Measurements of Protein Sequence-Structure Correlations

Gavin E. Crooks, Jason Wolfe, Steven E. Brenner
Department of Plant and Microbial Biology,
University of California, Berkeley, CA, USA 94720-3102

April 8, 2004

Keywords:
protein structure; contact potentials; mutual information; secondary structure; hydrophobicity
Running Title:
Protein Sequence-Structure Correlations
Corresponding Author:
Gavin E. Crooks,
Department of Plant and Microbial Biology,
111 Koshland Hall #3102,
University of California,
Berkeley, CA 94720-3102, USA.
Tel: +1 510-642-9614,
Fax: +1 208-279-8978,
Email: gec@compbio.berkeley.edu

ABSTRACT

Correlations between protein structures and amino acid sequences are widely used for protein structure prediction. For example, secondary structure predictors generally use correlations between a secondary structure sequence and corresponding primary structure sequence, whereas threading algorithms, and similar tertiary structure predictors, typically incorporate interresidue contact potentials. To investigate the relative importance of these interactions we measured the mutual information between the primary structure, secondary structure and side-chain surface exposure, both for adjacent residues along the amino acid sequence, and for tertiary structure contacts between residues distantly separated along the backbone. We find that local interactions along the amino acid chain are far more important than non-local contacts, and that correlations between proximate amino acids are essentially uninformative. This suggests that knowledge-based contact potentials may be less important for structure predication than is generally believed.

INTRODUCTION

The three-dimensional structure of a protein is determined by the protein's amino acid sequence via the process of protein folding. Conversely, the amino acid sequence is selected by evolution to fold into the required structure. At present, we are incapable of elucidating this protein sequence-structure mapping in detail. However, many coarse-grained structural features are directly correlated with the amino acid sequence. For example, hydrophobic valines tend to be buried in the protein core, alanines are over-represented in helixes, and cysteines are often located close in space so that they may form disulphide bonds.

Correlations between sequence and many different simplified structural features, such as burial, secondary structure and inter-residue proximity, have been extensively studied and used in a wide range of protein structure prediction algorithms and protein folding studies. Among these are secondary structure prediction[1,2], disorder prediction[3,4], various tertiary prediction methods[5], such as structure profiles [6,7] and threading[8,9,10], and simplified models of protein folding[11,12,13]. Unfortunately, it is often difficult to determine the relative importance of different interactions to the reported algorithmic performance. The selection and definition of potentials varies greatly, and the choices made are not always carefully cross-validated. Indeed, many of the more sophisticated methodologies are computationally expensive, rendering a detailed study or comprehensive comparison[14] problematic. Often-times the actual interactions are hidden beneath a neural network or other training layer, and very frequently structure-sequence interactions are intermingled with homology information, masking the relative effect of structure versus evolutionary history on the amino acid sequence.

The direct, quantitative measurements of sequence-structure correlations can elucidate the relative importance of different interactions to protein structure[15], and facilitate the rational design of structure prediction algorithms. For example, the frequency with which a particular amino acid is located in the center of proteins is directly related to the hydrophobicity of the amino acid side chain[16,17]. Recently, Cline et al.[18] investigated the strength of correlations between residues close in space but distant along the chain, and found them to be surprisingly weak. These correlations are the essential interaction encoded by contact potentials, which are commonly used by protein threading and allied structure prediction algorithms, and also in many idealized models of protein folding[12,13]. Similarly, we have recently examined the correlations between secondary structure and the local amino acid sequence, with the objective of better understanding secondary structure formation and secondary

structure prediction algorithms[2].

In this study, we examined the strength, organization and relative importance of correlations between amino acid identity, secondary structure, and fractional side chain burial, both for residues neighboring along the amino acid chain, and for residues proximate in space, but distantly located along the chain. These interactions are quantified using mutual information and similar measures. To ensure reliable results we employed a large, non-redundant and diverse collection of protein structures that contains 2,853 sequences. We find that local sequence-structure correlations along the chain are far stronger than tertiary structure interactions. Moreover, once local structural interactions have been accounted for, we find very little correlation between the side chain identity of spatially proximate residues.

MATERIALS AND METHODS

Correlations and Mutual Information

The correlation between two events can be quantified as the log likelihood ratio, the log ratio of observing the joint event $p(x, y)$, against the expected probability of observing the joint event if the events were uncorrelated $p(x)p(y)$;

$$\mathcal{L}(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}. \tag{1}$$

The base of the logarithm is arbitrary, but base 2 is conventional and convenient, resulting in bits as the unit of measurement.

These correlations represent effective interactions that can be used in prediction algorithms as an objective function or additive score. Alternatively, we might invoke a quasi-chemical approximation[12], and think of the log likelihood ratio as a negative interaction energy, or a free energy, in units of $k_B T \ln 2$, where T is the temperature, k_B is Boltzmann's constant and the factor $\ln 2 \approx 0.69$ converts between binary and natural logarithms. (Note that 1 bit $\approx$ 1.7 kJ mol^{-1}, or $\approx$ 0.4 kcal mol^{-1}, at ambient temperatures.)

