore, no fold- or class-specific trend in the scoring was observed. The only observed difference was in the number of amino acids that contribute negatively. In other words, high and low scoring sequences had similar positive contributions, but low scoring targets had more negative scores. These results indicate that there is not a fundamental flaw in the environmental descriptors, which fails for a class of proteins. Instead, the limitation is likely to be in the empirical nature of the environmental boundaries. It is often seen that in one environment, an amino acid will have little or no propensity, whereas in a neighboring environment, there is a large negative propensity (e.g., Pro in MHL versus HHL). Thus, the score for an amino acid can depend on the precise boundary, if that environment score is close to the boundary value. Indeed, the empirical nature of the boundaries being the primary source of the low scoring sequences is supported by the observation that changes in the precise boundaries of the thermodynamic environments do not significantly affect the number of sequences that score in the top fifth percentile. What is changed, however, are the target sequences, which, when matched with the correct structures, score in the top fifth percentile.

Two additional results pertain to secondary structure, highlighting and extending the notion that secondary structural elements cannot be obligatorily regarded as cooperative folding units (Wrabl et al. 2001). The observations that the same thermodynamic environment may occur in α -helices as well as β -strands, and that different thermodynamic environments occur within the same secondary structural element, show that different regions within a given secondary structural element can be stabilized through different energetic mechanisms. The ability of thermodynamic information from the all- α database to successfully thread all- β proteins indicates that thermodynamic environments represent a more fundamental property of fold specificity, a property that transcends secondary structural classifications. In this regard, we note that the threading scheme described in this work is significantly more successful at matching sequences with their native structures than is a simple threading scheme based only on secondary structure propensities (Wrabl et al. 2001). An additional point of note is that only

6% of the high scoring decoy sequences for each target fold were in the same SCOP structural class (i.e., all- α , all- β , $\alpha + \beta$, etc.) as the target fold. This highlights the notion that secondary structure (or more specifically, a property of secondary structure) is not the cause of the fold-recognition success. Instead, the rules are thermodynamic in nature, and identical for helix and strand.

Although the fold-recognition results presented above are used only as a “proof-of-principle” for the concept of thermodynamic environments, it is probable that ensemble-based structural energetics represents a new source of data for fold recognition and structure prediction that is orthogonal and complementary to the traditional sources of information from amino acid substitution matrices and secondary structure propensities. Energetic information is generally unexploited by prediction methods that conveniently portray protein targets as static collections of atoms, with all portions of a target structure being equally stable from the point of view of a fold-recognition scoring function. Ignorance of the experimental reality of time-averaged fluctuations around some minimum-energy conformation could be a major shortcoming of current fold-recognition methods. The results of this work could have practical application to fold-recognition algorithms in two ways. First, the COREX-derived thermodynamic signatures of target structures could be used to define regions that should be more heavily or less heavily weighted by a scoring function. Second, propensities of amino acid types for thermodynamic environments could be the basis of a scoring function relating sequence to structure, as explored above. These concepts remain to be tested through blind structure prediction and experimental protein design.

Finally, although the approach described here stresses the concept of thermodynamic rather than structural determinants for a specific fold, the significance and the success of this approach stems from the ensemble-based nature of the energetics. As emphasized above, the residue-specific thermodynamic descriptors do not represent the contribution of each amino acid to the stability of the protein. Instead, they report on the local energetics of each region, implicitly considering contributions of all other amino acids in the context of that structure. The approach described here provides a thermodynamically rigorous means of relating residue-specific quantities to the overall energetics of the system, wherein the physical determinants of each residue are weighted according to their energetic impact on the ensemble.

Materials and methods

Selection of proteins used in the data set

A database of 81 proteins, 5849 residues total (Table 1), was selected from the Protein Data Bank (Berman et al. 2000) on the basis of biological and computational criteria, as described previ-

ously (Wrabl et al. 2001). One criterion of note was that the proteins were nonhomologous with every other member of the set, as ascertained by SCOP (Murzin et al. 1995).

