The Dynameomics Entropy Dictionary: A Large-Scale Assessment of Conformational Entropy Across Protein Fold Space

Clare-Louise Towse[†§], Mikael Akke[‡] and Valerie Daggett[†]*

[†]Department of Bioengineering, University of Washington, Box 355013, Seattle, WA 98195-5013, USA

[‡]Department of Biophysical Chemistry, Lund University, PO Box 124, SE-22100 Lund, Sweden

[§]Current Address: Department of Chemistry and Biosciences, Faculty of Life Sciences, University of Bradford, Richmond Road, Bradford, BD7 1DP, United Kingdom

*Corresponding author: Valerie Daggett, Email: daggett@uw.edu

Running title: Conformational Entropy from Dynameomics

ABSTRACT

Molecular dynamics (MD) simulations contain considerable information with regard to the motions and fluctuations of a protein, the magnitude of which can be used to estimate conformational entropy. Here we survey conformational entropy across protein fold space using the Dynameomics database, which represents the largest existing dataset of protein MD simulations for representatives of essentially all known protein folds. We provide an overview of MD-derived entropies accounting for all possible degrees of dihedral freedom on an unprecedented scale. Although different side chains might be expected to impose varying restrictions on the conformational space that the backbone can sample, we found that the backbone entropy and side chain size are not strictly coupled. An outcome of these analyses is the Dynameomics Entropy Dictionary, the contents of which have been compared with entropies derived by other theoretical approaches and experiment. As might be expected, the conformational entropies scale linearly with the number of residues, demonstrating that conformational entropy is an extensive property of proteins. The calculated conformational entropies of folding agree well with previous estimates. Detailed analysis of specific cases identify deviations in conformational entropy from the average values that highlight how conformational entropy varies with sequence, secondary structure, and tertiary fold. Notably, α-helices have lower entropy on average than do β-sheets, and both are lower than coil regions.

INTRODUCTION

Conformational entropy is increasingly being recognized as a potentially important driving force for biologically relevant processes involving proteins and other macromolecules. Ligand binding and allostery have been shown to depend sensitively on changes in conformational entropy.[1-9] Methods to derive protein conformational entropy from structural and simulation data date back over thirty years.[10-11] Due to methodological challenges to experimentally characterize conformational entropy, most approaches have been theoretical, although some benchmarking against experimental data has been performed.[12] Attempts have been made to derive conformational entropy from dihedral angle distributions observed in a database of individual protein X-ray crystal structures.[13] It has been demonstrated that conformational entropy can be estimated from NMR relaxation studies and the derived generalized order parameter, S^2, which offers a unique source of experimental information on dihedral angle distributions.[1,14-16] As initially shown for cooperative binding of Ca^{2+} to calbindin D_{9K},[1] such studies can address the relationship between internal motions and biological function.

Given the continuing discussions regarding conformational entropy due to its importance in protein folding, ligand binding, allostery, and functionally relevant conformational changes, it is essential that the conformational entropy and the relationship to any experimental observables be well defined and characterized. Molecular dynamics (MD) simulations provide a critical complement to NMR order parameters, because both S^2 and conformational entropy can be calculated from the generated coordinate distributions, thereby enabling benchmarking between simulation and experiment, as well as a mechanistic interpretation of the underlying dynamics. For example, methyl-axis S^2 values have been shown to partition as a measure of rotamer popu-

lations.[17-19] A number of MD studies have investigated the relationship between S^2 and conformational entropy.[20-25] Most importantly, MD simulations yield the total conformational entropy of the sampled ensemble, while NMR is limited by probing specific degrees of freedom, e.g., via the ^{15}N backbone order parameter.

Several important studies based on MD simulations have contributed to our understanding of conformational entropy as an important factor in ligand binding.[9,26-27] Aside from the problem of correspondence between experimental variables and conformational entropy, a number of important questions arise regarding conformational entropy per se, such as the degree of variability within proteins as a function of amino-acid residue type, location in secondary structure elements, and the extent of coupling between dihedral angles. However, previous studies have used only a single protein or very small datasets that limit the generality of these results.

Here we present a large-scale characterization of conformational entropy in proteins, based on the Dynameomics MD dataset that represents 807 different proteins covering essentially all known protein folds. The dataset thereby reduces the potential topological bias and far exceeds the statistical power of the earlier studies performed on only a few proteins. As a result, we are able to address general features of conformational entropy across protein fold classes, secondary structure elements, and residue types. Dynameomics is a large-scale effort to explore protein dynamics and unfolding, in a broad and systematic manner, across protein fold space.[28,29] Protein fold space was defined by constructing a consensus domain dictionary to evaluate and rank protein domains into metafolds based on population in the Protein Data Bank (PDB); the top 807 metafolds were determined to cover 97% of the known autonomous globular protein folds.[30,31] Here we present the analysis of 2,421 simulations of these 807 representative proteins

at 298 K and 498 K, with a total simulation time of 80,700 ns (81 μs). In addition, we have determined benchmark entropies using an 'idealized' GGXGG dataset, commonly used as a random coil reference in NMR studies, to provide intrinsic amino acid propensities where 'X' is any of the 20 proteinogenic amino acids.[32] We also highlight some case studies comparing multiple members of highly populated metafolds and a heteromorphic high sequence-identity domain pair. From this large-scale analysis of dihedral angle distributions from simulations of the GGXGG host-guest series, and the 807 protein folds under native state (298 K) and elevated temperature (498 K) conditions, we have surveyed the conformational entropies of proteins and their constituent amino acids and local secondary structure to create a 'dictionary' of entropy values, the Dynameomics Entropy Dictionary (DED).

In total, three dictionaries were generated, one for intrinsic 'random coil' entropy values (GGXGG data), one for native state entropy values (298K protein data) and one for denatured state entropy values (498K protein data). The absolute entropies in these dictionaries were then used to determine ΔS values, yielding good agreement between calculated ΔS_{fold} values with previously determined experimental values. Naturally, fine detail is lost in the creation of such dictionaries that compute average entropies across such a large dataset, but we show that it can be teased out using entropy profiles of the deviations between the entropy of a given residue and the dictionary value highlighting relationships with secondary structure and topology.

METHODS

Simulation Datasets. Two datasets were used in this work: a host-guest series to model intrinsic propensities using GGXGG peptides,[32] 'X' representing each of the proteinogenic ami-

no acids, and a database of globular protein simulations. The latter dataset is the Dynameomics database[28,29] which houses simulations of 807 proteins (D807) that represent 97% of all known autonomous protein folds selected from our Consensus Domain Dictionary.[30,31] Additional simulation data were used for cases studies that included a heteromorphic protein pair, the GA88 and GB88 protein domains,[33] and related folds within the Dynameomics rank 1 and rank 5 metafold families (Table S1) for comparison with the Dynameomics dataset.

All proteins and peptides were simulated as microcanonical (NVE) ensembles with explicit, flexible F3C water models[34,35] at 298 K and 498 K using *in lucem* molecular mechanics (*il*mm)[36] and the Levitt *et al.* force field.[37] The D807 298 and 498 K data are currently available up to 51 ns, with duplicate simulations at 498 K for each target. For the GGXGG dataset, the peptides were simulated with their N and C-termini acetylated and amidated, respectively, for 101 ns. In all cases, the first nanosecond was discarded as an initial equilibration period. For the analysis, 50 ns of data were used from all the datasets. For the D807 dataset, all 50 ns of production dynamics from the 298 K simulations was used, 25 ns were taken from the latter portion of each duplicate set of 498 K simulations to represent the denatured state and 50 ns taken from the latter portion of the GGXGG simulations. The G_A88 and G_B88 domains simulations and have been presented in depth elsewhere.[38] The lengths of the simulations for the Rank1 and Rank 5 sets are detailed in Table S1. The first nanosecond was discarded as equilibration for all simulations. Cysteine residues are present in the D807 dataset in disulfide bonded and protonated forms, but the short length of the GGXGG peptides precludes modeling of the disulfide bonded form of Cys, hence only entropies for the protonated form (Cyh in our force field) are given for the GGXGG host-guest series.

