Structure
Article

Predicting the Energetics of Conformational Fluctuations in Proteins from Sequence: A Strategy for Profiling the Proteome

Jenny Gu[1,2] and Vincent J. Hilser[1,2,*]
[1]Department of Biochemistry and Molecular Biology
[2]Sealy Center for Structural Biology and Molecular Biophysics
University of Texas Medical Branch, Galveston, TX, 77555-1068, USA
*Correspondence: vjhilser@utmb.edu
DOI 10.1016/j.str.2008.08.016

SUMMARY

The abundance of dynamic and disordered regions in proteins suggests that structural determinants alone may not be sufficient to describe function. Instead, descriptors that account for the dynamic features of the energy landscape populated by the protein ensemble may be required. Here, we show that the thermodynamics of the dynamical complexity that imparts biological function can be largely reconstructed using sequence information alone, allowing thermodynamic characterization of entire proteomes without the need for structural analysis. We show that this tool can be used to analyze conserved energetic signatures within classes of proteins, as well as to compare the thermodynamic character of different proteomes.

INTRODUCTION

The paradigm that a stable structure is a prerequisite to protein function has been challenged by a growing body of evidence for the importance of flexible and intrinsically disordered (ID) regions in proteins (Xie et al., 2007a, 2007b). The functional roles for disorder in proteins can be grouped into at least four classes: (1) molecular recognition, (2) molecular assembly, (3) protein modification, and (4) entropic chain activities (Dunker et al., 2002; Radivojac et al., 2007), with a new role as an allosteric regulator having recently been described (Hilser and Thompson, 2007; Yi et al., 2007). It is critical to note that formation of structure is no longer considered to be a prerequisite for function (Uversky, 2002). In fact, dynamical properties can be subjected to varying selection pressures, as observed in intrinsically unstructured linker domains of a 70 kDa subunit or replication protein A (Daughdrill et al., 2007).

Of particular importance is the observation that ID is found in disproportionately higher amounts in transcription factors (TFs) (Liu et al., 2006). Because TFs are usually multidomain proteins whose functions are to integrate complex arrays of binding events in different functional domains and to translate those events into a transcription initiation signals (Liu et al., 2006; McEwan et al., 2007; Ward et al., 2004), the hyper-abundance of ID in TFs would suggest an important role in signaling. Indeed, the apparent conservation of ID to particular regions of the TF sequence suggests a common purpose, although this purpose is not yet known.

Recently, it was shown that coupling the folding of ID regions to the binding of ligands provides a general mechanism for optimizing allosteric communication (Hilser and Thompson, 2007). The findings suggest that the role of ID may be to facilitate signaling. A more general result of that study, however, was the observation that the magnitude of the coupling between different sites in a protein is a function of where the relevant conformational equilibria are poised prior to the signaling event. In effect, signaling is coupled to the native-to-denatured state energetic balances in proteins. The importance of this finding is that the regional differences in stability within a sequence may contain important functional clues and that methods designed to estimate the stability profile of a protein from sequence could be invaluable in efforts to annotate function within entire proteomes.

In this paper, we use an ensemble-based thermodynamic description of proteins to establish a relationship between the energy landscapes for a set of proteins and their respective sequences. We find that the reconstruction of the energetic landscape populated by protein ensembles can be achieved using constraints inherently encoded in the protein sequence. The success of this reconstruction suggests that variations of position-specific thermodynamics within a protein, although determined by both local and distal interactions, are nonetheless encoded at the local sequence level. This allows position-specific stability and dynamic information to be estimated for entire proteomes, thus providing a vehicle for evaluating properties of proteins where structural information is currently unavailable or, as in the case of ID segments, where direct observation may not possible.

RESULTS AND DISCUSSION

Sequence-Based Reconstruction of the Energetic Landscape: The eScape Algorithm

Our goal is to develop a sequence-based algorithm that reconstructs the protein stability profile for a protein sequence using thermodynamic information obtained from known structures. Our strategy to reconstruct protein stability profiles is to first perform COREX calculations (Hilser and Freire, 1996; Hilser et al., 2006) on a data set of nonredundant human proteins between

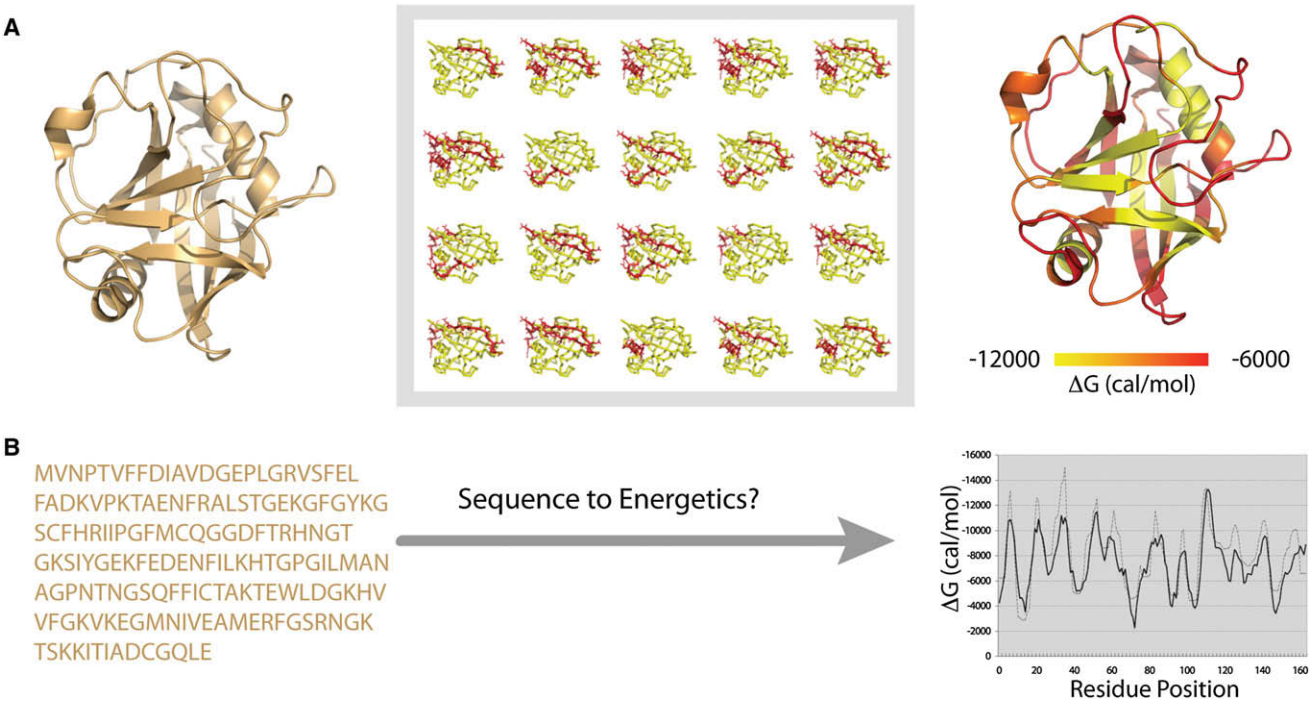

Figure 1. Schematic Representation of the COREX Method for Generating Ensembles and Basis for eScape

(A) COREX calculates position-specific stability by using a conformational ensemble of locally unfolded regions generated from a high-resolution structure. The ensemble of states for each protein (box) is obtained by systematically folding and unfolding all possible small residue segments in the protein. The probability of each state in the ensemble is calculated by using a surface area–based energy function that subsequently allows for the calculation of the Gibbs free energy of stability (ΔG). The apolar (ΔH_{ap}) and polar (ΔH_p) enthalpic as well as entropic (TΔS) contributions can also be calculated. COREX provides the foundation in which (B) eScape extrapolates a relationship between protein sequence and local conformational fluctuation in the native state. The prediction of eScape for a human cyclophilin A (PDB ID: 2CPL) is plotted (thick line) and compared to stability values calculated by COREX (thin dash line).