COREX algorithm and surface area calculations

The COREX algorithm has been described in detail (Hilser and Freire 1996). The calculations in this work were run at a simulated temperature of 25°C and a window size of five residues. The calorimetric enthalpy and entropy of solvation were parameterized from polar and apolar surface exposure (Hilser and Freire, 1996). The Boltzmann weight of each microstate (i.e., $K_i = e^{(-\Delta G_i/RT)}$) was used to calculate its probability. COREX uses empirical parameterizations to calculate the relative apolar and polar free energies of each microstate (Xie and Freire 1994; Gomez et al. 1995; D'Aquino et al. 1996; Hilser and Freire 1996):

$$\Delta G_{\text{apolar},i}(T) = 8.44 * \Delta \text{ASA}_{\text{apolar},i} + 0.45 * \Delta \text{ASA}_{\text{apolar},i} * (T - 333) - T * (0.45 * \Delta \text{ASA}_{\text{apolar},i} * \ln(T/385)) \quad (17)$$

$$\Delta G_{\text{polar},i}(T) = 31.44 * \Delta \text{ASA}_{\text{polar},i} - 0.26 * \Delta \text{ASA}_{\text{polar},i} * (T - 333) - T * (-0.26 * \Delta \text{ASA}_{\text{polar},i} * \ln(T/335)) \quad (18)$$

The three primary components used to calculate conformational entropies ($\Delta S_{i,\text{conf}}$) for each microstate are as follows: (1) $\Delta S_{\text{bu} \rightarrow \text{ex}}$, the entropy change associated with the transfer of a side-chain that is buried in the interior of the protein to its surface; (2) $\Delta S_{\text{ex} \rightarrow \text{u}}$, the entropy change gained by a surface-exposed side-chain when the peptide backbone unfolds; and (3) ΔS_{bb} , the entropy change gained by the backbone itself on unfolding (Hilser and Freire 1996). For fold-recognition calculations, the total ($\Delta S_{i,\text{conf}}$) of all proteins was multiplied by a scaling factor to eliminate the contribution of the completely unfolded microstate to the residue-specific thermodynamic parameters (Wrabl et al. 2001).

Binning of thermodynamic environments

Each of the 5849 residues in the database was binned into one of the 12 thermodynamic environment classes based on their stability ($\kappa_{f,j}$), enthalpy ($H_{\text{ratio},j}$), and entropy ($S_{\text{ratio},j}$) values (Equations 3, 16, and 15, respectively). These thermodynamic environments are denoted by the following abbreviations: LLL, LLH, LHL, LHH, MLL, MLH, MHL, MHH, HLL, HLH, HHL, and HHH. For example, residues in the LLH thermodynamic environment would have been binned into the low (L) stability ($\kappa_{f,j}$) class, the low (L) enthalpy ($H_{\text{ratio},j}$) class, and the high (H) entropy ($S_{\text{ratio},j}$) class. The cutoffs for each thermodynamic class were defined as

Stability ($\kappa_{f,j}$) class (L, M, or H):

$$\text{Low } \kappa_{f,j} \text{ (L)} \equiv [\ln \kappa_{f,j} < 7.95] \quad (19)$$

$$\text{Medium } \kappa_{f,j} \text{ (M)} \equiv [7.95 \leq \ln \kappa_{f,j} < 13.40] \quad (20)$$

$$\text{High } \kappa_{f,j} \text{ (H)} \equiv [13.40 \leq \ln \kappa_{f,j}] \quad (21)$$

Enthalpy ($H_{\text{ratio},j}$) class (L or H):

$$\text{Low } H_{\text{ratio},j} \text{ (L)} \equiv [-\Delta H_{\text{pol}} < -1.024 * \Delta H_{\text{ap}} - 2553 \text{ cal/mol}] \quad (22)$$

$$\text{High } H_{\text{ratio},j} \text{ (H)} \equiv [-\Delta H_{\text{pol}} \geq -1.024 * \Delta H_{\text{ap}} - 2553 \text{ cal/mol}] \quad (23)$$