The average strength of the effective interaction between two variables (the mean of the log odds score) is the mutual information, a measure of the knowledge that each variable carries about the other[19,20].

$$I(X;Y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

The mutual information is related to the entropy, $H(X)$.

$$I(X;Y) = H(X) + H(Y) - H(X, Y) \tag{3}$$

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{4}$$

A high mutual information is a result of strong correlations, whereas zero mutual information indicates uncorrelated variables. Mutual information has various advantages as a correlation measure: It is firmly grounded in information theory[19]; it is additive for independent contributions; and it has consistent, intuitive units (bits), which allow for quantitative comparisons between different features and problem domains. The main disadvantage of information measures is that they are difficult to calculate accurately, requiring large data sets and careful error analysis.

The interactions between three variables can be quantified by the the triplet mutual information, $I^{(3)}(X;Y;Z)$. This is the average information carried by the three way interactions, in excess of the information carried by the pairwise interactions[21,22].

$$\begin{aligned} I^{(3)}(X;Y;Z) &= -H(X) - H(Y) - H(Z) \\ &\quad + H(X, Y) + H(X, Z) + H(Y, Z) \\ &\quad - H(X, Y, Z) \end{aligned} \tag{5}$$

Alternatively, we can consider the conditional mutual information[20] $I(X;Y|Z)$, the information that X carries about Y, given Z.

$$\begin{aligned} I(X;Y|Z) &= -H(Z) + H(X, Z) + H(Y, Z) \\ &\quad - H(X, Y, Z) \\ &= I(X;Y) + I^{(3)}(X;Y;Z) \end{aligned} \tag{6}$$

The pairwise and conditional mutual informations are positive. The triplet mutual information can be positive or negative, and consequentially the conditional mutual information can be greater or smaller than the unconditioned mutual information.

Information Bias Correction

Estimating information from limited amounts of data leads to significant bias[23], resulting in a systematic underestimation of the entropy, or overestimation of the mutual information. We used bootstrap resampling[24] to correct for this bias, and to estimate standard statistical errors. We generated fifty replicas of the original data by sampling N sequences, with replacement, from the N available sequences. This resampling has associated systematic and random errors that are approximately the same as the errors introduced by the original finite random sampling of sequences from the true random distribution. These error estimates were not significantly improved when the number of replicas was increased from 50 to 500. The requisite pseudo-random numbers were drawn from the Mersenne Twister generator[25]. Note that bootstrap resampling can estimate errors caused by a finite amount of data, but it cannot correct for errors due to the inherent biases of our data set. Moreover, the bootstrap tends to underestimate the magnitude of the bias when provided with insufficient data. Our experience with limited sampling from known distributions suggests that whenever the estimated bias is comparable to, or greater than, the estimated standard error, the bootstrap results should be treated with skepticism.

Protein Structure Library

Our structure library is based upon a representative, high-quality and non-redundant subset of available protein structures. Although we have attempted to minimize sampling errors, this collection is still inevitably biased, since structures that are interesting and tractable are over-represented in the public data, and unordered protein regions are excluded from crystallographic data. The Protein Data Bank (PDB)[26] currently contains over 20,000 publicly accessible structures, but many of these are very similar, and many are of relatively low quality. The Structural Classification Of Proteins (SCOP)[27,28] database provides a convenient decomposition of PDB structures into 44,000 protein domains. The ASTRAL compendium[29] provides representative subsets of SCOP domains, filtered so that no two domains share more than a given percentage level of sequence identity. This filtering reduces the redundancy of the PDB while preferentially retaining higher quality structures, as judged by AEROSPACI scores[29], an agglomeration of several structure quality measures. We selected the ASTRAL 40% sequence identity subset of SCOP release 1.61, which was further filtered to remove multi-sequence domains, SCOP classes f (membrane and cell surface proteins) and g (small proteins), and to retain only those structures determined by X-ray diffraction at better than 2.5 Å resolution. The protein sequences were taken from the ASTRAL Rapid Access Format (RAF) sequence mappings[29], which provides a more reliable and convenient representation of the true sequence than the PDB ATOM or SEQRES records alone. The tertiary structure of each protein was simplified to just one coordinate for each residue, the location of the β-carbon. For glycine the location of an imaginary β-carbon was estimated from the backbone coordinates. The secondary structure and surface accessibility of each residue were determined using the program STRIDE[30]. STRIDE was unable to process a small number of SCOP domains, which were consequently removed from further consideration. The resulting structure template library contains 2,853 protein domains and 553,370 residues, and can be obtained from our website at `http://compbio.berkeley.edu/`.