50 and 250 residues in length. Briefly, COREX generates an ensemble of states for each protein by systematically folding and unfolding small 5-residue segments of the protein structure in all possible combinations. Use of a surface area–based energy function allows for the determination of the probability of each state. Shown in Figure 1A is a schematic representation of COREX-derived ensemble, as well as the corresponding position-specific energetics. Quintessential to the COREX approach is the determination of the position-specific stability within each protein. Defined as the position-specific free energy, this parameter describes the energy difference between the subensembles in which a particular residue *j* is folded and unfolded, as revealed by the following expression:

$$[\Delta G]_j = \langle \Delta G_{f,j} \rangle - \langle \Delta G_{nf,j} \rangle = -RT \ln\left(\frac{\sum P_{f,j}}{\sum P_{nf,j}}\right). \quad (1)$$

The fact that Equation 1 represents an ensemble-average property that is reported at each position means that the value obtained for each residue cannot be interpreted as either (1) a property of the amino acid at that position, or (2) the contribution of that residue to the overall stability of the protein. This important point was demonstrated previously by showing that the ensemble average energetic values reported at each position (Larson and Hilser, 2004) correlate with neither the properties of the amino acids at those positions nor static structural properties, such as accessible surface area at each position. In short, the position-specific free energy predictions obtained from COREX (Figure 1) are not simply recapitulating local structural parameters.

The most important aspect of Equation 1, however, is that under the appropriate exchange conditions, the computed values for each position can be compared to protection factors obtained from hydrogen-deuterium exchange experiments (Hilser and Freire, 1996). The agreement between the calculated values from Equation 1 and the free energies obtained from the experimentally measured protection factors from hydrogen exchange experiments ($\Delta G_{HX} = -RT \ln PF$, where PF is the protection factor) suggests that the COREX algorithm provides a reasonable model of the energy landscape of a protein (Hilser and Freire, 1996).

Because the goal of this work is to be able to reconstruct the energetic profile from sequence information alone (Figure 1B), a COREX analysis was performed on a database of proteins to sample the propensities of different amino acids to be in environments of different stability. From this database of COREX-derived, position-specific energetics, the information can be subdivided to account for the influence of neighboring residues on the stability of a particular position. Examination of the library of tripeptides reveals that the stability of a position with a particular amino acid depends on the identity of neighboring residues and not only the identity of the amino acid itself (Figure 2). For example, tripeptides containing alanine, arginine, and proline

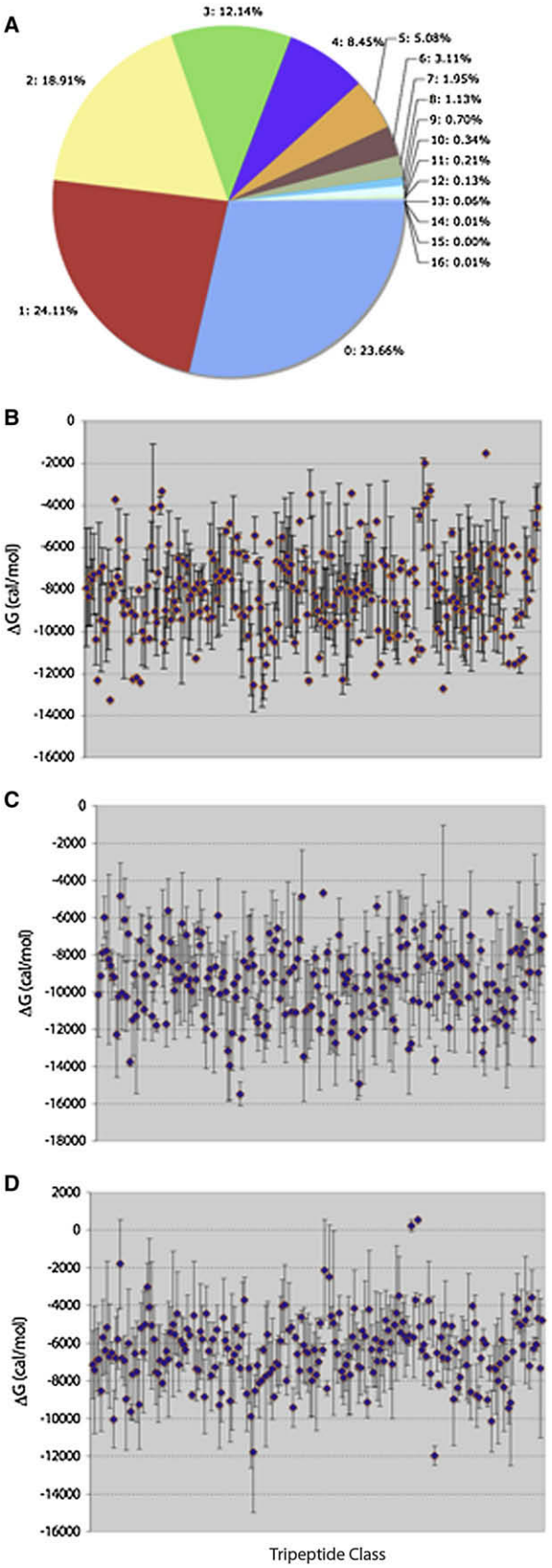

Figure 2. Parsing the Energetic Landscape with Higher Order Patterns

The energetic landscape for a nonredundant set of human proteins calculated using COREX is inspected for relationships between higher sequence order and position-specific thermodynamic descriptors. This inspection is important in the design of the eScape algorithm.

(A) The frequencies of observed tripeptides show that 23.66% of the possible 8800 tripeptides (including terminal end spacers) are not represented with this data set with some tripeptides having a higher sampling frequency than others.

(B–D) Examples of using higher order sequence information (tripeptides) to define initial energetic limits that will subsequently be used as input features for the trained linear regression models are shown for different tripeptides with (B) alanine, (C) arginine, and (D) proline as the central residue. Each point represents the mean ΔG observed for the different tripeptide combination containing the indicated central residue. Error bars represent one standard deviation from the mean energetic stability observed for the corresponding tripeptide pattern in the database.

as the central residue were extracted from the data set to demonstrate that the mean stability is found to vary between each tripeptide and amino acid type. Interestingly, although the energetic limits for some tripeptide sequences appear to span the entire range of data, this is not true for the majority of the tripeptides. Indeed, the subtle differences evident in the energetic mean and range will prove to be critical to the successful predictions. From this tripeptide partitioning, a library of energetic limits (i.e., the observed maximum and minimum values associated with each sequence pattern) can be obtained for the database of proteins. It should be noted that the information itself is insufficient to directly reconstruct the energetic landscape and requires the aid of an additional machine learning technique.

To reconstruct the energetic landscape, we developed eScape, shown schematically in Figure 3, which is a multistep algorithm that uses a trained linear regression model in the final step to make predictions. Reconstruction is achieved by first defining the energetic boundaries for each residue type in the protein sequence. Preliminary energetic boundaries are assigned from the library of COREX position-specific stability values calculated from the nonredundant set of proteins. To do this, the energetic ranges for the tripeptide that is centered on the residue in question (as found in the library) are used. The boundaries (i.e., the minimum and maximum of the range) of the energetics are then averaged across the sequence, with a sliding window of 5 residues. This step provides an estimate of the thermodynamic boundaries for that sequence. It is important to note that because the energetic ranges at each position were derived from ensemble-averaged properties, both global and local contributions to the residue stability are implicitly considered. The final prediction of residue stability is made using the estimated thermodynamic boundaries as input features to a trained linear model conducted with 10-fold cross-validation as described in Experimental Procedures.