Assessment of Datasets The convergence and suitability of using the latter 50 ns of the 100 ns GGXGG simulations for our comparative analyses with our Dynameomics dataset has been shown previously to sample conformational space sufficiently and consistently over the trajectories with negligible increase in the coverage of ϕ/ψ space above 50 ns.[39] We reconfirmed this by calculating ϕ/ψ entropies over the first and second 50 ns portions of the trajectories for the full GGXGG series and the D807 test set of 51 ns trajectories (Tables S2 and S3). In both instances, these calculations provided the same entropies for the first and latter trajectory portions with negligible differences and showed the use of 5° bin widths to be acceptable (Table S2-S5). The data for the 807 trajectories has previously been scrutinized and the dihedral angle distributions obtained for the 51 ns D807 trajectories were as previously reported.[18]

Calculation of Entropy. Entropies were estimated from the side chain and main chain dihedral angle motions throughout the simulations. All dihedral angle values were saved every 0.2 ps and, for a given dihedral angle, the entropy was evaluated as a Boltzmann sampling of states using equation (1) for a given angle, θ, where p_i is the fractional population of microstate i and, R, the gas constant to convert the units into cal mol^{-1} K^{-1}; the exact number used for R was 1.98721 cal mol^{-1} K^{-1} based on 1 cal = 4.184 J. Microstates were defined in 5° increments and the populations of these microstates were obtained using histograms of the dihedral angle distributions with a 5° bin width. The populations in each bin were normalized using the simulation length to give a fractional population for each bin, p_i. In this manner, the entropic contribution for amino acid side chains, S_{sc}, was taken as a simple summation of the entropies for the dihedral angles in the residue. Here, this summation of the dihedral angle entropies was used to estimate the entropy of amino acid side chain, S_{sc}, using the dihedral angles in the side chain group.

For calculation of the main chain entropy of each amino acid residue (S_{mc}) the coupling between the main chain dihedral angles was captured by determining the population of microstates using using three-dimensional histograms with 5° bin widths in each dimension, with the three dimensions pertaining to ϕ, ψ and ω dihedral angles. When using the three-dimensional histograms, the probability of the i,j,k^{th} 5°×5°×5° bin, p_{ijk}, replaces p_i. Total entropies for each residue, S_{res}, were then calculated as a summation of the constituent entropies of the relevant main chain and side chain dihedral angles, equation (2). For the 498 K data, as shorter segments of the simulations were used, histograms over duplicate simulations for a single angle were generated and normalized accordingly. All dihedral angles were included and are defined within the Levitt *et al*. force field (Figure S1).

$$S_\theta = -R \sum p_i ln p_i \quad (1)$$

$$S_{res} = \sum S_{mc} + S_{\chi 1} + \cdots + S_{\chi n} \quad (2)$$

For tractability and because the length of the side chains varies, the main chain and side chain entropies were computed independently and coupling between angles was only accounted for in calculating the main chain entropies. Terminal residues, not having a full complement of ϕ/ψ angles for computing the main chain entropy, were excluded from the calculations. Validation of the selected bin width and assessment of coupling between angles is contained in the Supplementary Information (Tables S2-S8 and Figure S2).

RESULTS AND DISCUSSION

The Dynameomics MD data were acquired using our in-house molecular modeling package *in lucem* molecular mechanics (*il*mm)[36] and an established force field and methods.[37,40-41] The available molecular mechanics packages have a common origin, beginning with CFF (Consistent Force Field), as has been described previously by Levitt.[42] For reference, our force field and methods are direct descendants of CFF, via EREF and ENCAD[43] from the Levitt Lab, then being re-written in the Daggett Lab using modern software engineering techniques. Thus, *il*mm is related to AMBER, CHARMM and other packages, but ENCAD and *il*mm focus on the simulation of fully flexible, unrestrained systems. One critical feature is the use of a simple, fully flexible water model (F3C) that provides excellent agreement with experimental diffusion constants and radial distribution functions even at elevated temperatures.[34-35,40] We also do not employ algorithms that restrict bond and angle motion, such as SHAKE or artificially high masses on hydrogens. We use the microcanonical ensemble (NVE, constant number of particles and energy) to avoid the use of temperature and pressure coupling, which allows for continuous trajectories for characterization of pathways, and we do not employ EWALD summations, which can lead to artificial periodicity.[44]

These approaches provide good agreement with a variety of experimental results, and in particular for our purposes here, they include native state dynamics and S^2 order parameters.[18,45-46] They have also been shown to have predictive power for moving beyond the native state, such as in protein folding/unfolding pathways. There include transition states (TS),[47-49] intermediate states,[50-52] denatured states.[53-55]. To better illustrate this, we focus briefly on one simple system, the 3-helix bundle engrailed homeodomain (EnHD), with experimental studies performed in the Fersht Lab. The unfolding time by MD is in good agreement with experiment,[56] the MD-

predicted TS structure is in good agreement with experiment,[56-58] the structure of the MD-predicted intermediate state[56] was confirmed by direct NMR studies 5 years later.[59] Furthermore, the unfolding pathway is independent of temperature and merely accelerated at high temperature,[56-58] which is an important benchmark for the high temperature simulations used in this study. In addition, microscopic reversibility is observed[60] as is refolding from the intermediate state.[61-62] Further details regarding comparisons and benchmarking against experiment have been reviewed.[63-65] Our Dynameomics project moves away from such detailed investigations of well-studied model systems to explore the properties of all protein folds, but the earlier studies presented above (and others) represent the groundwork for studies making use of the Dynameomics Database.

Determination of Conformational Entropy. Using the dihedral angle distributions from intrinsic propensity models (GGXGG), native state and elevated temperature simulations of 807 protein folds, three dictionaries of entropy values were created (Table 1). For any given amino acid residue there are a large number of main chain and side chain dihedral angles that contribute to a residue's entropy, S_{res}. Although some coupling between the angles can be expected,[10,66] there is evidence suggesting that the entropy of the side chain and protein backbone can be independent of one another with negligible coupling.[4,21,67-68] To determine how to best approach calculation of amino acid entropies, we established the extent of coupling between angles using a test set (Table S1)[69] and examined the effect of different bin widths (Tables S2-S5). The test set contained 11 proteins from the Dynameomics dataset spanning a range of different architectures and sizes from 29 to 399 residues[69] (Table S6). For every possible combination of main chain and side chain dihedral angles, the entropies of individual angles were calculated using 1D histo-

grams and summed, as described in the Methods section. To estimate the coupling, these additive entropies were then compared to those calculated using a 2D histogram for the same two angles (described in detail in the Supplementary Information, Table S7).

Although the coupling between angles could have a notable effect for an individual protein, *e.g.* at 298 K, $T\Delta S$ –0.6 kcal mol^{-1} for $\chi1/\chi2$ of Asn44 in the scorpion neurotoxin protein (Figure S2), in terms of general entropic propensities lower estimates were obtained, *e.g.* –0.1 kcal mol^{-1} for the coupling $\chi1/\chi2$ in Asn residues. Across the protein test set the coupling at most amounted to –0.4 kcal mol^{-1} (Pro $\chi3/\chi4$) with 72% of the estimated values falling below the mean value of 0.05 kcal mol^{-1}. The greatest coupling effects were between the backbone ϕ/ψ angles, with an average of 0.31 cal mol^{-1} K^{-1} ($T\Delta S$ –0.09 kcal mol^{-1}), and between certain side chain angles; for example, on average the coupling between $\chi1/\chi2$ was –0.13 kcal mol^{-1} and it was strongest for Leu and Pro and (Table S7). Considering the complexity of the interactions between side chain angles, the inequivalence of the degrees of freedom for different amino acids, and the significantly lower coupling observed for side chain dihedral angles than the backbone angles, coupling effects were not included in our calculations of side chain entropies (Tables 1 and S7). Hence, each side chain was treated as a set of independent χ-angles and the total entropy for a side chain, S_{sc}, was calculated as a sum of the individual angle entropies where microstate populations had been determined using one-dimensional histograms with 5° bin widths (see Methods).

However, as the coupling between backbone angles could be marginally larger and exhibit sensitivity to secondary structure, incorporated coupling into calculations of main chain en-

tropy. Although many studies only consider the ϕ/ψ angles to estimate the backbone conformational entropy, the dihedral angle for the more rigid peptide bond, ω, has a contribution of 3.3 cal mol^{-1} K^{-1} $residue^{-1}$ (D807) to 4.5 cal mol^{-1} K^{-1} $residue^{-1}$ (GGXGG), which is equivalent to some of the side chain contributions of more restricted angles (Table 1). Although S_ω was relatively invariant to the amino acid identity when averaged over the D807 dataset, is was incorporated into the descriptor for the main chain, S_{mc}, due to its sensitivity to the ψ angle[70] and due to some deviations >30° from planarity observed in our MD simulations and in the PDB.[71] For this reason, 3-dimensional histograms were used for computing main chain entropies, S_{mc}. The total entropy of an amino acid residue, S_{res}, was then obtained from a sum of the S_{mc} and S_{sc} values for each amino acid (Table 1). Coupling effects between angles become less pronounced with the size of the data set (Table S7), and, for the purposes of providing a dictionary of amino acid entropic propensities across protein fold space, our approach to coupling should be valid. While it should be kept in mind that in certain instances there can be protein-specific and site-specific instances of notable differences in entropy when neglecting the coupling between angles (Figure S2), the error from neglecting dihedral angle correlations is on the whole small, as has also been determined previously.[21,67-68]

Main Chain and Side Chain Contributions to Residue Entropy. The contribution of a single residue to the total conformational entropy of a protein is heavily influenced by the side chain contribution (Table 1). In the D807 native state (298 K) dataset, the main chain contributes S_{mc} = 12.2 – 14.0 cal mol^{-1} K^{-1} towards the total entropy, S_{res}, of an amino acid, which equates to relative contributions between 26% (Arg) to 73% (disulfide-bonded Cys). S_{mc} obtained from the intrinsic (GGXGG) dataset and denatured 498 K dataset were greater than those obtained from the

native state simulations, reflecting the greater freedom of the backbone in the Gly-based peptides and unfolded states. For many residues, the difference between the GGXGG and 498 K datasets was marginal, confirming that the GGXGG simulations can provide a good approximation of denatured state entropy. In all cases the main chain entropy was relatively invariant when considered against the range of side chain entropies determined (Figure 1a).