Secondary Structure

Secondary structure is a concise description of a residue's backbone torsional angles and pattern of hydrogen bonding. STRIDE[30] assigns each residue to one of 8 classes; α-helix (H), 3_{10} helix (G), π-helix (I), β-strand (E), β-bridge (B or b), Coil (C, L, or space), Turn (T) or Bend (S). Unstructured or poorly resolved regions of the protein are unassigned (X). These 8 classes were reduced to the three letter alphabet, E (Extended strand), H (Helix), and L (Loop/Other) using the common mapping[2] E→E, H→H, all others→L. This substantial reduction of the state space only reduces the singlet amino acid to secondary structure mutual information by 0.008 bits.

	bits	std. err. (bias corr.)
$I(R;S)$ (8 states)	0.0893	±0.0007 (+0.00014)
$I(R;S)$ (3 states)	0.0812	±0.0007 (+0.00007)

Here, R is the amino acid identity of a single residue, and S the secondary structure assignment of that residue.

Burial/Surface Accessibility

STRIDE also calculates the exposed surface area of each residue. These areas are converted to percentages using the nominal maximum solvent accessibility of a residue in the tripeptide G-X-G[16]. We aggregate these surface exposures into a finite number of burial bins (B), and use dynamic programming to determine the exposure thresholds that maximize the mutual information $I(R;B)$ between amino acid identity and burial. The algorithm utilizes the fact that once a single partition has been chosen, the optimal partitionings of its left and right sides (into some specified numbers of bins) are independent of one-another. Thus, a simple recursive algorithm suffices (which can be memorized or 'unrolled' for speed), which finds the optimal n partitions by choosing the leftmost optimal partition, given the optimal $(n-1)$ partitions for the remaining space to the right. We settled upon 4 burial bins as a reasonable compromise between cardinality and information gain.

	bits	std. err. (bias corr.)
$I(R;B)$ (100 bins)	0.214	±0.001 (+0.0001)
$I(R;B)$ (8 bins)	0.205	±0.001 (+0.0002)
$I(R;B)$ (4 bins)	0.190	±0.001 (+0.0003)
$I(R;B)$ (2 bins)	0.150	±0.001 (+0.0001)

The optimal partitions are 0.0-1.5%, 1.5-17.5%, 17.5-40.0% and 40.0-100%. Due to algorithmic differences and post-translational chemical modification, a small number of residues are assigned an erroneously high surface accessibility of greater than 100%. These residues are conglomerated into the largest surface exposure bin.

Interresidue contacts

We define two residues to be in non-local contact if they are separated by at least 6 residues in the chain, and the distance between the side chain β-carbons is less than 8 Å. The number of contacts per residue ranges from 0 to 15, but more than a few contacts is uncommon. The average is 2.1 pairings per residue. We are reluctant to heavily optimize our definition of inter-residue contact, or to employ finely partitioned contact distances, lest we inadvertently directly encode information about the side chain volume or amino acid identity into the contact potential, in addition to generic properties of the local structural environment.

Summary

To recapitulate, our data set consists of 2,853 diverse protein domains containing more than one half million residues.

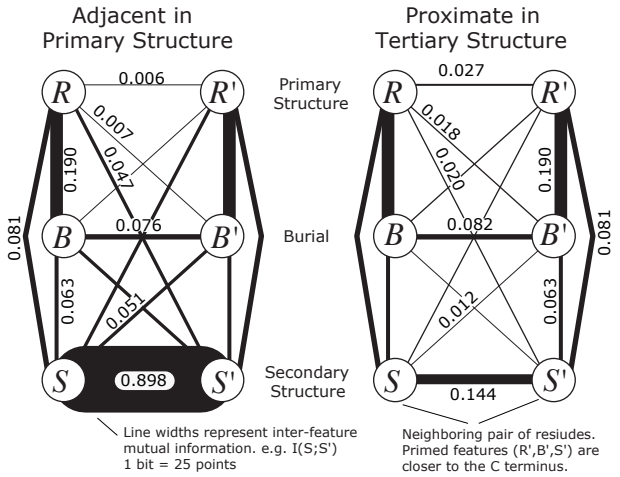


Figure 1: Summary of inter- and intra- residue mutual information. Correlations between neighboring structural features are relatively strong, but sequence-structure and sequence-sequence interactions are weak. R: amino acid identity (20 canonical types); S: secondary structure (Helix, strand or loop); B: surface exposure (4 side chain burial states). For clarity, we only show one numerical value of each asymmetric interaction pair (e.g. $I(R;B')$, but not $I(R';B)$), since the asymmetry is small, as can be seen in Fig. 2.