Results of eScape Model

The average performance values and the results from each 10-fold cross-validation are reported for the database of proteins (Table 1). The energetic profiles determined for the native ensembles (Larson and Hilser, 2004; Wrabl et al., 2001, 2002) can be reconstructed with eScape, resulting in an adjusted R^2

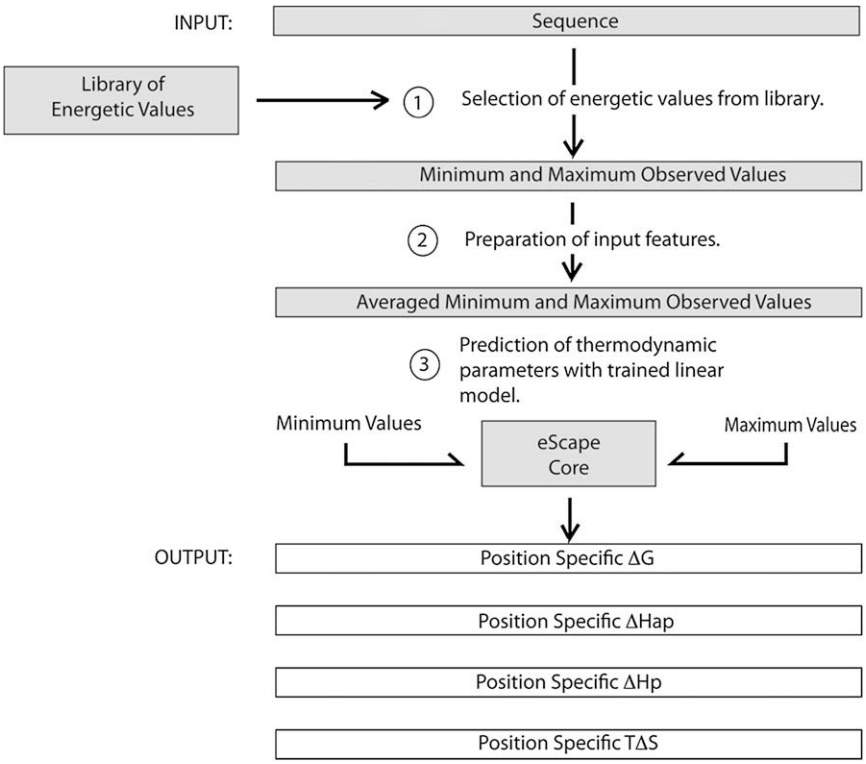

Figure 3. Scheme of the eScape Algorithm
A procedural scheme illustrating the underlying infrastructure of eScape. First, corresponding tripeptide energetic boundaries are drawn from the library of COREX calculated thermodynamic descriptors for each position in the sequence. The limits of both extremes are then averaged across the sequence with a sliding window of 5 residues. The smoothed boundary is used as input features for a trained linear model to give a final estimate of the contributing thermodynamic values that describe the energetic contributions to ensemble stability at the specific position.

value of 0.70 and a Pearson's correlation coefficient of 83.63%. Although eScape predicts residue energetic values at ~80% accuracy with an error margin of ± 2 kcal/mol (Figure 4), a potential weakness of the approach is the reliance on data sampling. As noted in Figure 2, the sequence space is not fully represented with this data set, which contains 6283 triplet patterns out of the possible 8800 combinations (including the "spacer" to accommodate terminal ends) that are observed in public sequence databases. This leaves nearly a quarter (23.66%) of the sequence space that is otherwise observed in sequence databases unsampled. For example, within the data set, we observe 349, 224, and 214 triplet patterns out of the 440 possible combinations for alanine, arginine, and proline, respectively. A recent analysis of the Protein Data Bank (PDB)(Berman et al., 2000) concluded that only ~47.7% of the entire sequence space is covered by a domain with available structural information (Marsden et al., 2007), suggesting that the limited sequence coverage is reflecting the incomplete structural coverage and is not a limitation of our database. Nonetheless, the potential impact of incomplete data sampling on eScape predictions was addressed in two ways. First, the correlation between the prediction error and the sampling frequency of tripeptides was investigated (Figure 4). Second, the effects of spiking data sequences with randomly selected positions designated to be unrepresented in the library (although in actuality the information is available) were examined (Figure 5). In the former case, we found that the sampling frequency of tripeptides has no significant impact on the magnitude of the error. In the latter case, default values were randomly assigned to positions within the protein sequence (instead of using values corresponding to the tripeptide in the library). Assessment of potential error arising from differences in sampling frequency was determined without retraining the eScape algorithm. The assessment, therefore, reflected the potential error encountered for new sequences containing unrepresented tripeptides in the library.

Three strategies for default value assignment were used to measure the potential performance impact resulting from tripeptides that were not sampled in the library (Figure 2A). Energetic ranges for randomly chosen, hypothetically unsampled, tripeptides were assigned on the basis of the values observed for the (1) position N-terminal to the unrepresented position, (2) energetic range observed for the central amino acid for all tripeptides, and (3) energetic range observed for all residues. In all cases, our results indicate that the impact on the predictions is minimal (mean error $\leq$ 0.1 kcal/mol) when energetic ranges from the previous position is repeated, compared with using known energetic ranges for the amino acid of interest or all amino acids. For all default values, the variance in introduced error is similar until 40% of each sequence contains unsampled tripeptides (Figure 5). This result indicates that the information is both robustly encoded in the protein sequences and redundant within the library of tripeptides.

The effectiveness of the eScape algorithm can be demonstrated by examining three randomly selected proteins: lithostathine, α-lactalbumin, and immunoglobin receptor FcγRIIIb (Figure 6). The predicted ΔG values are compared to those calculated from a COREX analysis of the high-resolution PDB structures. In general, the results show excellent agreement (quantitative in most cases) between the structure-based description of the energy landscape from COREX and the sequence-based reconstruction provided by eScape.

Despite the simplicity of the approach and apparent paucity of data density, the successful reconstruction of the energy landscape with eScape has several noteworthy implications. First, the results suggest that there is a small set of underlying thermodynamic mechanisms that relate sequence to fold, since the successful reconstruction must result from the inherent redundancy of these mechanisms in the data set. Second and equally important, the sequence space that determines whether a residue occupies a certain position is largely coded locally. In fact, previous work has shown that just eight types of thermodynamic

Table 1. Cross-Validated Results for eScape

Adjusted R^2	ΔG	ΔH_{ap}	ΔH_p	TΔS
	0.702	0.695	0.652	0.600
	0.702	0.698	0.651	0.606
	0.701	0.696	0.653	0.606
	0.703	0.692	0.652	0.595
	0.697	0.695	0.650	0.610
	0.705	0.696	0.657	0.607
	0.699	0.695	0.656	0.596
	0.702	0.696	0.655	0.603
	0.702	0.694	0.652	0.596
	0.705	0.694	0.657	0.605
AVERAGE	0.702	0.695	0.653	0.602
PEARSON	0.834	0.837	0.821	0.802
	0.839	0.821	0.818	0.746
	0.843	0.828	0.809	0.756
	0.835	0.847	0.815	0.808
	0.859	0.834	0.825	0.733
	0.811	0.828	0.786	0.748
	0.849	0.836	0.795	0.799
	0.835	0.826	0.801	0.775
	0.836	0.839	0.818	0.807
	0.823	0.837	0.791	0.756
AVERAGE	0.836	0.833	0.808	0.773

The 10-fold cross-validated performance results for eScape are reported for the model trained on natively folded ensembles. Both the adjusted R^2 values and Pearson correlation coefficient is reported for each round.