An interesting aspect of side chain entropy is the variation in dihedral fluctuations along a side chain. Trbovic *et al.* have demonstrated dynamic decoupling of the terminal bond vector motions from the rest of the side chain for the RNase H protein.[22] For residues with longer side chains, such decoupling can allow misinterpretation of the side chain conformational state from experimental data. In the case of Arg, salt bridges confer similar experimental order parameters regardless of whether the rest of the side chain is disordered or not.[22] They found this condition was shared with other amino acids (Lys, Glu, Gln, and Met) and hypothesized that this behavior should hold for amino acids with side chains containing more than two dihedral angles. Our inspection of side chain entropy showed wider distributions on increasing side chain length or bulk. For longer side chains, this variation in side chain entropy was independent of the main chain entropies, most notably for Lys and Arg due to their greater degrees of freedom (Figure 1b). Generally, the range of side chain entropies narrowed as the main chain entropy increased. From this we inferred that not only was there a relative decoupling between the main chain and side chain entropy, but that the longer side chains sampled alternative states regardless of the dynamics of the main chain.

Our findings regarding the variation in entropy along a side chain follow the expected pattern where, the average entropy of a dihedral angle decreased with proximity to the backbone,

the largest entropies being for the terminal dihedral angles (Figure 2a).[72] The side chain angles closest to the main chain can be restricted by the adherence to the ordered structure in the native state, whereas those angles furthest from the main chain can add a greater contribution to the total entropy. Within the proximity of the main chain, the entropies of the dihedral angles, particularly the χ_1 angles, are similar between residue types, with the exceptions of Ala and Pro (Figure 2b). The invariance in the χ_1 angles could be a reflection of the similar steric restraints experienced close to the main chain regardless of amino acid identity. By contrast, the entropies of the χ_2 angles show more variation between residues types (Figure 2b). The increase in entropy moving out along the side chain was most pronounced in the longer, aliphatic side chains, such as Ile (Figure 2c). However, this was not always the case. Within Arg the entropy of the side-chain angles initially increased on moving away from the main chain but then decreased for the terminal angles, χ_{61} and χ_{62} (Figure 2c). This is similar to what was seen by Trbvoic *et al.*,[22] possibly a result of the guanidinium group being involved in salt bridges. However, lysine is also commonly found in salt bridges and yet there is a distinct increase in the average entropy of the terminal χ_5 angle. This difference might reflect the fact that the guanidinium group is planar and exhibits charge distribution, while the lysine amine is tetrahedral, possibly making the former relatively more restricted in its mobility upon interaction with other molecular moieties.

The total entropy per residue, S_{res}, was calculated for the 807 proteins at 298 K and was found to increase in direct proportion to the side chain length (Table 1, Figure 1). In calculating the total conformational entropy of an entire protein, we find a linear relationship between the entropy and the number of residues, N_{res}, with correlations above 0.98 (Figure 3a); this conclusion holds also for the subtotals contributed by the main chain, S_{mc}, or side chain entropies, S_{sc}.

This result is consistent with the previous argument of Karplus *et al.*[10] that the residual vibrational entropy of a folded globular protein is an extensive property and that proteins are large enough that the residual conformational entropy is similar or equal to $n\langle S_{res}\rangle$, with $\langle S_{res}\rangle$ being the average configurational entropy over all amino acids in a single conformational state and *n (or* N_{res}*)* being the number of residues for that protein. However, the spread of entropy values for a given protein size is significant, ~10–20% of S_{res}, indicating that the variation in conformational entropy with sequence and structure leaves ample room for evolution to tap into conformational entropy as a contribution to protein stability and function, and in these cases, distinctions may be determined by the number χ angles, N_{chi}.

Assessment of total protein entropies for the Dynameomics dataset demonstrated that entropies were distributed relatively evenly across protein fold space. Proteomic Ramachandran plots (PRplot)[73] position protein structures in ϕ/ψ space based on the average of all main chain angles, thereby allowing properties, such as conformational entropy, to be surveyed across protein fold space (Figure 3b). Viewing the total entropies in this manner indicated that there are no distinct trends in the entropy of a protein and its predominant secondary structure. This behavior was also seen for calculated entropy changes on folding (ΔS_{fold}) with a similar relationship with sequence length and no clear trend in protein fold space, although there was some concentration of structures in the lower tail of the PRplot (all-α region) with larger ΔS (Figures 3c and 3d).

Comparison of Predicted Entropy Changes with Experiment. The absolute entropy values listed in Table 1 comprise our 'dictionary' values, however they are physically meaningless on their own as these entropies can vary depending on the bin widths used (Tables S2,S3,S6). How-

ever, the relative entropies and entropy changes upon folding/unfolding or mutation calculated using these dictionary values should be valid, being independent of bin width (Tables S2,S3,S6), and can be compared with experimental data. The variation in ΔS_{fold} for each of the 20 amino acids as determined across the D807 dataset are listed in Table 2. In the case of Cys residues, the entropy changes due to the removal of the disulfide bonds at high temperature are shown; the Cys ΔS_{fold} values are more than double that seen for free Cys (termed Cyh here) residues (Table 2).

Protein ΔS_{fold} values calculated from the dihedral distributions ranged from -235.5 (for a 35 residue protein) to -3367.7 cal mol^{-1} K^{-1} (for 410 residues), the magnitudes being 15–37% of the total protein entropy. The ΔS_{fold} ranged from -4.6 to -10.5 cal mol^{-1} K^{-1} $residue^{-1}$ (mean ΔS_{fold}, -7.8 ± 0.8 cal mol^{-1} K^{-1} $residue^{-1}$), corresponding to folding free energies of 1.4 – 3.1 kcal mol^{-1} $residue^{-1}$ (298 K). These data are consistent with initial calculations by Karplus *et al.* using BPTI,[10] showing that the ΔS_{fold} of a protein can be an order of magnitude smaller than the conformational entropy in the folded state with a range of 4 to 6 cal mol^{-1} K^{-1} $residue^{-1}$.

Experimental estimates of the entropy change on folding, ΔS_{fold}, give a comparable range of -2.6 to -9 cal mol^{-1} K^{-1} per residue.[12,74,75] The lower value of -2.6 cal mol^{-1} K^{-1} $residue^{-1}$ was determined for ps dynamics from neutron spectroscopy experiments of the thermal unfolding of α-amylase[75] and, as expected, is smaller than our lower limit of -4.6 cal mol^{-1} K^{-1} $residue^{-1}$ due to our sampling of ns dynamics. The higher end of this range was the result of entropies calculated by subtraction of estimates of the solvation entropy from the total entropy.[74] Calorimetric studies at high temperatures where there is no hydration contribution to the entropy include barnase,

with a ΔS_{fold} of -4.3 cal mol^{-1} K^{-1} residue^{-1}, and ubiquitin, 3.3 cal mol^{-1} K^{-1} residue^{-1}.[76] D'Aquino *et al.*[12] estimated an appropriate value for the entropic contribution of a residue to be 1.8 kcal mol^{-1} at 298 K (~6.0 cal mol^{-1} K^{-1} residue^{-1}) in determining the backbone conformational entropy for mutants of a leucine zipper peptide consistent with entropies ($T\Delta S_{fold}$ 0.4 – 1.3 kcal mol^{-1} residue^{-1}) determined from backbone dynamics using order parameters by Yang and Kay.[15] These experimental estimates are in close agreement with our $T\Delta S_{res}$ average of 2.4 ± 0.6 kcal mol^{-1} residue^{-1} at 298 K, with an average contribution from the backbone of 1.3 ± 0.3 kcal mol^{-1} residue^{-1} ($T\Delta S_{mc}$).