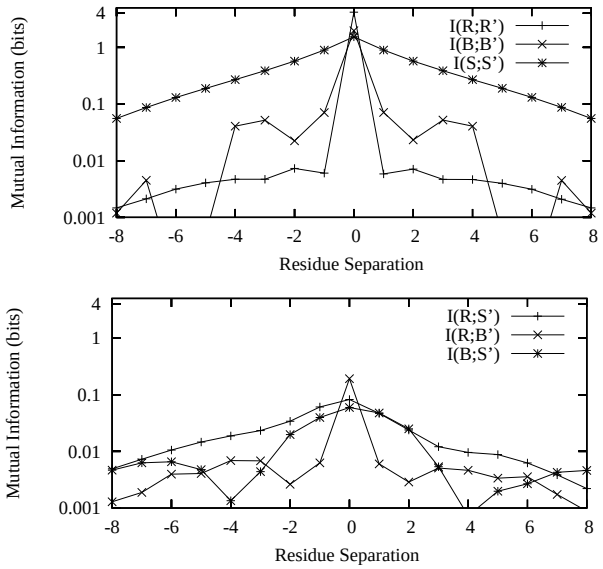


Figure 2: Intra-feature (top) and inter-feature (bottom) mutual information for residues at various separations along the amino acid chain. Note that the self information for the same feature on the same residue is the entropy, $I(X;X) = H(X)$. Interestingly, neighboring amino acids are approximately independent, and the correlations between more distantly separated residues are also very small. This implies that, to a good approximation, amino acid sequences of ordered protein structures are essentially random and uncorrelated[31,2]. This is despite the fact that such sequences have been selected by evolution to fold into compact and functional shapes. Apparently these special sequences are randomly scattered throughout sequence space[32]. Secondary structure is strongly correlated over extended distances, both to itself and to the amino acid sequence. This is to be expected, since secondary structure consists of long helixes and stands interspersed with loop regions. We have previously determined that these non-local interactions increase the effective sequence-secondary structure mutual information from about 0.08 bits (Table 1) to about 0.16 bits per residue[2].

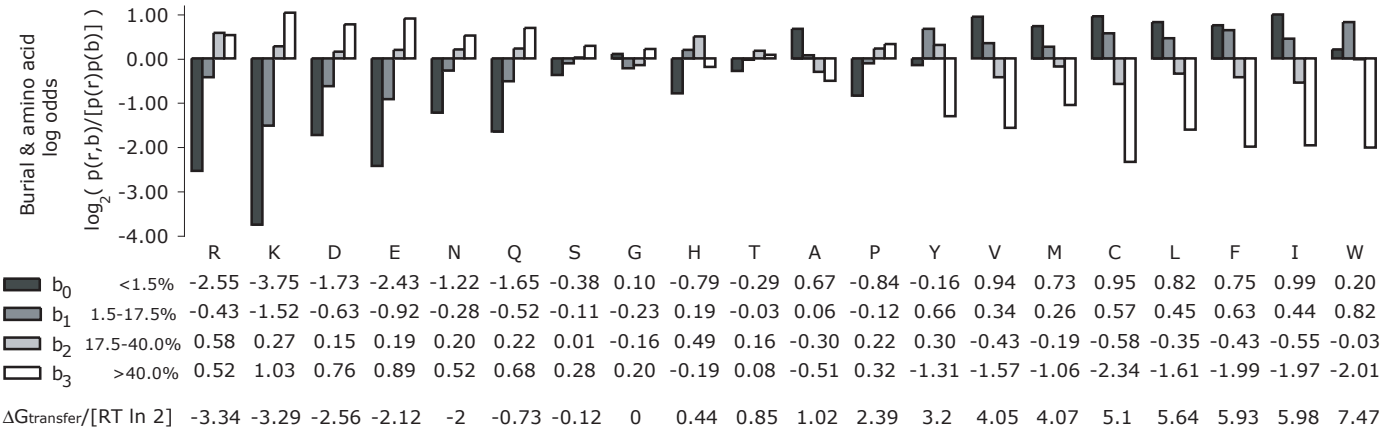
Each residue is identified with one of the 20 canonical amino acids, one of 3 secondary structure classes (helix, strand or loop), and one of 4 side chain burial states, ranging from almost entirely buried to mostly exposed. Residues are non-local neighbors in the tertiary structure if their side chain β-carbons are separated by less than 8 Å, provided that they are separated by at least 6 other residues in the sequence.

RESULTS

Figure 1 summaries the mutual information between amino acid identity (R), burial (B) and secondary structure (S), both for single residues, and for pairs of neighboring residues that are either adjacent along the chain, or proximate in space, but distantly separated along the chain. Tables 1 and 2 show the same data in more detail. The correlations between features of the same residue, and structural features on neighboring residues, are relatively strong, but inter-residue sequence-structure and sequence-sequence interactions are surprisingly weak.

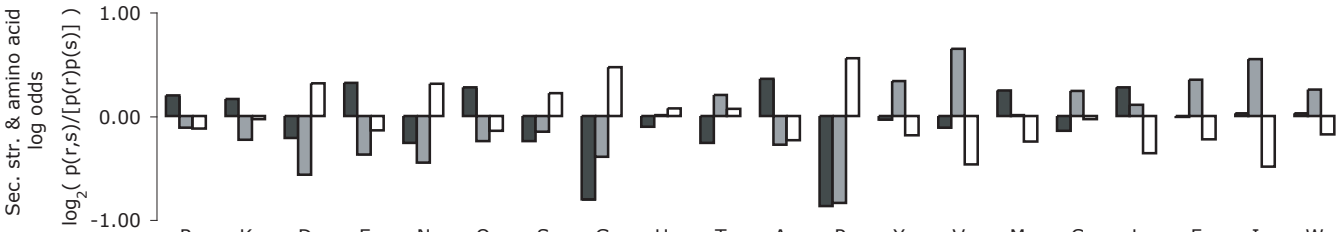
Figure 2 shows correlations for residues separated by varying distances along the chain. It is notable that neighboring amino acids are almost entirely uncorrelated, and therefore that the primary structure of proteins in the PDB is random at the level of linear amino acid sequences. In contrast, secondary structure is strongly correlated along the chain, and exhibits significant delocalized correlations with both primary structure and surface exposure. Secondary structure features also tend to cluster in space, as is evident in Figure 1; β-sheets are hydrogen bonded across strands, and many protein domains are built from mostly strands or mostly helixes[36].