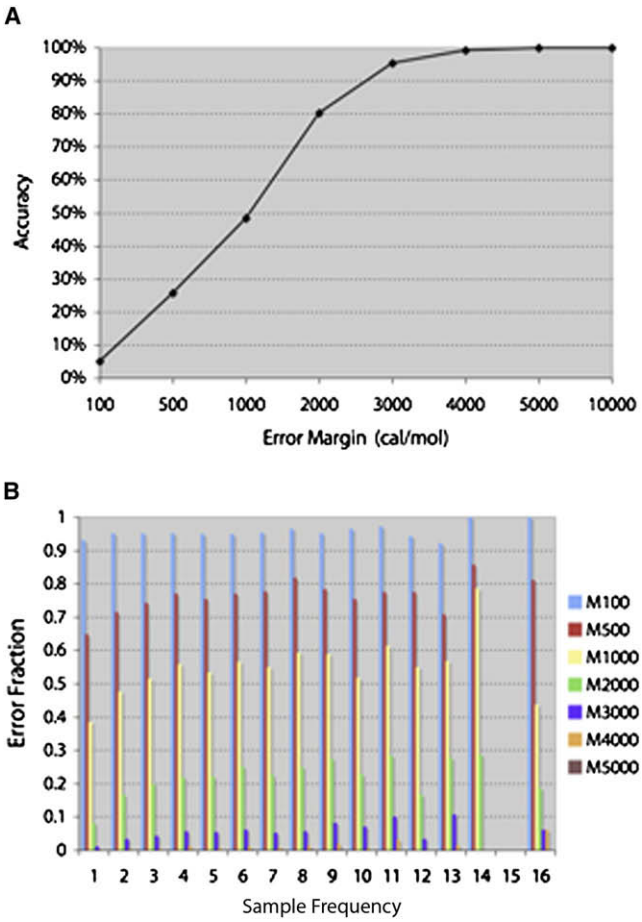

Figure 4. eScape Performance

(A) Prediction accuracy is calculated for incremental error margins defining correct predictions. Approximately 80% of the residues in the dataset can be correctly predicted within an error range of ±2 kcal/mol.

(B) The effect of sampling frequency of tripeptides in the energetic library on predictions was explored. Findings for various error margins (M100 = ±100 cal/mol) show that the sampling frequency has little impact on the performance of the predictor. Error fraction is the fraction of residues where the predicted value is not within the error margin of the COREX calculated value.

environments could serve as elementary energetic building blocks within an entire database of proteins (Larson and Hilser, 2004). Taken together, these findings suggest that the energetic landscapes of proteins can be understood in terms of the physicochemical limits of the local environments in which a residue appears, rather than the direct contribution of that residue to the regional stability of the protein where it is located. The importance of this distinction is that it allows us to establish a quantitative relationship between sequences and the energetic landscapes of proteins, even in the absence of an understanding of how a protein uses its energy landscape to conduct its biological function. Indeed, this approach allows for a high-throughput study to probe for stability patterns within proteomes, studies that could help identify energetic signatures of function. It should be noted, however, that although eScape provides a reasonable estimate of thermodynamic descriptors using only sequence information, the ability of cooperativity and long-range communication to be captured from sequence has not been demonstrated. As such, COREX analysis of the high-resolution structure is still required to obtain this information. Whether such information is ultimately attainable from sequence is currently under investigation.

eScape to Explore Energetic Landscapes of Proteomes

The successful reconstruction of the energetic landscape suggests that the energetic ground rules governing amino acid propensities for different thermodynamic environments (Larson and Hilser, 2004) are sufficiently represented using energy information content encoded in tripeptides (and averaged over a 5-residue sliding window). This leaves open the possibility that eScape can be used as a general tool to evaluate both global thermodynamic properties of proteomes, as well as more local thermodynamic properties within individual proteins, using only sequence information.

To examine differences between the thermodynamics of different proteomes, eScape was used to reconstruct the energetic landscapes for *Arabidopsis thaliana, Caenorhabditis elegans, Escherichia coli, Homo sapiens, Methanococcoides burtonii, Pyrococcus furiosis,* and *Saccharomyces cerevisiae* proteomes obtained from Integr8 (Kersey et al., 2005). The prediction results were then parsed on the basis of shared tripeptide distribution between species (Table 2). Previous comparative analysis between proteomes showed little difference in amino acid

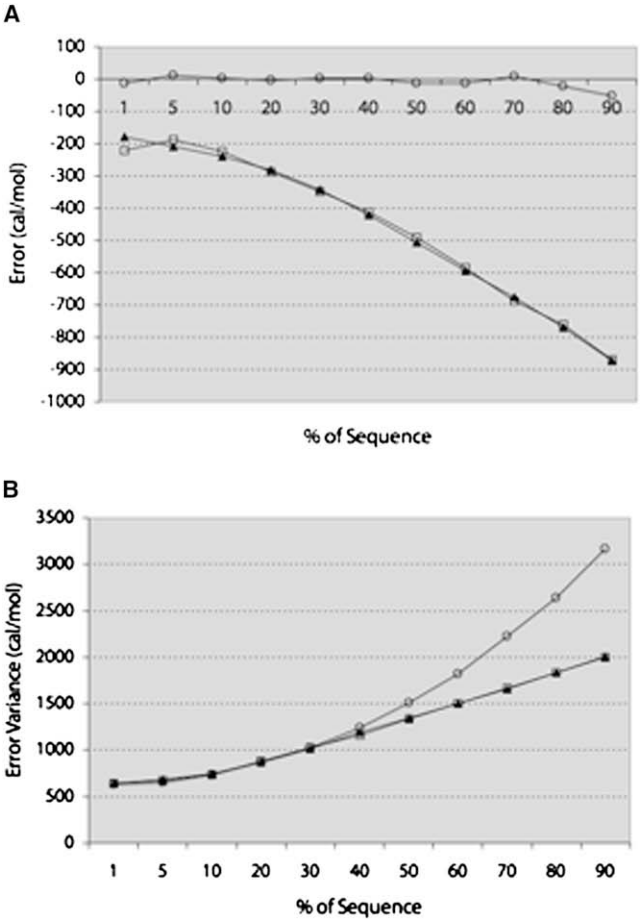

Figure 5. Impact of Unsampled Sequence Space on Prediction Results

The effects of unsampled sequence space were addressed by randomly assigning positions within the sequence (a range from 1% to 90%) designated to be unrepresented in the library. Several replacement strategies were used to assign default values for these randomly selected instances of unrepresented tripeptides by assigning energetic boundaries observed for (1) the position N-terminal to the unrepresented tripeptide (open circles), (2) the amino acid of interest in the absence of higher order information (closed triangle), and (3) that observed for the entire energetic space (open square). The (A) average error and (B) variance show that constraining energetic boundaries to those observed for neighboring residues is better than that based on the central amino acid.

composition between species (Liu and Rost, 2001). Nonetheless, as shown in Table 2, there are distinct preferences for higher order sequence information, and the differences in higher order information are associated with differences in the energy landscapes. Although a substantial portion of the sequence space is shared between all species in Table 2 (i.e., 8270 tripeptides overlap), 166 tripeptides were observed only in either *A. thaliana* (2) or *H. sapiens* (164). Interestingly, the observed stabilities for each category show a general destabilizing trend, with the mean relative stability of the universal set being higher than those that are preferred by a select species. This analysis reveals that *H. sapiens* and *A. thaliana* proteomes have a larger repertoire of unique tripeptides associated with more destabilized regions of the energetic landscape. The observed destabilizing trends agree qualitatively with previous reports indicating that higher organisms contain more intrinsically disordered regions (Liu and Rost, 2001; Ward et al., 2004).

Understanding changes in the energetic landscape through sequence-based approaches can help provide new insights into evolutionary processes, particularly those facilitated by the presence of disordered and highly flexible regions. The findings may help clarify previous observations that used structure-based strategies to understand protein sequence evolution (Deeds et al., 2003; Shakhnovich et al., 2005). For example, a structure-based analysis conducted by Shakhnovich and colleagues reported that proteins with high contact order are more robust and therefore tolerant to high mutation rates. However, Bloom and colleagues reported that the observed increase in mutation rate is largely attributed to regions that are not densely packed in the yeast proteome (Bloom et al., 2006). Clearly, function and evolutionary robustness are linked in complex ways to the global and local stabilities within each protein and within the proteome as a whole. The ability to examine energetic differences within proteins and to explore the conservation of these differences across proteomes provides a unique opportunity to explore (1) the relationship between stability and function in isolated proteins, (2) how mutations affect those relationships, and (3) how changes in other proteins affect these processes.