Entropy ($S_{\text{ratio},j}$) class (L or H):

$$\text{Low } S_{\text{ratio},j} \text{ (L)} \equiv [-T\Delta S_{\text{conf}} < 0.125 * \Delta G_{\text{solv}} - 3053 \text{ cal/mol}] \quad (24)$$

$$\text{High } S_{\text{ratio},j} \text{ (H)} \equiv [-T\Delta S_{\text{conf}} \geq 0.125 * \Delta G_{\text{solv}} - 3053 \text{ cal/mol}] \quad (25)$$

Visual inspection of the segregation of amino acid types as a function of various thermodynamic parameters extracted from the 81-protein COREX database, guided by the development outlined in Equations 5 through 16, indicated that the general classifications of stability, enthalpy, and entropy would reasonably divide thermodynamic space (as indicated in Fig. 2). The exact cutoffs for the 12 residue-specific thermodynamic environments used in the threading calculations were determined automatically by an exhaustive grid search of all possible cutoffs (data not shown). The utility of each trial set of cutoffs was initially determined from a coarse search of cutoff space by threading a constant subset of eight targets in the protein database and recording sets of cutoffs that maximized the Z-scores and percentiles for each target. Then, a finer grid search over the best sets of cutoffs, threading against a subset of 20 targets for each trial set of cutoffs, resulted in the optimized set of cutoffs used for the threading experiments shown in this work. Identical cutoffs were used for the α/β -threading calculations; namely, no special optimization was performed for the scoring of the α/β -experiment.

Fold-recognition details

The profiling method of Eisenberg and coworkers (Gribskov et al. 1987; Bowie et al. 1991) was used to create thermodynamic environment profiles for each of the 81 proteins in the database. A thermodynamic profile is a one-dimensional string of thermodynamic environments (i.e., LLH) as a function of residue position. Next, a three-dimensional-to-one-dimensional scoring matrix for each protein in the database was calculated, in which the scoring matrix data was simply the log-odds probabilities of finding amino acid types in one of the thermodynamic environment classes (Equation 27; below). The resulting profile of the target protein was then optimally aligned to each member of a library of amino acid sequences (i.e., 3858 decoy sequences) by maximizing the score between the sequence and the profile using a local alignment algorithm based on the Smith-Waterman algorithm (Smith and Waterman 1981) as implemented in PROFILESEARCH (Bowie et al. 1991). No attempt was made to optimize the gap opening and extension penalties for the local algorithm; in all cases, these were the default values given in the PROFILESEARCH package, 5.00 and 0.05, respectively. Z-scores were taken from the PROFILESEARCH output for each threading result. PROFILESEARCH computes a Z-score using Equation (26):

$$\text{Z-score} = (s - \langle S \rangle) / \sigma \quad (26)$$

In Equation 26, s is the PROFILESEARCH threading score of sequence i (adjusted for the length of sequence i according to routines internal to the PROFILESEARCH package) when threaded against the target structure; $\langle S \rangle$ is the average length-adjusted threading score of all sequences in the database threaded against the target structure; and σ is the standard deviation of the length-adjusted scores of all sequences in the database threaded against the target structure. Thus, the Z-score is the number of standard deviations above (or below, if the Z-score is negative) the mean that sequence i scored against its target. The 3858-member sequence library consisted of all protein chains of length 35 to 100 residues (inclusive) corresponding to structures found in the Protein Data Bank (Berman et al. 2000).