The total, single-residue sequence-structure mutual information is $I(R;B,S) \approx 0.26$ bits (Table 1), indicating significant correlation between local structure and amino acid identity. This information is approximately additive across burial and secondary structure, $I(R;S,B) \approx I(R;S) + I(R;B)$. The discrepancy is the triplet mutual information $I^{(3)}(R;S;B) = I(R;S,B) - I(R;S) - I(R;B) \approx -0.014$ bits (Eq. 5), which is small compared to the pairwise interactions. This suggests that our definitions of secondary structure and burial are non-redundant, and that the inter-



		R	K	D	E	N	Q	S	G	H	T	A	P	Y	V	M	C	L	F	I	W
b_0	<1.5%	-2.55	-3.75	-1.73	-2.43	-1.22	-1.65	-0.38	0.10	-0.79	-0.29	0.67	-0.84	-0.16	0.94	0.73	0.95	0.82	0.75	0.99	0.20
b_1	1.5-17.5%	-0.43	-1.52	-0.63	-0.92	-0.28	-0.52	-0.11	-0.23	0.19	-0.03	0.06	-0.12	0.66	0.34	0.26	0.57	0.45	0.63	0.44	0.82
b_2	17.5-40.0%	0.58	0.27	0.15	0.19	0.20	0.22	0.01	-0.16	0.49	0.16	-0.30	0.22	0.30	-0.43	-0.19	-0.58	-0.35	-0.43	-0.55	-0.03
b_3	>40.0%	0.52	1.03	0.76	0.89	0.52	0.68	0.28	0.20	-0.19	0.08	-0.51	0.32	-1.31	-1.57	-1.06	-2.34	-1.61	-1.99	-1.97	-2.01
$\Delta G_{transfer}/[RT \ln 2]$		-3.34	-3.29	-2.56	-2.12	-2	-0.73	-0.12	0	0.44	0.85	1.02	2.39	3.2	4.05	4.07	5.1	5.64	5.93	5.98	7.47

Figure 3: Log likelihood ratio for amino acid side chain burial. The most hydrophobic and hydrophilic residues strongly segregate between the surface and interior of the protein, the most extreme case being lysine (K), which is 10 times more likely to be exposed than buried. However, most amino acids have only a weak preference for burial or surface exposure, resulting in a net burial-sequence mutual information of only 0.19 bits. Interestingly, tryptophan (W), histidine (H) and tyrosine (Y) have a preference to be partially exposed rather than totally buried or totally exposed. These residues all have large side chains containing both hydrophobic and hydrophilic moieties. Residues are ordered by $\Delta G_{\text{transfer}}/[RT \ln 2]$, the experimental free energy for transferring the amino acid between water and octanol[33,34] (converted to bits), a conventional measure of hydrophobicity. (See also Fig. 5)



	R	K	D	E	N	Q	S	G	H	T	A	P	Y	V	M	C	L	F	I	W
Helix	0.20	0.16	-0.21	0.32	-0.26	0.27	-0.24	-0.81	-0.11	-0.26	0.36	-0.87	-0.04	-0.11	0.24	-0.15	0.27	-0.01	0.03	0.02
Strand	-0.11	-0.23	-0.57	-0.37	-0.45	-0.24	-0.15	-0.40	0.01	0.20	-0.28	-0.84	0.33	0.64	0.01	0.24	0.11	0.35	0.55	0.25
Loop	-0.12	-0.03	0.31	-0.14	0.31	-0.15	0.22	0.47	0.07	0.07	-0.24	0.56	-0.19	-0.47	-0.25	-0.03	-0.36	-0.22	-0.49	-0.18

Figure 4: Residue propensity for secondary structure elements. Some residues clearly favor particular secondary structure features; glycine (G) and proline (P) are over-represented in loops[35], valine (V) and isoleucine (I) in strands, and alanine (A) in helixes. However, these preferences are not particularly strong, resulting in a net single-site secondary structure to amino acid identity mutual information of only 0.08 bits. This sparsity of information is responsible for the poor performance of algorithms that attempt to directly predict secondary structure from the primary structure alone[2]. Residues are ordered by their nominal hydrophobicity, as in figure 3. This reveals a slight preference for hydrophilic residues in helixes and loops, and hydrophobic residues in strands.

actions between these different structural features and the amino acid sequence are approximately independent. Figures 3, 4 and 5 explore these single-site correlations in more detail.