A Tool for Calculating Stability Profiles in Proteins

As noted previously, disordered or highly dynamic regions have been found to be important for many aspects of protein function, such as catalysis, recognition, and regulation (Eisenmesser et al., 2005; Xiao and Kaltashov, 2005; Xie et al., 2007b). In the case of enzymes, for example, this observation can be reconciled by noting that, for a protein to achieve optimal activity, at least two criteria must be met. First, there must be sufficient population of the active conformation such that it is able to bind its ligand under physiological conditions. Second, the conformational transitions necessary to facilitate a protein's function must be poised at the correct point in the equilibrium so as to respond to its physiological ligand. As a consequence of these demands, proteins have evolved to be marginally stable (DePristo et al., 2005), although little is known about the relationship between stability and function, or whether different proteins share common mechanisms. The observed increase in the number of low stability tripeptides in higher organisms, as demonstrated in Table 2, leaves open the possibility that eScape can be used to (1) locate stretches of both low (i.e., ID) and high stability in proteins and (2) identify thermodynamic signatures that relate stability to function. To demonstrate that eScape can be used to identify conserved regional differences in stability in proteins, we examined the stability of three prominent members of the steroid hormone receptor (SHR) family of TFs. The SHRs are ligand-activated TFs with a domain structure arrangement typical of the nuclear hormone receptor (NHR) superfamily (Figure 7) (Beato, 1989; Evans, 1988; Kumar et al., 1999; Yamamoto, 1985). The ligand-binding domain (LBD) binds the respective steroid, the DNA-binding domain (DBD) binds to its putative response element on the DNA, the hinge region (HR) connects the LBD with the DBD, and the N-terminal domain (NTD) is responsible for transcriptional activation. Importantly, these steroid hormone receptors have been well-established

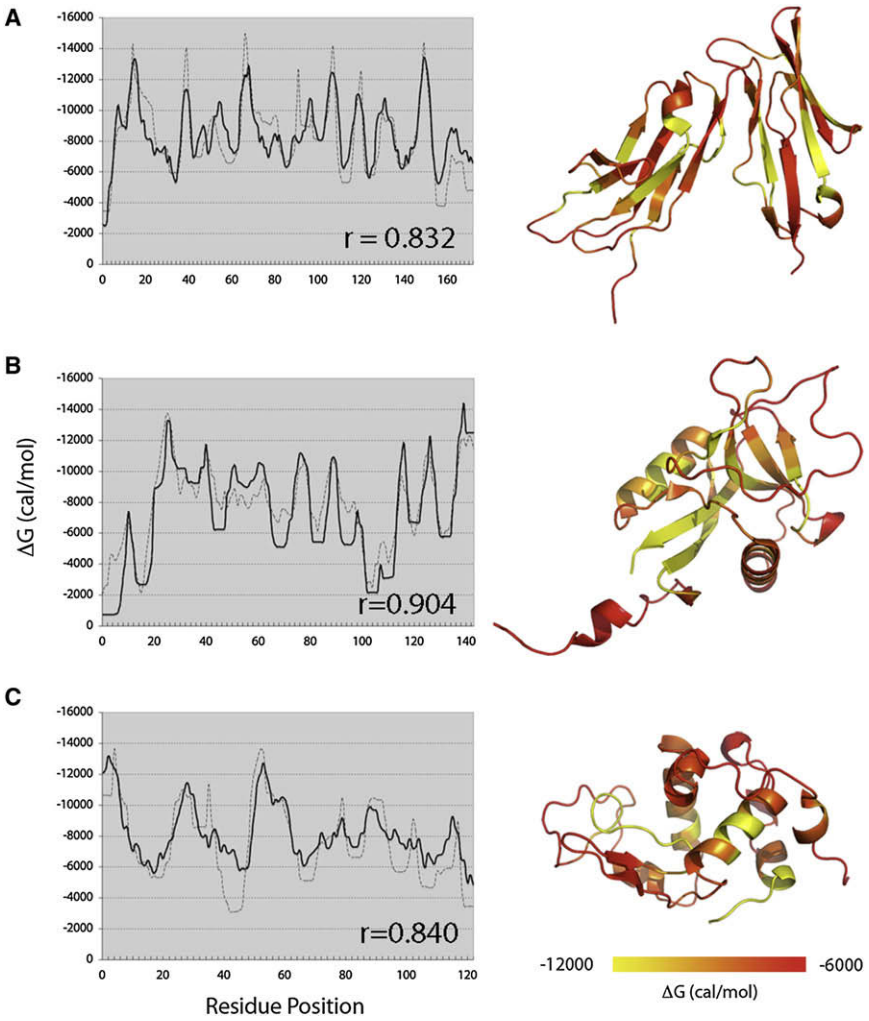

Figure 6. Performance of eScape on Randomly Selected Case Examples

(A–C) Three randomly selected proteins are shown here to provide case examples of eScape performance compared to target values. Stability profiles reported in ΔG values are shown for the native ensemble of (A) immunoglobin receptor FcγRIIIb, (B) lithostathine, (C) α-lactalbumin (PDB IDs: 1FNL, 1QDD, and 1B9O, respectively). The Pearson correlation coefficient (r) is also reported. Comparisons between COREX calculated values (dash lines) and eScape predictions (thick lines) are plotted.

experimentally to contain significant ID regions in the NTD (Kumar et al., 1999, 2007; McEwan et al., 2007).

Shown in Figure 8 is a comparison of the eScape analysis of three SHRs to several disorder predictors. As is clear, similar qualitative trends are observed when predicted values are averaged for each functional domain in estrogen, glucocorticoid, and progesterone receptors. Unlike other disordered predictors, however, eScape provides an estimate of the position-specific stability in energetic terms rather than probabilistic scales and empirical thresholds that define disordered regions. Interestingly, inspection of the eScape results for the SHR family indicates that the average stability is least for the NTD of each SHR, followed by the hinge region. Higher average stabilities are found in the DBDs and LBDs, which are known to be stable, folded structures (Baumann et al., 1993; Kauppi et al., 2003; Lavery and McEwan, 2005; Luisi et al., 1991). Also of interest is that these patterns exist even in the absence of sequence conservation in the hinge and NTD regions.

We note that the comparison of the average stability across the domain is intended only to provide a simplified metric by which to compare different domains, and to compare different metrics of disorder/stability in proteins. Not withstanding this oversimplification, the results do highlight the ability of eScape to provide an energetic description of a protein sequence, and such a metric captures known coarse-grain stability differences. Given that interdomain allosteric coupling has been both proposed on theoretical grounds (Hilser and Thompson, 2007) and demonstrated experimentally to be correlated to domain stability in several proteins (Gekko et al., 2004; Laine et al., 2008), it is possible that eScape will prove to be valuable for efforts geared toward identifying such functions as allostery or signaling from sequence. A test of this hypothesis, however, will require the analysis of multiple proteins families and is outside of the scope of the current study.

Finally, we note that a recent survey of the PDB showed that only 7% of structures have no disordered regions, with 25% of the proteins containing structural information for only 95% of the sequence (Le Gall et al., 2007). This finding indicates that, rather than being a unique facet of SHRs or TFs in general, ID is ubiquitous in the proteome. In short, many proteins have functions involving ID regions, although exactly how ID is used and whether it is quantitatively related to functional properties is not known. Clearly, methods that can quantify these stability differences would provide a valuable tool in understanding the relationship between stability and function.