Similarly, when examining secondary structure formation, experimental estimates for the folding/unfolding of a Gly-based α-helix have ranged between 4.5 – 6.5 cal mol^{-1} K^{-1} residue^{-1}, with a wider 2 – 6.5 cal mol^{-1} K^{-1} residue^{-1} range for ΔS_{fold} for Gly in any secondary structure.[16] The same range has also been applied to sheet and turn formation.[66] As Figure 4 and Table S8 illustrate, we also observed a similar pattern of entropy differences across secondary structure types with a lower mean entropy when a residue is in helical structure. Naturally, where there was a lack of structure the average entropies were greater than when secondary structure was present (Figure 4, Table S8). In close agreement with experimental determinations, if we consider the value of Gly from D807 in the 'unassigned' structure category, pertaining to random coil and loop regions, we get a smaller entropy change of 3.1 cal mol^{-1} K^{-1} residue^{-1} on folding to an α-helix. However, if we use the average entropy for a Gly residue, regardless of the secondary structure assignment in the folded state, we estimate ΔS_{fold} = 4.6 cal mol^{-1} K^{-1} residue^{-1}. Again,

these estimates derived from our Dynameomics dataset fall into the experimentally derived ΔS_{fold} range of 2.0 – 6.5 cal mol^{-1} K^{-1} $residue^{-1}$ for Gly.[66]

Our dictionary values are also useful for predicting entropy changes on mutation. For example, D'Aquino *et al.*[12] indirectly estimated from experimental mutagenesis studies on a segment of GCN4 that the difference in entropy between Ala and Gly was 2.46 ± 0.2 cal mol^{-1} K^{-1} in a solvent-exposed position, a difference attributed solely to the lowering of entropy due to restriction of the backbone by the methyl side chain of Ala. We similarly observed a reduction in backbone entropy by 1.5 ± 2.3 cal mol^{-1} K^{-1} (D807 298K) to 1.1 cal mol^{-1} K^{-1} (GGXGG). However, when the Ala side chain was considered, an increase of 5.5 ± 2.4 cal mol^{-1} K^{-1} for the $\Delta S_{Gly \rightarrow Ala}$ was predicted (Table 1). Using intrinsic entropies from our GGXGG dataset to take into account the solvent-exposure in the D'Aquino study, the $\Delta S_{Gly \rightarrow Ala}$ increased to 6.1 cal mol^{-1} K^{-1} illustrating greater sampling by the exposed Ala side chain (Table 1). The calculation of a value for the change in main chain entropy lower than that obtained experimentally is to be expected, the experimental study of mutagenesis not being able to unequivocally assign the change in entropy to the backbone alone.[12] Moreover, an entropy difference of 1.5 to 2.0 cal mol^{-1} K^{-1} $residue^{-1}$ has been observed between the Gly and Ala backbone,[77-79] supported by an estimate of 2.5 cal mol^{-1} K^{-1} from experiment. Again, this is close to the difference of ~1.5 cal mol^{-1} K^{-1} between Gly and Ala computed here.

Low Entropy Depressions, α-Helices and Residual Structure. The mean main chain entropies calculated for each amino acid type were insensitive to the amino acid identity and tightly distributed between 12.2 to 14.0 cal mol^{-1} K^{-1} (Figure 1). However, the distribution of main chain entropies calculated for individual residues across our D807 dataset covered a wider

8.6 cal mol^{-1} K^{-1} range, from 9.4 to 18.1 cal mol^{-1} K^{-1}. At the same time, we have shown that within the folded, native state a significant amount of entropy is present in the side chain angles furthest from the main chain (Figure 2). The extent of this contribution will likely be affected by the local environment and hydrogen bonding networks for a given residue. Again, examination of the entropy dependence on secondary structure type showed variations in the average conformational entropy of the amino acids to be smaller than anticipated with the entropy of residues in α-helices lower than their corollary in β-sheets (Figure 4). Given the intrinsic stabilities of α-helix and β-sheet structures, with β-sheets often less stable due to the necessity for interactions between adjacent strands rather than the more localized hydrogen bonding in α-helices,[80] larger differences had been expected. The difference is marginal also when the dataset is split into the different topologies (Figure 5; Figure S3). However, the ranges of entropy values are wider and more variable between residue types than the mean values, indicating that the pertinent information might be buried in the finer details of our D807 dataset. As Figure 5 shows, there is greater variability among residues in the α and β structures than indicated by the average and standard deviation over the dataset (Figure 1, Table 1). Hence, the averages calculated are useful for quantifying the relative amino acid entropies in globular proteins, but they mask the vast variability in the dataset.

An ideal case to test the relationship between secondary structure and entropy is a pair of heteromorphic proteins, G_A88 and G_B88, that share high sequence identity (88%) but adopt different folds.[33] Although only differing by eight residues, G_A88 forms an all-α fold and G_B88 a β grasp fold (Figure 6). The differences between the G_A88 and G_B88 main chain entropies correlate most strongly with differences in secondary structure (Figure 6a), particularly in the regions 9–

22 and 39–47, despite the N-terminal portion of the sequence being identical up to residue 24 and there being only 1 substitution in the latter region. The mid-section of the sequence in which both structures contain helical structure shows the smallest difference between the main chain entropies (Figure 6). In fact, only one (L/Y45) of the eight non-identical residues displayed a large difference between the main chain entropies, indicating that secondary structure has a greater effect on main chain entropy than residue identity. When viewed for an individual protein, there is discrimination between the entropies of different structural elements. This discrimination was less pronounced when incorporating side chain entropies (Figure 6), but individual conserved side chains show sizeable differences that most likely reflect changes in the tertiary fold. The residue entropies (S_{mc} + S_{sc}) have the largest difference at residue 30, due to a large difference between the side chains entropies, which is a non-identical residue (I/F30). Strangely, a very similar pair at residue 45 (L/Y45) does not show a change in the total S_{res}, but this results from compensation between main chain and side chain entropies, where ΔS_{mc} is large due to S_{mc} being lower in α3 of G_A88, yet this is offset by the higher S_{sc}, compared with G_B88, resulting in the S_{res} of residue 45 being the same for the two proteins. Overall, when side chain entropies were included, the directionality of the differences observed at the backbone level was retained, showing the relationship between secondary structure and entropy is small but detectable.

There are fewer differences between G_A88 and G_B88 in the denatured state entropy profiles (as shown mapped onto the structures in Figure 6c), with 8 residues displaying discernable differences observed. Interestingly, 7 of the 8 positions (residues 24, 25, 30, 33, 45, 49, 50) are where the sequences differ, with I/F30 showing greatest difference. The only residue that is not conserved between the sequences and does not show an appreciable difference in entropy is E48,

although its neighboring residue D47 did show an entropy difference and was the only conserved residue that did. Thus, once native structure has been abolished, the sequence entropies become comparable and have simple dependence on the degrees of freedom of the side chains.

The fine structure was explored by plotting the difference between the entropy values from the G_A88 and G_B88 simulations at 298 K (S_{298}) and average value at 298 K over the full 807 proteins (S_{dict}) from the DED (Figure 6c). This produces a profile of peaks and troughs, showing how individual segments of secondary structure and loops in the G_A and G_B proteins differ from the average values across fold space. The troughs correspond to α-helix in both proteins, as well as β1 in G_B88. The other β-sheet regions are higher in entropy (Figure 6c). Interestingly, these low entropy sinks also correlate with regions structured in the transition states of both proteins (Figure 6). This includes the β1 strand that, unlike the other β stands, forms a low entropy depression compared with our D807 dictionary values and forms helical structure in the transition state. Consequently, depressions may not just relate to stable structure, like α-helices, but also regions where nucleation of folding occurs, which is also illustrated through the lower entropy regions mapped onto representative denatured state structures in Figure 6c.

The residues where forming an α-helix results in the largest entropic cost are Met and Pro, and to a lesser extent Lys. Residues where there is an increase in entropy upon forming α-helix are Phe, Leu, Ser, Trp, His. In fact, $S_{298}–S_{dict}$, reveals a correlation between helical regions and low entropy depressions (Figure 6c).

Entropy in the context of protein fold space. Given the differences in the entropy profiles of two domains with 88% sequence identity, a question arises whether entropy profiles are

conserved across metafolds where the structural topology is shared but the sequence identity varies. Metafolds within our consensus domain dictionary are ranked by population, the most populated metafold being the immunoglobulin-like β-sandwich (Rank 1). A contrast between entropy conservation in different topologies was drawn using this Rank 1 metafold family and an all-α metafold, the DNA/RNA-binding 3-helical bundles (Rank 5). Examination of 18 members of the Rank 1 (7 members) and Rank 5 (11 members) metafolds (listed in Table S8) demonstrated similar patterns in the entropy profiles to those observed for the G_A88/G_B88 pair (Figure 7). Depressions below the average entropic propensity from the Dynameomics 298 K dictionary are again observed for α-helices. One outlier is the MATA-1 homeodomain (1f43), which does not exhibit the low entropy depression (Figure S4). Interestingly, this MATA-1 domain has a higher number of homorepeats in its sequence than the other rank 5 members, indicating some regions may have a tendency towards disorder, which is the consensus from disorder prediction programs (MobiDB, Uniprot: POCY10).