The additional sequence-structure information that can be obtained from neighboring residues, given the local structural context, is the conditional mutual information $I(R; S', B'|S, B)$. Unfortunately, a direct calculation of this mutual information is impractical due to the large dimension of the underlying probability distribution ($20 \times 3^2 \times 4^2$), and the finite size of our data set. This results in a large mutual information bias that is beyond the ability of our bootstrap error analysis to correct. A more effective bias correction can be obtained by performing a full Bayesian analysis[38], but this requires the determination of a reasonable prior distribution. To avoid this additional complication, we take a more direct approach and approximate the full conditional information as the sum of (ap-

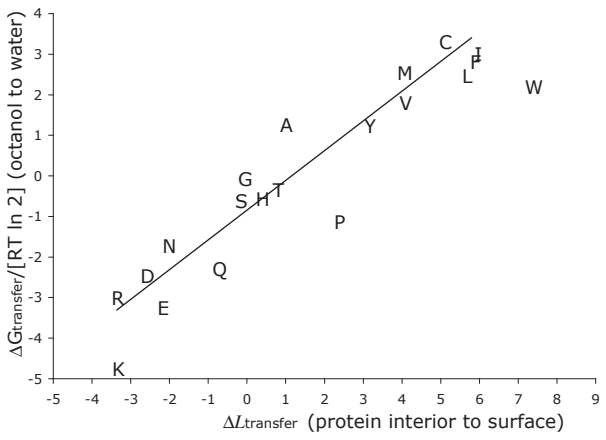


Figure 5: The segregation of residues between the interior and surface of a protein is principally due to hydrophobic effects[16,17]. The vertical axis is the free energy change for transferring N-acetyl-amino-acid amides from water to octanol[33,34]. The transfer free energy of glycine (G) has been set to zero, and we have converted from units of kcal/mol to free entropies measured in bits. The horizontal axis (also measured in bits) is the log odds difference between buried and exposed residues, $\Delta\mathcal{L} = \log_2 p(r,b_0)/[p(r)p(b_0)] - \log_2 p(r,b_3)/[p(r)p(b_3)]$, where, b_0 is a surface exposure of less than 1.5%, and b_3 is a surface exposure of greater than 40%. This is essentially the free entropy for transferring a residue from the interior to the exterior of a protein. The transfer free entropy between octanol and water measures the hydrophobicity of the residue. The log odds difference includes hydrophobic effects, but is also influenced by protein structure and evolution. Nonetheless, there is a near quantitative agreement between these two measures, indicating that hydrophobicity is the primary determinate of the frequency with which different residues segregate to the interiors of proteins. The trend line has a slope of 0.70. It has been suggested that proline is frequently exposed due to steric effects and, that cysteine is often buried due to disulphide bonding[37,34]. However, this figure indicates that the burial frequency of both residues can be adequately explained by hydrophobicity alone.

proximately) independent parts, as suggested by the results for the single site sequence-structure mutual information, $I(R;S,B) \approx I(R;S) + I(R;B)$. Thus, the additional information obtained from considering the structure of a neighboring residue, given the local structural context, can be split into approximately independent burial and secondary structure contributions

$$I(R;S',B'|S,B) \approx I(R;S'|S) + I(R;B'|B)$$

These quantities are tabulated in Table 3. Therefore, the additional sequence-structure information is about 0.030 bits per residue for local neighbors and 0.015 bits for contacts, in excess of the previously determined single site mutual information of 0.26 bits. Note that almost every residue has two nearest neighbors along the chain, and an average of about 2 neighbors through space. Thus the total information content of these sequence-structure interactions is about 0.35 bits per residue.

Table 1: Single residue mutual information.

	bits	std. err. (bias corr.)
$I(R;S,B)$	0.257	±0.0014 (+0.0006)
$I(R;S)$ (3 states)	0.0812	±0.0007 (+0.0001)
$I(R;B)$	0.1900	±0.0013 (+0.0002)
$I(B;S)$	0.0633	±0.0010 (−0.0001)
$I^{(3)}(R;S;B)$	−0.014	±0.0021 (+0.0010)

Having dealt with structure-structure and sequence-structure correlations, the remaining pairwise interactions to consider are those between amino acid identity at neighboring sites. The use of contact sequence correlations for structure template matching is the central idea underlying protein threading, yet it is clear from Fig. 1 that the direct mutual information is small: $I(R;R') \approx 0.006$ for primary structure neighbors, and 0.027 bits for contacts. As with structural features, we should consider the inter-amino acid correlations given the local structure, $I(R;R'|S',B',S,B)$. This is an estimate of the information contained in inter-residue contact potentials. As above, we can approximate this mutual information by a collection of simpler, approximately independent contributions;

$$\begin{aligned} I(R;R'|S',B',S,B) \approx\ & I(R;R') \\ & +(I(R;R'|S',S) - I(R;R')) \\ & +(I(R;R'|B,B') - I(R;R'))\,. \end{aligned}$$

From Table 3, this inter-sequence information is about 0.01 bits for local neighbors and 0.02 bits for contacts. The correlation between amino acid identity at neighboring sites is very small, whether or not those correlations are conditioned on the local structure.