Conclusions

We have shown that the energetic landscapes of proteins can be reconstructed from sequence information alone and that the apparent thermodynamic complexities seen in protein folds are defined by a common set of underlying energetic determinants. These determinants define the thermodynamic basis of robustness to sequence divergence and allow us to identify conserved energetic signatures for protein sequences. For example, preliminary analysis of the SHR family of transcription factors reveals that significant heterogeneity in the stability of the different domains across the family belies a robustly conserved energetic hierarchy of stability between the domains. Furthermore, eScape

Table 2. Comparison of Tripeptide Distribution and eScape Predictions Between Proteomes

Category	Species	No. of Observances of Unique Tripeptide	Average Stability for Category (cal/mol)
1 Species (166)	*A. thaliana*	2	−7004.82 ±2373.42
	H. sapiens	164	
2 Species (184)	*A. thaliana*	161	−7562.52 ±2304.96
	C. elegans	21	
	E. coli	2	
	H. sapiens	184	
3 Species (26)	*A. thaliana*	26	−7476.64 ±1988.95
	C. elegans	20	
	E. coli	4	
	H. sapiens	26	
	S. cerevisiae	2	
4 Species (10)	*A. thaliana*	10	−8928.64 ±3108.52
	C. elegans	10	
	E. coli	5	
	H. sapiens	10	
	S. cerevisiae	5	
5 Species (30)	*A. thaliana*	30	−8627.47 ±3695.56
	C. elegans	30	
	E. coli	20	
	H. sapiens	30	
	M. burtonii	7	
	P. furiosus	5	
	S. cerevisiae	28	
6 Species (112)	*A. thaliana*	112	−9117.29 ±2876.70
	C. elegans	112	
	E. coli	109	
	H. sapiens	112	
	M. burtonii	73	
	P. furiosus	42	
	S. cerevisiae	112	
7 Species (8270)	*A. thaliana*	8270	−8103.5643 ±2018.48213
	C. elegans	8270	
	E. coli	8270	
	H. sapiens	8270	
	M. burtonii	8270	
	P. furiosus	8270	
	S. cerevisiae	8270	

eScape predictions were performed for proteomes of 7 species (*A. thaliana, C. elegans, E. coli, H. sapiens, M. burtonii, P. furiosis,* and *S. cerevisiae*) with representation for all three kingdoms of life. Of the 8800 possible tripeptides, this set was parsed into 7 categories on the basis of the number of species in which the tripeptide was observed; 8270 tripeptides were observed in all seven species where as 166 tripeptides were only observed in *A. thaliana* (2) and *H. sapiens* (164). The predicted stability values were partitioned on the basis of shared tripeptides between species and averages for each category show an increasingly destabilizing trend from the universal set to those additionally sampled by higher organisms.

provides a means of evaluating differences in average thermodynamic behavior across entire proteomes. Our analysis reveals that thermodynamic signatures utilizing ID are a property of higher organisms, a result that is in agreement with numerous disorder predictors. It is hoped that the ability to look at regional differences in stability within proteins, as well as changes in stability across entire proteomes, will provide opportunities to understand the complex interplay between the response of individual proteins to selective pressures and the response of the entire proteome.

EXPERIMENTAL PROCEDURES

eScape Training Data Set

A nonredundant set of human proteins (Larson and Hilser, 2004) was used for the development of eScape to reconstruct protein stability profiles using only sequence information. A total of 122 X-ray crystal structures with no missing residues, terminal ends exempted from this criterion, were used. These structures are 50–250 residues in length with <2.5 Å resolution and a maximum of 50% sequence identity within the data set.

The COREX Algorithm

The COREX algorithm is a statistical thermodynamic model in which a native protein is systematically unfolded to depict an ensemble of states rather than be represented as a single static structure (D'Aquino et al., 1996). The ensemble comprises the native protein, ranging from the fully folded to denatured conformational states. The energetics of each of the 122 proteins in the eScape training data set was calculated using the COREX algorithm. For proteins larger than 80 residues, Monte Carlo sampling (50,000 states/partition) was used to generate ensembles for consideration of computational demands, followed by full COREX enumeration. For proteins less than 80 residues, all states in the ensemble were fully enumerated.

We briefly describe the COREX algorithm and ask readers to refer to references (Hilser and Freire, 1996; Hilser et al., 2006) for more details. Under equilibrium conditions, the probability of any given conformational microstate, *i*, in the ensemble is given by the following equation:

$$P_i = \frac{K_i}{\sum_{i=1}^{N_{states}} K_i} = \frac{K_i}{Q}, \quad (2)$$

where $K_i = e^{(-\Delta G_i/RT)}$ is the statistical weight of each microstate, and R is the gas constant, for a given absolute temperature T. The summation in the denominator is the partition function, Q, for the system. The Gibbs free energy for each microstate, ΔG_i, is calculated as:

$$\Delta G_i = \Delta H_{i,solvation} - T(\Delta S_{i,solvation} + W\Delta S_{i,conformational}), \quad (3)$$

where W is an entropy-weighting factor used to control the contributions of the natively folded state. This entropy-weighting factor enables us to perturb the ensemble to favor denature or natively folded states, thus allowing us to investigate thermodynamic properties under natively folded or denaturing conditions. During the calculation of position stability, an entropy-weighting factor of W = 0.5 is used to increase natively folded states in the population and consider, for the most part, contributions to local stability from the native conformation of the protein.

The equilibrium of the natively folded and unfolded states of proteins can be evaluated for each residue using a statistical descriptor defined as the residue stability constant, $\kappa_{f,j}$ (D'Aquino et al., 1996). 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 ($\Sigma P_{f,j}$) to the summed probability of all states in which j is in an unfolded conformation ($\Sigma P_{nf,j}$):

$$k_{f,j} = \frac{\sum P_{f,j}}{\sum P_{nf,j}}. \quad (4)$$

From the stability constant, the position-specific free energy expressed in units of cal/mol can be written as:

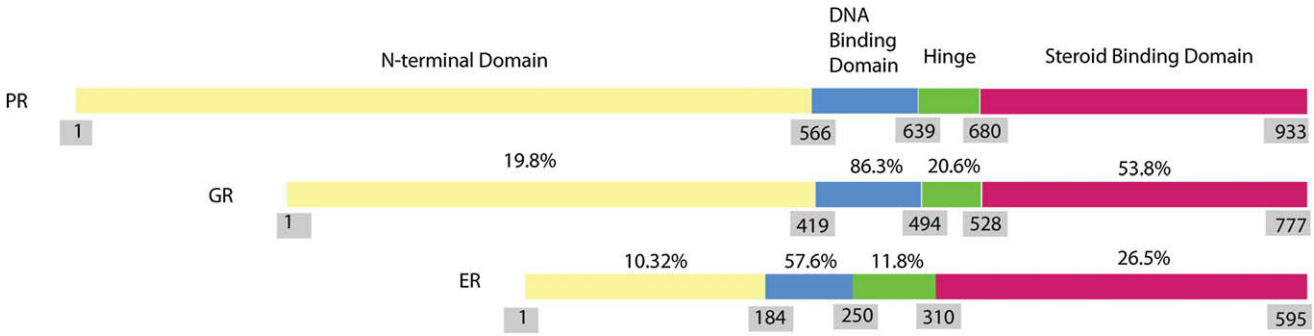

Figure 7. Schematic Representation of Steroid Hormone Receptors

A schematic representation of the steroid hormone receptor architecture containing four domains: the N-terminal domain, DNA-binding domain, hinge, and steroid-binding domain. Differences between sequence identities for the respective four domains between the progesterone receptor (PR, 933 residues), glucocorticoid receptor (GR, 777 residues, α variant), and estrogen receptor (ER, 595 residues, α isoform) are reported. These domain boundaries are used to identify energetic imbalances between domains that may be important for allosteric communication.

$$[\Delta G]_j = -RT \cdot \ln \frac{\sum P_{f,j}}{\sum P_{nf,j}} = \langle \Delta G_{f,j} \rangle - \langle \Delta G_{nf,j} \rangle. \quad (5)$$

The importance of the stability constant and its good agreement with experimental data had been shown before with hydrogen deuterium exchange comparisons (Hilser and Freire, 1996).