The packing interfaces between the helices of the rank 5 bundles in particular appear to harbor the lowest entropy within the folds, the termini of helices and chain termini having higher entropies. Entropy, however, is not necessarily always lowest in the hydrophobic core, as evidenced by the 1wit and 1fna β-sandwich folds, which display generally higher entropies across all β-strands. Figure 7 includes a non-canonical member, 2efj, of the Rank 1 family, which contains helices and loops in addition to the core immunoglobulin domain. In this larger structure, the helices, again, have low entropy depressions but so do some of the β-strands. This leads us to consider another question of entropy compensation within protein structures. Based on the extensive relationship that appears to exist between the conformational entropy and number of resi-

dues (Figure 3), it appears that, for globular proteins, high entropy regions are offset by lower entropy segments. This may explain why more low entropy depressions are observed as the β-sandwich folds grow larger in size, although this is not always the case, *e.g.* 1ahm (Figure S4). Nonetheless, our results suggest a limiting sequence length at which low entropy depressions become necessary to maintain the fold (Figure S4).

Conclusions

We have generated a 'dictionary' of entropy values, the DED, which can be used in computing ΔS of folding/unfolding and mutation. The relative entropies computed using these dictionary values are in agreement with experiment. Unlike previous studies of protein conformational entropy, these data were obtained from dihedral angle distributions captured over MD simulations of 807 representative proteins that span nearly all known protein folds. By using all-atom solvated MD simulations, we have also been able to include all degrees of freedom. In performing this study, we have shown the total conformational entropy of a protein is an extensive property that can be estimated based on the number of residues alone. However, when examined on a per residue basis there is also fine structure and a relationship between secondary structure and the entropy of a given amino acid. These details are lost when creating an entropy dictionary, but by determining the intrinsic entropies using the ideal GGXGG model and the average entropic propensities of the amino acids at 298 K and 498 K we have been able to tease out fine structure reflected in the deviations from the dictionary values. We found that α-helices frequently coincide with low entropy depressions and that the entropy of a helix under native state conditions can be an indication of how prevalent it will be in transition states, with higher entropy hel-

ices forming later and that β-strands are generally higher in entropy than the Dynameomics dictionary values for a given residue.

AUTHOR CONTRIBUTIONS

C-L. T. performed the research and analysis. M. A. and V. D. devised the research. All authors contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

SUPPLEMENTARY INFORMATION

4 figures and 8 tables are contained in the supporting material.

ACKNOWLEDGMENTS

We are grateful for support from the National Institutes of Health (GM 50789 to V. D.) and the Swedish Research Council (621-2010-4912 to M.A.) and the Knut and Alice Wallenberg Foundation (2013.0022 to M.A.). The MD trajectories contained in the data warehouse were produced using computer time through the DOE Office of Biological Research as provided by the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U. S. Department of Energy under Contract No. DE-AC02-05CH11231. All figures of proteins were produced using UCSF Chimera.[81]

References

(1) Akke, M.; Brüschweiler, R.; Palmer, A. G. NMR Order Parameters and Free Energy: An Analytical Approach and its Application to Calbindin D9k. *J. Am. Chem. Soc.* **1993**, *115*, 9832-9833.

(2) Bracken, C.; Carr, P. A.; Cavanagh, J.; Palmer, A. G. Temperature dependence of intramolecular dynamics of the basic leucine zipper of GCN4: implications for the entropy of association with DNA. *J. Mol. Biol.* **1999**, *285*, 2133-2146.

(3) Zidek, L.; Novotny, M. V.; Stone, M. J. Increased Protein Backbone Conformational Entropy upon Hydrophobic Ligand Binding. *Nat. Struct. Biol.* **1999**, *6*, 1118-1121.

(4) Lee, A. L.; Kinnear, S. A.; Wand, A. J. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nature Struct. Biol.* **2000**, *7*, 72-77.

(5) Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **2007**, *448*, 325-329.

(6) Diehl, C.; Engström, O.; Delaine, T.; Håkansson, M.; Genheden, S.; Modig, K.; Leffler, H.; Ryde, U.; Nilsson, U. J.; Akke, M. Protein Flexibility and Conformational Entropy in Ligand Design Targeting the Carbohydrate Recognition Domain of Galectin-3. *J. Am. Chem. Soc.* **2010**, *132*, 14577-14589.

(7) Tzeng, S.-R.; Kalodimos, C. G. Dynamic activation of an allosteric regulatory protein. *Nature* **2009**, *462*, 368-372.

(8) Tzeng, S. R.; Kalodimos, C. G. Protein Activity Regulation by Conformational Esntropy. *Nature* **2012**, *488*, 236-240.

(9) Baron, R.; McCammon, J. A. Molecular Recognition and Ligand Association. *Ann. Rev. Phys. Chem.,* **2013**, *64*, 151-175.

(10) Karplus, M.; Ichiye, T.; Pettitt, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083-1085.

(11) Karplus, M.; Kushick, J. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14*, 325-332.

(12) D'Aquino, J. A.; Gómez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E. The Magnitude of the Backbone Conformational Entropy Change in Protein Folding. *Proteins* **1996**, *25*, 143-156.

(13) Stites, W. E.; Pranata, J. Empirical Evaluation of the Influence of Side Chains on the Conformational Entropy of the Polypeptide Backbone. *Proteins: Struct, Funct, Genetics* **1995**, *22*, 132-140.

(14) Mine, S.; Ueda, T.; Hashimoto, Y.; Imoto, T. Analysis of the Internal Motion of Free and Ligand-Bound Human Lysozyme by Use of NMR Relaxation Measurement: A Comparison with Those of Hen Lysozyme. *Prot. Sci*. **2000,** *9*, 1669-1684.

(15) Yang, D.; Kay, L. E. Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-Derived Order Parameters: Application to Protein Folding. *J. Mol. Biol.* **1996**, *263*, 369-382.

(16) Li, Z.; Raychaudhuri, S.; Wand, A. J. Insights into the Local Residual Entropy of Proteins Provided by NMR Relaxation. *Protein Sci.* **1996**, *5*, 2647-2650.

(17) Best, R. B.; Clarke, J.; Karplus, M. The Origin of Protein Sidechain Order Parameter Distributions. *J. Am. Chem. Soc.* **2004**, *126*, 7734-7735.

(18) Scouras, A. D.; Daggett, V. The Dynameomics Rotamer Library: Amino Acid Side Chain Conformations and Dynamics from Comprehensive Molecular Dynamics Simulations in Water. *Protein Sci* **2011**, *20*, 341-352.
(19) Lee, A. L.; Wand, A. J. Microscopic Origins of Entropy, Heat Capacity and the Glass Transition in Proteins. *Nature* **2001**, *411*, 501-504.
(20) Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. Conformational Entropy Changes upon Lactose Binding to the Carbohydrate Recognition Domain of Galectin-3. *J. Biomol. NMR* **2009**, *45*, 157-169.
(21) Li, D.W.; Brüschweiler, R. A Dictionary for Protein Side-Chain Entropies from NMR Order Parameters. *J. Am. Chem. Soc.* **2009**, *131*, 7226-7227.
(22) Trbovic, N.; Cho, J.-H.; Abel, R.; Friesner, R. A.; Rance, M.; Palmer, A. G., III. Protein Side-Chain Dynamics and Residual Conformational Entropy. *J. Am. Chem. Soc.* **2009**, *131*, 615-622.
(23) Genheden, S.; Akke, M.; Ryde, U. Conformational Entropies and Order Parameters: Convergence, Reproducibility, and Transferability. *J. Chem. Theory Comput.* **2014**, *10,* 432-438.
(24) Kasinath, V.; Sharp, K. A.; Wand, A. J. Microscopic Insights into the NMR Relaxation Based Protein Conformational Entropy Meter. *J. Am. Chem. Soc.* **2013**, *135*, 15092-15100.
(25) Sharp, K. A.; O'Brien, E.; Kasinath, V.; Wand, A. J. On the Relationship Between NMR- Derived Amide Order Parameters and Protein Backbone Entropy Changes. *Proteins: Struct., Funct., Bioinf.* **2015**, *83*, 922-930.
(26) Fenley, A. T.; Muddana, H. S.; Gilson, M. K. Entropy-Enthalpy Transduction Caused by Conformational Shifts can Obscure the Forces Driving Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 20006-20011.
(27) Forrey, C.; Douglas, J. F.; Gilson, M. K. The Fundamental Role of Flexibility on the Strength of Molecular Binding. *Soft Matter* **2012**, *8*, 6385-6392.
(28) Beck, D. A. C.; Jonsson, A. L.; Schaeffer, R. D.; Scott, K. A.; Day, R.; Toofanny, R.; Alonso, D. O. V.; Daggett, V. Dynameomics: Mass Annotation of Protein Dynamics and Unfolding in Water by High-Throughput Atomistic Molecular Dynamics Simulations. *Protein Eng. Des. Sel.* **2008**, *21*, 353-368.
(29) van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S. *et al.* Dynameomics: A Comprehensive Database of Protein Dynamics. *Structure* **2010**, *18*, 423-435.
(30) Day, R.; Beck, D. A. C.; Armen, R. S.; Daggett, V. Consensus View of Fold Space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* **2003**, *12*, 2150-2160.
(31) Schaeffer, R. D.; Jonsson, A. L.; Simms, A. M.; Daggett, V. Generation of a Consensus Protein Domain Dictionary. *Bioinformatics* **2011**, *27*, 46-54.
(32) Beck, D. A. C.; Alonso, D. O. V.; Inoyama, D.; Daggett, V. The Intrinsic Conformational Propensities of the 20 Naturally Occurring Amino Acids and Reflection of these Propensities in Proteins. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 12259-12264.