In a previous investigation Cline et al.[18] measured an inter-amino acid contact correlation, given that neither residue is located in a loop, of 0.04 bits per contact. As a comparison we have measured this same correlation on our dataset, using Cline's parameters. We agree qualitatively, but find that this inter-residue correlation in our data is only 0.02 bits per contact, about half as strong as in the Cline data set. This degree of difference is not unexpected, since our dataset is more recent, contains an order of magnitude more sequence data, and is more structurally diverse. However, this comparison suggests that partitioning the protein data into structural classes may marginally increase the apparent inter-residue mutual information.

In summary, the mutual information between the amino acid sequence and the local protein structure is a few tenths

Table 2: Mutual information between neighboring residues.

	Adjacent in Primary Structure		Proximate in Tertiary Structure	
	bits	std. err. (bias corr.)	bits	std. err. (bias corr.)
$I(R;R')$	0.0058	±0.0002 (+0.0005)	0.0268	±0.0003 (+0.0002)
$I(R;B')$	0.0066	±0.0002 (+0.00009)	0.0183	±0.0003 (+0.00005)
$I(R;S')$	0.0468	±0.0004 (+0.00006)	0.0203	±0.0003 (+0.00006)
$I(B;B')$	0.0755	±0.0011 (+0.0001)	0.0822	±0.0009 (+0.00001)
$I(S;S')$	0.898	±0.0017 (−0.0001)	0.1443	±0.0028 (−0.00015)
$I(S;B')$	0.0511	±0.0007 (+0.00009)	0.0117	±0.0003 (+0.00006)

Table 3: Additional mutual information from neighboring residues, conditioned upon local structure.

	Adjacent in Primary Structure		Proximate in Tertiary Structure	
	bits	std. err. (bias corr.)	bits	std. err. (bias corr.)
$I(R;S'\|S)$	0.0153	±0.00098 (−0.00001)	0.0120	±0.0014 (+0.00083)
$I(R;B'\|B)$	0.0148	±0.00190 (+0.00002)	0.0032	±0.0022 (+0.00115)
$I(R;S',B'\|S,B)$	0.0301	±0.00214 (+0.00001)	0.0152	±0.0027 (+0.00198)
$I(R;R'\|S,S')$	0.0160	±0.0037 (+0.0037)	0.0189	±0.0003 (+0.0021)
$I(R;R'\|B,B')$	0.0100	±0.0004 (+0.0077)	0.0207	±0.0003 (+0.0037)
$I(R;R'\|S,S',B,B')$	0.0112	±0.0038 (+0.0109)	0.0202	±0.0005 (+0.0056)

of a bit per residue, whereas the information content of amino acid correlations is only a few hundredths of a bit per proximate pair. The strongest observed determinant of local protein structure was hydrophobicity. The precise numerical values of these measurements will obviously depend on the exact procedure for coarse-graining the protein structure, the choice of local structure features, and also upon the selection of protein structures to study. However, it is implausible that any reasonable change in our methodology, such as adjusting the definition of amino acid contacts, could substantially alter the relative magnitudes of sequence-structure and sequence-sequence interactions.

DISCUSSION

Amino acid contact potentials have been widely used in many simplified protein structure prediction algorithms, and in simplified models of proteins folding[8,9,10,11,12,13]. However, the signal from these tertiary contact correlations is far less than that embedded in burial, secondary structure, and other local structural features. We might reasonably question whether contact potentials materially contribute to the performance of these structure prediction algorithms in practice. One alternative to contact potentials is to embrace greater molecular detail in protein structure prediction, while still keeping the model simple enough to be computationally tractable[5]. On the other hand, if we disregard direct amino acid contact correlations altogether, then threading, and many similar structure prediction methods, can be recast in term of protein structure profiles[6], which are computationally efficient, and conceptually very similar to the sequence profile methods used for homology detection.[39]

Many effective structure prediction methods incorporate evolutionary information, as well as structure-sequence correlations. For detectable remote pairwise similarity, the mutual information for aligned amino acids is about 2/3 bit per residue (entropy of BLOSUM62 substitution matrix[40]), whereas multiple sequence alignments typically provide 1-3 bits per position[41]. It appears that amino acid correlations due to evolutionary constraints are stronger than those imposed by structural constraints alone. Thus, many structure prediction methods that incorporate homology might be more accurately thought of as complex homology detection algorithms that are enhanced by limited amounts of structural information.