Position-Specific Thermodynamic Descriptors

Position-specific thermodynamic descriptors were calculated by taking the difference in folded and unfolded subensemble quantities.

$$[\Delta H]_{pol,\,j} = < \Delta H_{pol,\,f,\,j} > - < \Delta H_{pol,\,nf,\,j} > \quad (6)$$

$$[\Delta H]_{apol,\,j} = < \Delta H_{apol,\,f,\,j} > - < \Delta H_{apol,\,nf,\,j} > \quad (7)$$

$$[\Delta S]_{conf,\,j} = < \Delta S_{conf,\,f,\,j} > - < \Delta S_{conf,\,nf,\,j} > \quad (8)$$

Quantities in folded and unfolded subensembles were calculated (Wrabl et al., 2002) as:

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

$$\langle \Delta S \rangle = \sum_{i=1}^{N_{states}} P_i \bullet \Delta S_i = \sum_{i=1}^{N_{states}} \frac{K_i \bullet \Delta S_i}{Q}. \quad (10)$$

Development of eScape

Analysis of higher order sequence association with COREX calculated position-specific stability showed that the energetic landscape can be partitioned on the basis of tripeptide patterns. This suggested that the heuristic information could be leveraged in the reconstruction of protein stability profiles using only sequence information. Several different input features and regression strategies were explored in the pursuit of a sequence-based energetic predictor. The ultimate strategy adopted in eScape leverages the concept that the range of stability values in which residues are found in the energetic landscape is limited. The procedure in which the predictions are made is illustrated in Figure 3 and here described in detail.

Predictions are made by first drawing energetic boundaries for each position of the sequence from the library of COREX calculated energetic values for the nonredundant set of proteins. Energetic boundaries (i.e., the minimum and maximum values) observed for the corresponding tripeptide at the position are then respectively averaged across the sequence with a sliding window size of 5. The final averaged values are used as input features for a trained linear regression model, the core of eScape, to make predictions about the position-specific thermodynamic descriptors that have been modeled with COREX. The trained linear regression model which best fit the training data for each of the four position-specific thermodynamic descriptors are:

$$\Delta G = [(0.8195 * min_{i,\,\Delta G}) + (0.7492 * max_{i,\,\Delta G})] + 4696;$$

$$\Delta H_{ap} = [(0.7665 * min_{i,\,\Delta Hap}) + (0.7632 * max_{i,\,\Delta Hap})] - 5068;$$

$$\Delta H_{p} = [(0.7791 * min_{i,\,\Delta Hp}) + (0.7524 * max_{i,\,\Delta Hp})] + 6195;$$

$$T\Delta S = [(0.7047 * min_{i,\,T\Delta S}) + (0.7507 * max_{i,\,T\Delta S})] + 1998;$$

where $min_{i,\,x}$ and $max_{i,\,x}$ corresponds to the minimum and maximum observed range of the thermodynamic descriptor for the corresponding tripeptide at the given position. eScape was conducted with 10-fold cross-validation with performance reported at an average adjusted R^2 value of 0.70 and an average Pearson's correlation coefficient of 83.63% (Table 1). Training using the cross-validation strategy ensures that the subset of the training set (10%) used to test the performance of the predictor was not included in the training set. Therefore, predicted values are not based on data to which it was trained. Finally, we note that because only the energetic boundaries (i.e., extrema) were used to select position-specific stability values, not all values in the library are used.

Proteomic Analysis

Entire proteomes of *A. thaliana, C. elegans, E. coli, H. sapiens, M. burtonii, P. furiosis,* and *S. cerevisiae,* were obtained from Integr8 (http://www.ebi.ac.uk/integr8) (Kersey et al., 2005). Position-specific stabilities were predicted using eScape method described here. To accommodate unsampled tripeptides that are not represented in the eScape library, input values used for the previous position (N-terminal to the position of interest) were used for subsequent reconstruction of energetic profiles. Proteomes were partitioned into tripeptide fragments to examine the sampling coverage between different species and associated position-specific stability values.

Steroid Hormone Receptor Case Study

The estrogen, glucocorticoid, and progesterone steroid hormone receptors were used as a case study to demonstrate the applicability of eScape in identifying energetic profiles within multidomain proteins. The accession numbers for sequences retrieved from NCBI used in this study are gi544257, gi121069, and gi90110048. Alignments to calculate sequence identities between domains were performed using ProbCons (Do et al., 2005).

ACKNOWLEDGMENTS

This study was supported by grants from the NIH (GM13747) and the Robert A. Welch Foundation (H-1461). J.G. would like to thank the Jeanne Kempner Scholar Foundation for financial support. We would also like to thank James

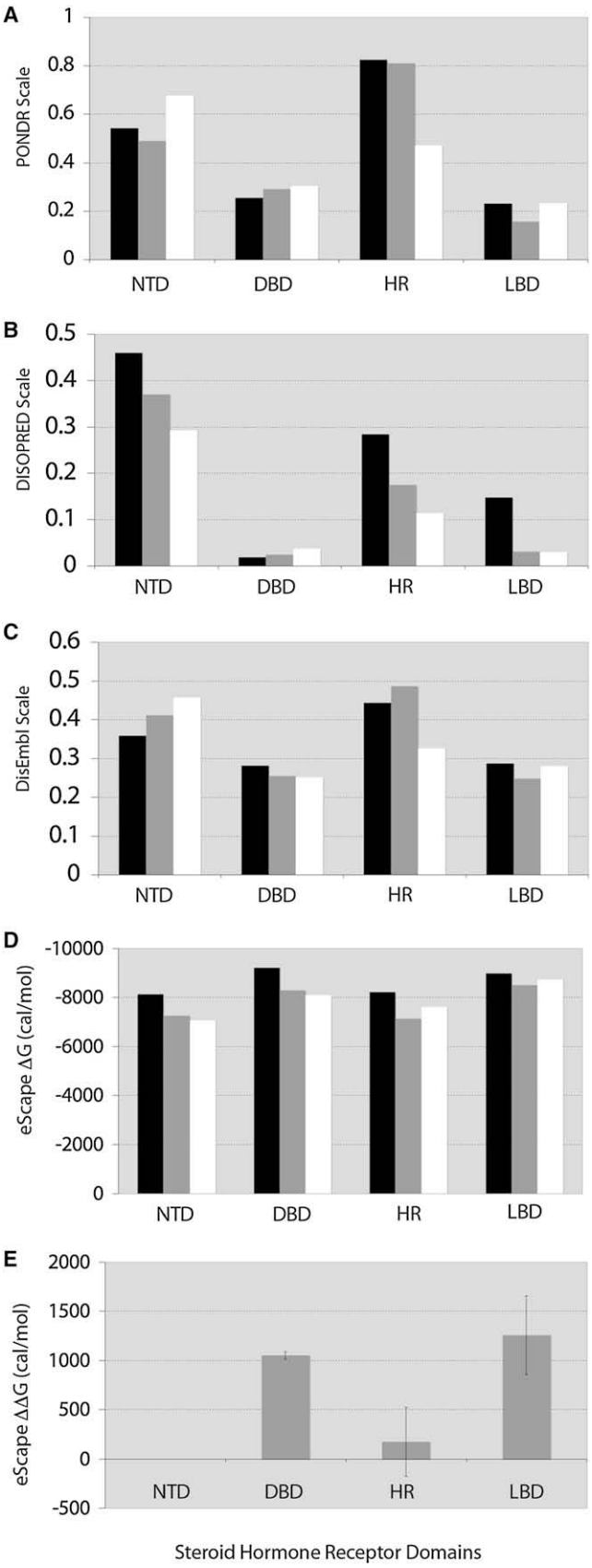

Figure 8. Application of eScape for Steroid Hormone Receptors

(A–E) Comparison of eScape to several disorder predictors shows the same qualitative trends when applied to estrogen (black bars), glucocorticoid (gray bars), and progesterone (white bars) receptors. The prediction values of (A) PONDR (Romero et al., 1997a, 1997b), (B) DISOPRED (Jones and Ward, 2003), (C) DisEMBL (Linding et al., 2003), and (D) eScape were averaged on the basis of domain boundaries. On the basis of values estimated by eScape, (E) the average energetic differences between domains ($\Delta\Delta G$) were normalized relative to the N-terminal domain for the three nuclear receptors. Note that for the eScape predictions, (1) the stabilities are presented relative to the unfolded state (i.e., negative is more stable), and (2) the stabilities have been determined using an entropy weighting factor of W = 0.5 (see Equation 3). Although the eScape ΔGs (D) are sensitive to the value of W used for the calculation, eScape $\Delta\Delta G$s are relatively insensitive. Error bars represent one standard deviation of the differences for each respective domain averaged across the three steroid hormone receptors.