(33) Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N. The Design and Characterization of Two Proteins with 88% Sequence Identity but Different Structure and Function. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11963-11968.

(34) Levitt, M.; Hirshberg, M.; Sharon, R.; Laidig, K. E.; Daggett, V. Calibration and Testing of a Water Model for Simulation of the Molecular Dynamics of Proteins and Nucleic Acids in Solution. *J. Phys. Chem. B* **1997**, *101*, 5051-5061.

(35) Beck, D. A. C.; Alonso, D. O. V.; Daggett, V. A Microscopic View of Peptide and Protein Solvation. *Biophys Chem* **2003**, *100*, 221-237.

(36) Beck, D. A. C.; McCully, M. E.; Alonso, D. O. V.; Daggett, V. *il*mm --- *in lucem* molecular mechanics, Computer Program, **2000-2017**, University of Washington, Seattle, USA.

(37) Levitt, M.; Hirshberg, M.; Sharon, R.; Daggett, V. Potential Energy Function and Parameters for Simulations of the Molecular Dynamics of Proteins and Nucleic Acids in Solution. *Comput. Phys. Commun.* **1995**, *91*, 215-231.

(38) Morrone, A.; McCully, M. E.; Bryan, P. N.; Brunori, M.; Daggett, V.; Gianni, S.; Traglini-Allocatelli, C. The Denatured State Dictates the Topology of Two Proteins with Almost Identical Sequence but Different Native Structure and Function. *J. Biol. Chem.* **2011**, *286* (5), 3863.

(39) Scott, K. A.; Alonso, D. O. V.; Sato, S.; Fersht, A. R.; Daggett, V. Conformational Entropy of Alanine versus Glycine in Protein Denatured States. *Proc. Natl. Acad. Sci. USA* **2007**, *104* (8), 2661.

(40) Daggett, V.; Levitt, M. Realistic Simulations of Native-Protein Dynamics in Solution and Beyond. *Ann. Rev. of Biophys. and Biomol. Struct.* **1993**, ***22*** 353-380.

(41) Beck, D.A.C.; Daggett, V. Methods for Molecular Dynamics Simulations of Protein Folding/Unfolding in Solution. *Methods* **2004**, *34*, 112-120.

(42) Levitt, M. The Birth of Computational Structural Biology. *Nature* **2001**, *8*, 392-393.

(43) Levitt, M. ENCAD --- Energy Calculations and Dynamics, Computer Program, Molecular Applications Group, Stanford University, Stanford, CA and Yeda, Rehovot, Israel.

(44) Beck, D.A.C.; Armen, R.S.; Daggett, V. Cutoff size need not strongly influence molecular dynamics results on solvated polypeptides. *Biochemistry* **2005**, *44*, 609-616.

(45) Li, A.; Daggett, V. Investigation of the Solution Structure of Chymotrypsin Inhibitor 2 using Molecular Dynamics: Comparison to X-ray Crystallographic and NMR Data. *Protein Engineering*, **1995**, *8*: 1117-1128.

(46) Wong, K.B.; Daggett, V. Barstar Has A Highly Dynamic Hydrophobic Core: Evidence From Molecular Dynamics Simulation And NMR Relaxation Data. *Biochemistry*, **1998,** *87*, 11182-11192.

(47) Li, A.; Daggett, V. Characterization of the Transition State of Protein Unfolding Using Molecular Dynamics: Chymotrypsin Inhibitor 2. *Proc. Natl. Acad. Sci. USA*, **1994** *91*, 10430-10434.

(48) Li, A.; Daggett, V. Identification and Characterization of the Unfolding Transition State of Chymotrypsin Inhibitor 2 by Molecular Dynamics Simulations. *J. Mol. Biol.,* **1996**, *257*, 412-429.

(49) Daggett, V.; Li, A.; Itzhaki, L.S.; Otzen, D.E.; Fersht, A.R. Structure of the Transition State for Folding of a Protein Derived from Experiment and Simulation. *J. Mol. Biol.*, **1996,** *257*, 430-440.

(50) Alonso, D.O.V.; Alm, E.; Daggett, V. The Unfolding Pathway of the Cell Cycle Protein P13suc1: Implications for Domain Swapping**.** *Structure*, **2000**, *8*, 101-110.
(51) White, G.W.N.; Gianni, S.; Grossman, J.G.; Jemth, P.; Fersht, A.R.; Daggett, V. Simulation and Experiment Conspire to reveal Cryptic Intermediates and the Slide from the Nucleation-Condensation to Framework Mechanism of Folding. *J. Mol. Biol*., **2005**, *350*, 757-775.
(52) Jemth, P.; Gianni, S.; Day, R.; Li, B.; Johnson, C.M.; Daggett, V.; Fersht, A.R. Demonstration of a low energy on-pathway intermediate in a fast-folding protein by kinetics, protein engineering, and simulation. *Proc. Natl. Acad. Sci. USA*, **2004**, *101*, 6450-6455.
(53) Kazmirski, S.L.; Wong, K.B.; Freund, S.M.V.; Tan, Y.J.; Fersht, A.R.; Daggett, V. Protein Folding from a Highly Disordered Denatured State: The Folding Pathway of Chymotrypsin Inhibitor 2 at Atomic Resolution. *Proc. Natl. Acad. Sci. USA*, **2001**, *98*, 4349-4354.
(54) Bond, C.J.; Wong, K.; Clarke, J.; Fersht, A.R. Daggett, V. Characterization of Residual Structure in the Thermally Denatured State of Barnase by Simulation and Experiment: Description of the Folding Pathway. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 13409-13413.
(55) Wong, K.B.; Clarke, J.; Bond, C.J.; Neira, J.L.; Freund, S.M.V.; Fersht, A.R.F.; Daggett, V. Towards Complete Characterization of the Structural and Dynamic Properties of the Denatured State of Barnase and the Role of Residual Structure in Folding. *J. Mol. Biol*. **2000**, *296*, 1257-1282.
(56) Mayor, U.; Johnson, C.M.; Daggett , V.;. Fersht, A.R. Protein Folding and Unfolding in Microseconds to Nanoseconds by Experiment and Simulation. *Proc. Natl. Acad. Sci. USA* **2000**, *97***,** 13518-13522.
(57) Mayor, U., Johnson, C.M., Grossmann, J.G.; Sato, S.; Jas, G.S.; Freund, S.M.V.; Guydosh, N.R.; Alonso, D.O.V.; Daggett, V.; Fersht, A.R.F. *Nature* **2003**, *421*, 863-867.
(58) Gianni, S., Guydosh, N.R., Khan, F., Caldas, T.D., Mayor, U., White, G.W.N., DeMarco, M.L., Daggett, V.; Fersht, A.R. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 13286-13291.
(59) Religa, T.L.; Markson, J.S.; Mayor, U.; Freund, S.M.V.; Fersht, A.R. Solution structure of a protein denatured state and folding intermediate. *Nature* **2005**, *437*, 1053-1056.
(60) McCully, M.E.; Beck, D.A.C.; V. Daggett, V. Microscopic Reversibility of Protein Folding in Molecular Dynamics Simulations of the Engrailed Homeodomain. *Biochemistry* **2008**, *47*, 4079-7089.
(61) McCully, M.E.; Beck, D.A.C.; Fersht, A.R.; Daggett, V. Refolding Engrailed Homeodomain: Structural Basis for the Accumulation of a Folding Intermediate. *Biophys. J*. **2010**, *99*, 1628-1636.
(62) McCully M.E., Beck D.A.C., Daggett V. Multimolecule test-tube simulations of protein unfolding and aggregation. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17851-17856.
(63) Fersht**,** A.R.; Daggett, V. Protein Folding and Unfolding at Atomic Resolution. *Cell*, **2002**, *108*, 573-582.
(64) Daggett, V., Protein Folding --- Simulation, *Chemical Reviews*, **2006,** *106*, 1898-1916.
(65) Daggett, V.; Fersht, A.R. The Present View of the Mechanism of Protein Folding. *Nat. Rev. Mol. Cell. Biol*., **2003**, *4,* 497-502.
(66) Brady, G. P.; Sharp, K. A. Entropy in Protein Folding and in Protein-Protein Interactions. *Curr. Opin. Struct. Biol.* **1997**, *7*, 215-221.