In principle, direct interactions between amino acids determine the native structure of a protein. Therefore, it is somewhat puzzling that these interactions do not lead to strong inter-residue amino acid correlations. It is possible that the important details of the molecular interactions are washed out when we use a fixed template structure or otherwise coarse-grain the protein structure. However, it is inter-

esting to ask how strong we should expect any primary-to-tertiary structure interaction to be. One available datapoint is the free energy of protein folding, an approximate measure of tertiary structure stability. The stability of small, single domain proteins is typically about[42] $\Delta G \approx 20$–60 kJ mol^{-1}. At ambient temperatures, this corresponds to a free entropy change of $\Delta\psi \approx 600$ kJ mol^{-1} K^{-1}. The conversion factor between traditional thermodynamic entropy units and bits is $R \ln 2$, where $R = k_B N_A$ is the gas constant, and $\ln 2$ converts between natural and base 2 logarithms. Thus, the typical free entropy difference between native and denatured protein is about $\Delta\psi_{\text{folding}} \approx 10$–$30$ bits per macromolecule, or a few tenths of a bit per residue. This entropy of folding represents the cumulative effect of many molecular interactions, and does not directly correspond to any of the coarse-grained structural entropies studied in this work. However, it does provide a pertinent sense of scale, and reaffirms that protein structure determination is difficult because the interactions between amino acid sequence and macro-molecular structure are subtle.

ACKNOWLEDGMENTS

We would like to thank Richard E. Green, Melissa Cline and Phillip Geissler for helpful discussions and suggestions; and R. C. Edgar and Marcin Joachimiak for critical readings of this manuscript. The protocol for constructing a non-redundant, high quality protein structure data set was originally devised by K. Tomii. This work was supported by the National Institutes of Health (1-K22-HG00056). GEC received funding from the Sloan/DOE postdoctoral fellowship in computational molecular biology. SEB is a Searle Scholar (1-L-110).

References

1. Rost B. Review: Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.

2. Crooks GE, Brenner SE. Protein Secondary Structure: Entropy, Correlations and Prediction. Bioinformatics 2004;.

3. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequences. Proc IEEE International Conference on Neural Networks 1997;:9095.

4. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: Implications for structural proteomics. Structure Camb 2003;11:1453–1459.

5. Simons K, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.

6. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.

7. Alexandrov NN, Nussinov R, Zimmer RM. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. Pac Symp Biocomput 1996;:53–72.

8. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.

9. Mirny LA, Shakhnovich EI. Protein structure prediction by threading. Why it works and why it does not. J Mol Biol 1998;283:507–526.

10. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19:874–881.

11. Levitt M, Warshel A. Computer simulation of protein folding. Nature 1975;253:694–698.

12. Miyazawa S, Jernigan RL. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.

13. Abkevich VI, Gutin AM, Shakhnovich EI. Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. J Chem Phys 1994;101:6052–6062.

14. Green RE, Brenner SE. Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE 2002;9:1834–1847.

15. Sander C, Scharf M, Schneider R. Design of protein structures. In: Rees A, Sternberg M, Wetzel R, editors, Protein Engineering, a Practical Approach. Oxford University Press 1992.

16. Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–12.

17. Wolfenden RV, Cullis PM, Southgate CC. Water, protein folding, and the genetic code. Science 1979;206:575–577.

18. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG, Haussler D. Information-theoretic dissection of pairwise contact potentials. Proteins 2002;49:7–14.

19. Shannon CE. A mathematical theory of communication. Bell Sys Tech Journal 1948;27:379–423, 623–656.

20. Cover TM, Thomas JA. Elements of Information Theory. Wiley 1991.

21. Fisher IZ. Statistical Theory of Liquids. University of Chicago Press 1964.

22. Wolf DR. Information and correlation in statistical mechanical systems. PhD thesis University of Texas, Austin 1996.

23. Miller GA. Note on the bias of information estimates. In: Information Theory in Psychology; Problems and Methods II-B. Free Press 1955; 95–100.

24. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press 1993.

25. Matsumoto M, Nishimura T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Trans Model Comput Simul 1998;8:3–30.

26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.

27. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.

28. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:226–229.

29. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32:189–192.

30. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.

31. Weiss O, Jimenez-Montano MA, Herzel H. Information content of protein sequences. J Theor Biol 2000;206:379–386.

32. Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF. Exploring protein sequence space using knowledge-based potentials. J Theor Biol 2001;212:35–46.

33. Fauchere J, Pliska V. Hydrophobic parameters π of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. Eur J Med Chem 1983;18:369375.

34. Karplus PA. Hydrophobicity regained. Protein Sci 1997;6:1302–1307.

35. Leszczynski J, Rose GD. Loops in globular proteins. Science 1986;234:849–855.

36. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.

37. Ponnuswamy P. Hydrophobic characteristics of folded proteins. Prog Biophys Mol Biol 1993;59:57–103.

38. Wolpert DH, Wolf DR. Estimating functions of probability distributions from a finite set of samples. Phys Rev E 1995;52:6841–6854.

39. Gribskov M, McLachlan A, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 1987;84:4355–4358.

40. Henikoff S, Henikoff JG. Amino-acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.

41. Karplus K. Regularizers for estimating distributions of amino acids from small samples. In: 3rd International Conference on Intelligent Systems for Molecular Biology, Menlo Park, CA. AAAI/MIT Press July 1995. http://www.cse.ucsc.edu/research/compbio/dirichlets/.

42. Brandon C, Tooze J. Introduction to Protein Structure. Garland 2nd edition 1998.