Wrabl and Steven Whitten for helpful discussions and critical review of the manuscript.

Received: June 10, 2008
Revised: August 7, 2008
Accepted: August 19, 2008
Published: November 11, 2008

REFERENCES

Baumann, H., Paulsen, K., Kovacs, H., Berglund, H., Wright, A.P., Gustafsson, J.A., and Hard, T. (1993). Refined solution structure of the glucocorticoid receptor DNA-binding domain. Biochemistry *32*, 13463–13471.

Beato, M. (1989). Gene regulation by steroid hormones. Cell *56*, 335–344.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bloom, J.D., Drummond, D.A., Arnold, F.H., and Wilke, C.O. (2006). Structural determinants of the rate of protein evolution in yeast. Mol. Biol. Evol. *23*, 1751–1761.

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.

Daughdrill, G.W., Narayanaswami, P., Gilmore, S.H., Belczyk, A., and Brown, C.J. (2007). Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. J. Mol. Evol. *65*, 277–288.

Deeds, E.J., Dokholyan, N.V., and Shakhnovich, E.I. (2003). Protein evolution within a structural space. Biophys. J. *85*, 2962–2972.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. *6*, 678–687.

Do, C.B., Mahabhashyam, M.S., Brudno, M., and Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. *15*, 330–340.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z. (2002). Intrinsic disorder and protein function. Biochemistry *41*, 6573–6582.

Eisenmesser, E.Z., Millet, O., Labeikovsky, W., Korzhnev, D.M., Wolf-Watz, M., Bosco, D.A., Skalicky, J.J., Kay, L.E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. Nature *438*, 117–121.

Evans, R.M. (1988). The steroid and thyroid hormone receptor superfamily. Science *240*, 889–895.

Gekko, K., Obu, N., Li, J., and Lee, J.C. (2004). A linear correlation between the energetics of allosteric communication and protein flexibility in the *Escherichia coli* cyclic AMP receptor protein revealed by mutation-induced changes in compressibility and amide hydrogen-deuterium exchange. Biochemistry *43*, 3844–3852.

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., and Thompson, E.B. (2007). Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. Proc. Natl. Acad. Sci. USA *104*, 8311–8315.

Hilser, V.J., Garcia-Moreno, E.B., Oas, T.G., Kapp, G., and Whitten, S.T. (2006). A statistical thermodynamic model of the protein ensemble. Chem. Rev. *106*, 1545–1558.

Jones, D.T., and Ward, J.J. (2003). Prediction of disordered regions in proteins from position specific score matrices. Proteins *53* (*Suppl 6*), 573–578.

Kauppi, B., Jakob, C., Farnegardh, M., Yang, J., Ahola, H., Alarcon, M., Calles, K., Engstrom, O., Harlan, J., Muchmore, S., et al. (2003). The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. J. Biol. Chem. *278*, 22748–22754.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., et al. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. Nucleic Acids Res. *33*, D297–D302.

Kumar, R., Baskakov, I.V., Srinivasan, G., Bolen, D.W., Lee, J.C., and Thompson, E.B. (1999). Interdomain signaling in a two-domain fragment of the human glucocorticoid receptor. J. Biol. Chem. *274*, 24737–24741.

Kumar, R., Serrette, J.M., Khan, S.H., Miller, A.L., and Thompson, E.B. (2007). Effects of different osmolytes on the induced folding of the N-terminal activation domain (AF1) of the glucocorticoid receptor. Arch. Biochem. Biophys. *465*, 452–460.

Laine, O., Streaker, E.D., Nabavi, M., Fenselau, C.C., and Beckett, D. (2008). Allosteric signaling in the biotin repressor occurs via local folding coupled to global dampening of protein dynamics. J. Mol. Biol. *381*, 89–101.

Larson, S.A., and Hilser, V.J. (2004). Analysis of the "thermodynamic information content" of a *Homo sapiens* structural database reveals hierarchical thermodynamic organization. Protein Sci. *13*, 1787–1801.

Lavery, D.N., and McEwan, I.J. (2005). Structure and function of steroid receptor AF1 transactivation domains: induction of active conformations. Biochem. J. *391*, 449–464.

Le Gall, T., Romero, P.R., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2007). Intrinsic disorder in the Protein Data Bank. J. Biomol. Struct. Dyn. *24*, 325–342.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003). Protein disorder prediction: implications for structural proteomics. Structure *11*, 1453–1459.

Liu, J., and Rost, B. (2001). Comparing function and structure between entire proteomes. Protein Sci. *10*, 1970–1979.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006). Intrinsic disorder in transcription factors. Biochemistry *45*, 6873–6888.

Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., and Sigler, P.B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. Nature *352*, 497–505.

Marsden, R.L., Lewis, T.A., and Orengo, C.A. (2007). Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. BMC Bioinformatics *8*, 86.

McEwan, I.J., Lavery, D., Fischer, K., and Watt, K. (2007). Natural disordered sequences in the amino terminal domain of nuclear receptors: lessons from the androgen and glucocorticoid receptors. Nucl Recept Signal *5*, e001.

Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N., and Dunker, A.K. (2007). Intrinsic disorder and functional proteomics. Biophys. J. *92*, 1439–1456.

Romero, P., Obradovic, Z., and Dunker, A.K. (1997a). Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Inform. Ser. Workshop Genome Inform. *8*, 110–124.

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K. (1997b). Identifying disordered regions in proteins from amino acid sequences. Proc. I.E.E.E. International Conference on Neural Networks *1*, 90–95.

Shakhnovich, B.E., Deeds, E., Delisi, C., and Shakhnovich, E. (2005). Protein structure and evolutionary history determine sequence space topology. Genome Res. *15*, 385–392.

Uversky, V.N. (2002). Natively unfolded proteins: a point where biology waits for physics. Protein Sci. *11*, 739–756.

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.

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.

Wrabl, J.O., Larson, S.A., and Hilser, V.J. (2002). Thermodynamic environments in proteins: fundamental determinants of fold specificity. Protein Sci. *11*, 1945–1957.

Xiao, H., and Kaltashov, I.A. (2005). Transient structural disorder as a facilitator of protein-ligand binding: native H/D exchange-mass spectrometry study of cellular retinoic acid binding protein I. J. Am. Soc. Mass Spectrom. *16*, 869–879.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., and Uversky, V.N. (2007a). Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J. Proteome Res. *6*, 1917–1932.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., and Obradovic, Z. (2007b). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. *6*, 1882–1898.

Yamamoto, K.R. (1985). Steroid receptor regulated transcription of specific genes and gene networks. Annu. Rev. Genet. *19*, 209–252.

Yi, S., Boys, B.L., Brickenden, A., Konermann, L., and Choy, W.Y. (2007). Effects of zinc binding on the structure and dynamics of the intrinsically disordered protein prothymosin alpha: evidence for metalation as an entropic switch. Biochemistry *46*, 13120–13130.