(67) Li, D.W.; Meng, D.; Brüschweiler, R. Short-Range Coherence of Internal Protein Dynamics Revealed by High-Precision in Silico Study. *J. Am. Chem. Soc.* **2009**, *131*, 14610-14611.
(68) Li, D.W.; Showalter, S.A.; Brüschweiler, R. Entropy Localization in Proteins. *J. Phys. Chem. B* **2010**, *114*, 16036-16044.
(69) Simms, A. M.; Beck, D. A. C.; Daggett, V. Implementation of 3D Spatial Indexing and Compression in a Large-Scale Molecular Dynamics Simulation Database for Rapid Atomic Contact Detection. *BMC Bioinformatics* **2011**, *12*, 334.
(70) Esposito, L.; De Simone, A.; Zagari, A.; Vitagliano, L. Correlation Between Omega and Psi Dihedral Angles in Protein Structures. *J Mol. Biol.* **2005**, *347*, 483-487.
(71) Berkholz, D. S.; Driggers, C. M.; Shapovalov, M. V.; Dunbrack, R. L.; Karplus, P. A. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 449-453.
(72) Wittebort, R.J.; Szabo, A. Theory of NMR Relaxation in Macromolecules: Restricted Diffusion and Jump Models for Multiple Internal Rotations in Amino Acid Side Chains. *J. Chem. Phys.* **1978**, *69,1722-1736*.
(73) Carugo, O.; Djinović-Carugo, K. A Proteomic Ramachandran Plot (PRplot). *Amino Acids* **2013**, *44*, 781-790.
(74) Makhatadze, G. I.; Privalov, P. L. On the Entropy of Protein Folding. *Protein Sci* **1996**, *5*, 507-510.
(75) Fitter, J. A Measure of Conformational Entropy Change during Thermal Protein Unfolding Using Neutron Spectroscopy. *Biophys. J.* **2003**, *84*, 3924.
(76) Makhatadze, G. I.; Privalov, P. L. Energetics of Protein Structure. *Advances in Protein Chemistry*, **1995**, *47*, 307–425.
(77) Yang, A. S.; Honig, B. Free Energy Determinants of Secondary Structure Formation: I. α-Helices. *J Mol. Biol.* **1995**, *252*, 351-365.
(78) Yang, A. S.; Honig, B. Free Energy Determinants of Secondary Structure Formation: II. Anti-Parallel Beta-Sheets. *J Mol. Biol.* **1995**, *252*, 366-376.
(79) Yang, A. S.; Hitz, B.; Honig, B. Free Energy Determinants of Secondary Structure Formation: III. Beta-Turns and Their Role in Protein Folding. *J Mol. Biol.* **1996**, *259*, 873-882.
(80) Honig, B. Protein folding: From the Levinthal Paradox to Structure Prediction. *J Mol. Biol.* **1999**, *293*, 283-293.
(81) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605-1612.

Figure Legends

Figure 1. Comparison of the D807 and GGXGG backbone and side chain entropies. (a) Relative distributions of the GGXGG and D807 298 K entropies highlighting the substantially higher entropies obtained from the more extensive conformational sampling possible by the central guest residues in the GGXGG peptides. The shaded regions highlight the differences between the mean values from each of the GGXGG and D807 data sets. (b) Distribution of entropies determined for amino acids demonstrating larger spread of possible side chain entropies as the degrees of freedom increase for amino acids with longer side chains. This reflects the complexity that longer side chains can exhibit a wide range of entropies. The breadth of the distribution for the main chain entropies remains insensitive to the amino acid identity and length of side chain.

Figure 2. Variation in entropy along side chains with respect to distance from the main chain. (a) Average conformational side chain entropy of the side-chain angles. Proline has been excluded to remove the bias from including the substantially lower entropies obtained for the dihedral angles in the sterically constrained cyclic side chain. (b) Conformational entropy for the dihedral angles closest to the main chain determined by amino acid identity. For branched residues, χ_{21} and χ_{22} values are used due to positional equivalence to χ_2. (c) Variation in side chain entropy for the amino acids Arg, Ile and Lys demonstrating different behaviors related to the chain length and salt bridge affinity.

Figure 3. Relationship between entropy, protein size and protein topology. (a) Distribution of total residue entropies for the Dynameomics set of 807 proteins in terms of number of residues. (b) Distribution of total entropy in protein fold space (PRplot) (c) Distribution of ΔS_{fold} for the Dynameomics dataset in relation to number of residues, and (d) ΔS_{fold} across protein fold space in terms of the PRplot using the average ϕ/ψ angle to show spread of different protein topologies.

Figure 4. Relationship between entropy and secondary structure. Entropies calculated for different secondary structure types for population > 90% of a trajectory (legend inset). Entropies labeled "coil" are where no formal secondary structure could be assigned for > 90% of the trajectory. See also Table S7 for detailed information for each amino acid as a function of secondary structure.

Figure 5. Calculated, predicted and experimental estimates of entropy change on unfolding and relation to protein size. Box and whisker plots of calculated ΔS_{unfold} for Ala, Arg and Gly residues in different topology and protein classes showing the extremes of the data in addition to the mean (blue points), median (red bars) and the lower and upper quartiles: all-α, all-α metafold topology by SCOP classification; all-β, all-β metafold topology; HP, hyperthermophile; L, large (>=150) residues; M, medium-sized (50-150 residues); MP, mesophile; α/β, mixed-α/β metafold topology; S, small (<=50 residues); NS, non-standard topology, TP, thermophile. Shaded area indicates negative ΔS_{unfold} values. Figure S3 shows the corresponding results for all 20 amino acid types.

Figure 6. Entropy profiles for the G_A88 and G_B88 heteromorphic domains. (a) Sequences of the G_A88 and G_B88 domains showing 8 non-identical residues. (b) Main chain and total residue entropy differences at 298 K and 498 K for the G_A88 and G_B88 domains. (c) Entropy profiles of the difference between the dictionary values and G_A88 and G_B88 entropies illustrating pockets of low and high entropy that correlate with early secondary structure formation highlighted on native and transition state structures inset below.

Figure 7. Conservation of entropy profiles across CDD metafold members (S_{298}) showing positive (blue) and negative (red) deviations from the Dynameomics 298 K entropy dictionary (S_{dict}). (a) Entropy profiles for all-α folds from the DNA/RNA-binding 3-helical bundle metafold family (rank 5) showing entropy depressions due to helix presence. (b) Entropy profiles for the immunoglobulin-like β-sandwich metafold family (rank 1) illustrating conservation of higher entropy in β-strands in the canonical 1wit and 1fna folds with more frequent entropy depressions observed for the non-canonical immunoglobulin-like fold of 2fej. Structures inset are also colored by the entropy deviation from the average D807 298 K values. Figure S4 shows the corresponding results for all analyzed proteins in the rank 1 and rank 5 metafold families.

Table 1. The Dynameomics Entropy Dictionary. Contribution of dihedral angles to the residue conformational entropy (S_{res}) by amino acid identity calculated from dihedral angle distributions from simulations of the GGXGG series at 298 K and the 807 Dynameomics targets (D807) at 298 K and 498 K. Cysteine residues are present in the D807 dataset in disulfide bonded (Cys) and protonated forms (Cyh), but the short length of the GGXGG peptides precludes modeling of the disulfide bonded form of Cys, hence only entropies for the protonated form are given for the GGXGG host-guest series.