Construction of scoring matrices

The scoring matrices were calculated as log-odds probabilities of finding residue type j in structural environment k , as described

below (Bowie et al. 1991; Wrabl et al. 2001). The matrix score, $S_{j,k}$, was defined as

$$S_{j,k} = \ln \frac{P_{j|k}}{P_k} \quad (27)$$

$P_{j|k}$ is the probability of finding a residue of type j in stability class k (i.e., number of counts of residue type j in stability class k divided by the total number of counts of residue type j), and P_k is the probability of finding any residue in the database in stability environment k (i.e., number of residues in stability class k , regardless of amino acid type, divided by the total number of residues in the entire database, regardless of amino acid type). The structural environment used was one of the 12 COREX thermodynamic environments (LHH, LHL, LLH, LLL, MHH, MHL, MLH, MLL, HHH, HHL, HLH, and HLL), as described above. The fold-recognition target was removed from the database, and the remaining 80 proteins were used to calculate the probabilities. Therefore, information about the target was never included in the scoring matrix.

Electronic supplemental material

One table (tab-delimited text) listing the COREX-calculated thermodynamic parameters for the 5849 residues of the 81 proteins of Table 1: $\ln\kappa_{f,j}$, ΔH_{ap} , ΔH_{pol} , ΔS_{ap} , ΔS_{pol} , and ΔS_{conf} , and the latter three quantities multiplied by the simulated temperature (298.15 K).

Acknowledgments

Supported by National Science Foundation grant MCB-9875689, National Institutes of Health grant RO1-GM13747, Welch Award H-1461, and GSE Systems.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181**: 223–230.
- Baldwin, R.L. 1986. Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci.* **83**: 8069–8072.
- Baldwin, R.L. and Rose, G.D. 1999. Is protein folding hierarchic? I: Local structure and peptide folding. *Trends Biochem. Sci.* **24**: 26–33.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242.
- Bonneau, R., Tsai, J., Ruczinski, I., and Baker, D. 2001. Functional Inferences from blind ab initio protein structure predictions. *J. Struct. Biol.* **134**: 186–190.
- Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.
- D'Aquino, J.A., Gomez, J., Hilser, V.J., Lee, K.H., Amzel, L.M., and Freire, E. 1996. The magnitude of the backbone conformational entropy change in protein folding. *Proteins* **25**: 143–156.
- Gomez, J., Hilser, V.J., Xie, D., and Freire, E. 1995. The heat capacity of proteins. *Proteins* **22**: 404–412.
- Gribkov, M., McLachlan, A.D., and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci.* **84**: 4355–4358.
- Habermann, S.M. and Murphy, K.P. 1996. Energetics of hydrogen bonding in proteins: A model compound study. *Protein Sci.* **5**: 1229–1239.
- Hilser, V.J. and Freire, E. 1996. Structure-based calculation of the equilibrium folding pathway of proteins: Correlation with hydrogen exchange protection factors. *J. Mol. Biol.* **262**: 756–772.
- Hilser, V.J., Dowdy, D., Oas, T.G., and Freire, E. 1998. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc. Natl. Acad. Sci.* **95**: 9903–9908.
- Jones, D.T., Tress, M., Bryson, K., and Hadley, C. 1999. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins* **37**: 104–111.
- Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
- Koonin, E.V., Wolf, Y.I., and Aravind, L. 2000. Protein fold recognition using sequence profiles and its application in structural genomics. *Adv. Prot. Chem.* **54**: 245–275.
- Lee, K.H., Xie, D., Freire, E., and Amzel, L.M. 1994. Estimation of changes in side chain configurational entropy in binding and folding: General methods and application to helix formation. *Proteins* **20**: 68–84.
- Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- Panchenko, A.R., Marchler-Bauer, A., and Bryant, S.H. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296**: 1319–1331.
- Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**: 195–197.
- Srinivasan, R. and Rose, G.D. 1999. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci.* **96**: 14258–14263.
- Wrabl, J.O., Larson, S.A., and Hilser, V.J. 2001. Thermodynamic propensities of amino acids in the native state ensemble: Implications for fold recognition. *Protein Sci.* **10**: 1032–1045.
- Xie, D. and Freire, E. 1994. Structure-based prediction of protein folding intermediates. *J. Mol. Biol.* **242**: 62–80.