	GGXGG (cal.mol^{-1} .K^{-1})				D807, 298 K (cal.mol^{-1} .K^{-1})				D807, 498 K (cal.mol^{-1} .K^{-1})			
Res.	$S\phi/\psi$	S_{mc}	S_{sc}	S_{res}	$S\phi/\psi$	S_{mc}	S_{sc}	S_{res}	$S\phi/\psi$	S_{mc}	S_{sc}	S_{res}
Ala	13.8	17.1	7.2	24.2	9.4 ± 1.5	12.5 ± 1.7	7.1 ± 0.2	19.6 ± 1.7	14.0 ± 1.2	17.5 ± 1.2	7.7 ± 0.0	25.3 ± 1.2
Arg	13.2	16.5	40.7	57.2	9.3 ± 1.3	12.4 ± 1.5	36.1 ± 3.2	48.5 ± 3.8	13.3 ± 1.1	16.9 ± 1.1	44.2 ± 0.9	61.2 ± 1.6
Asn	12.2	15.5	18.2	33.7	9.6 ± 1.3	12.8 ± 1.4	17.0 ± 1.2	29.8 ± 2.1	13.5 ± 1.1	17.1 ± 1.1	20.5 ± 0.4	37.6 ± 1.3
Asp	12.3	15.7	11.4	27.1	9.6 ± 1.3	12.8 ± 1.4	10.7 ± 1.3	23.5 ± 2.2	13.2 ± 1.0	16.8 ± 1.0	13.4 ± 0.7	30.2 ± 1.4
Cys	-	-	-	-	9.8 ± 1.3	13.0 ± 1.4	4.8 ± 0.9	17.7 ± 2.0	-	-	-	-
Cyh	13.1	16.4	14.1	30.5	9.7 ± 1.4	12.8 ± 1.6	12.9 ± 0.9	25.7 ± 2.1	13.8 ± 1.0	17.4 ± 1.0	15.2 ± 0.2	32.5 ± 1.1
Gln	13.4	16.7	24.6	41.3	9.4 ± 1.4	12.5 ± 1.5	23.3 ± 1.4	35.8 ± 2.4	13.5 ± 1.1	17.1 ± 1.1	27.3 ± 0.3	44.5 ± 1.3
Glu	13.5	16.8	18.2	35.0	9.3 ± 1.3	12.5 ± 1.5	17.4 ± 1.8	29.9 ± 2.5	13.4 ± 1.0	17.0 ± 1.0	21.0 ± 0.8	38.0 ± 1.5
Gly	15.1	18.1	-	18.1	10.8 ± 1.5	14.0 ± 1.6	-	14.0 ± 1.6	15.2 ± 1.0	18.6 ± 1.0	-	18.6 ± 1.0
Hid	13.1	16.4	14.3	30.8	9.7 ± 1.4	12.9 ± 1.5	11.9 ± 1.7	24.8 ± 2.5	13.5 ± 1.1	17.1 ± 1.1	15.3 ± 0.2	32.4 ± 1.2
Hie	12.9	16.2	14.8	31.0	9.5 ± 1.3	12.7 ± 1.5	12.0 ± 1.7	24.6 ± 2.5	13.7 ± 1.0	17.3 ± 1.0	15.4 ± 0.2	32.6 ± 1.1
Ile	12.8	16.2	24.6	40.7	9.3 ± 1.2	12.4 ± 1.4	24.3 ± 1.3	36.7 ± 2.1	13.1 ± 1.0	16.7 ± 1.0	28.3 ± 0.3	45.0 ± 1.0
Leu	13.0	16.3	25.7	42.0	9.1 ± 1.3	12.2 ± 1.4	24.8 ± 1.4	37.0 ± 2.2	13.0 ± 1.2	16.6 ± 1.3	28.7 ± 0.2	45.3 ± 1.3
Lys	13.0	16.3	32.7	49.0	9.4 ± 1.3	12.6 ± 1.4	30.1 ± 2.6	42.7 ± 3.1	13.2 ± 1.1	16.8 ± 1.1	35.8 ± 1.0	52.7 ± 1.7
Met	13.3	16.6	26.7	43.3	9.3 ± 1.3	12.4 ± 1.5	25.4 ± 1.3	37.8 ± 2.3	13.2 ± 1.2	16.8 ± 1.2	29.4 ± 0.2	46.2 ± 1.3
Phe	12.8	16.2	13.8	30.0	9.5 ± 1.4	12.6 ± 1.5	11.2 ± 1.6	23.8 ± 2.4	13.3 ± 1.2	16.9 ± 1.2	14.9 ± 0.2	31.8 ± 1.4
Pro	10.5	14.8	20.5	35.3	8.8 ± 0.8	12.6 ± 1.0	19.9 ± 0.7	32.5 ± 1.5	11.4 ± 0.5	16.0 ± 0.6	23.0 ± 0.2	39.0 ± 0.7
Ser	13.6	16.9	13.7	30.6	9.8 ± 1.4	13.0 ± 1.5	12.0 ± 1.6	24.9 ± 2.7	13.4 ± 1.1	17.0 ± 1.1	14.7 ± 0.7	31.6 ± 1.6
Thr	12.6	16.0	20.4	36.3	9.6 ± 1.2	12.7 ± 1.4	18.4 ± 1.7	31.1 ± 2.7	12.6 ± 1.1	16.3 ± 1.1	21.4 ± 0.9	37.7 ± 1.8
Trp	13.0	16.3	14.2	30.5	9.3 ± 1.3	12.5 ± 1.5	10.0 ± 1.7	22.5 ± 2.6	13.3 ± 1.2	16.9 ± 1.2	14.9 ± 0.3	31.9 ± 1.4
Tyr	13.0	16.3	21.9	38.2	9.4 ± 1.3	12.6 ± 1.5	18.1 ± 2.2	30.7 ± 2.9	13.4 ± 1.1	17.0 ± 1.1	23.1 ± 0.3	40.1 ± 1.3
Val	12.9	16.3	20.2	36.5	9.4 ± 1.2	12.6 ± 1.4	19.1 ± 0.8	31.6 ± 1.9	13.2 ± 1.0	16.9 ± 1.0	21.8 ± 0.2	38.6 ± 1.0

*Cyh is the protonated version of Cys. Hie is the dominant tautomer of histidine with the Nε protonated and Hid has Nδ protonated.

Table 2. Mean conformational entropy change on unfolding, Native State (298K) → Denatured State (498K), calculated for individual amino acid residues in the D807 dataset and the predicted entropy change using the average 298 K and 498 K residue entropies for an amino acid type given in Table 1. Two entropy changes are given for cysteine residues, one for free, protonated cysteine residues (Cyh) and cystine residues (Cys) where the disulfide bond present in the native state simulation and note that the disulfide bonds were reduced in the high temperature simulations.

	ΔS_{mc} (cal.mol^{-1}.K^{-1})	ΔS_{res} (cal.mol^{-1}.K^{-1})	$T\Delta S_{mc}$ 298K (kcal.mol^{-1})	$T\Delta S_{res}$ 298K (kcal.mol^{-1})	ΔS_{res} (cal.mol^{-1}.K^{-1})
ALA	5.03 ± 1.72	5.63 ± 1.73	1.50	1.68	5.63
ARG	4.53 ± 1.49	12.65 ± 3.84	1.35	3.77	12.65
ASN	4.30 ± 1.50	7.82 ± 2.16	1.28	2.33	7.82
ASP	4.00 ± 1.46	6.68 ± 2.31	1.19	1.99	6.68
CYH	4.46 ± 1.52	6.76 ± 2.03	1.33	2.01	6.86
CYS*	4.50 ± 1.50	14.91 ± 2.03	1.34	4.44	14.79
GLN	4.59 ± 1.57	8.61 ± 2.45	1.37	2.57	8.61
GLU	4.58 ± 1.53	8.14 ± 2.67	1.36	2.43	8.14
GLY	4.56 ± 1.66	4.56 ± 1.66	1.36	1.36	4.56
HID*	4.27 ± 1.54	7.58 ± 2.60	1.27	2.26	7.59
HIE*	4.62 ± 1.57	7.98 ± 2.60	1.38	2.38	8.00
ILE	4.36 ± 1.42	8.35 ± 2.12	1.30	2.49	8.35
LEU	4.37 ± 1.54	8.33 ± 2.21	1.30	2.48	8.33
LYS	4.26 ± 1.49	9.97 ± 3.27	1.27	2.97	9.97
MET	4.42 ± 1.60	8.41 ± 2.32	1.32	2.51	8.42
PHE	4.34 ± 1.55	7.99 ± 2.53	1.29	2.38	7.99
PRO	3.42 ± 1.08	6.51 ± 1.58	1.02	1.94	6.51
SER	4.03 ± 1.58	6.70 ± 2.82	1.20	2.00	6.70
THR	3.61 ± 1.49	6.64 ± 2.97	1.07	1.98	6.63
TRP	4.46 ± 1.61	9.36 ± 2.66	1.33	2.79	9.36
TYR	4.41 ± 1.53	9.38 ± 2.99	1.31	2.80	9.39
VAL	4.29 ± 1.42	7.01 ± 1.91	1.28	2.09	7.01

*Cyh is the protonated version of Cys. Hie is the dominant tautomer of histidine with the Nε protonated and Hid has Nδ protonated.