BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 21 2012, pages 2755–2762
doi:10.1093/bioinformatics/bts512

Structural bioinformatics

Advance Access publication August 24, 2012

Information-theoretic analysis and prediction of protein atomic burials: on the search for an informational intermediate between sequence and structure

Juliana R. Rocha†, Marx G. van der Linden†, Diogo C. Ferreira, Paulo H. Azevêdo and Antônio F. Pereira de Araújo*

Laboratório de Biologia Teórica e Computacional, Departamento de Biologia Celular, Universidade de Brasília, Brasília-DF 70910-900, Brazil

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: It has been recently suggested that atomic burials, as expressed by molecular central distances, contain sufficient information to determine the tertiary structure of small globular proteins. A possible approach to structural determination from sequence could therefore involve a sequence-to-burial intermediate prediction step whose accuracy, however, is theoretically limited by the mutual information between these two variables. We use a non-redundant set of globular protein structures to estimate the mutual information between local amino acid sequence and atomic burials. Discretizing central distances of C_α or C_β atoms in equiprobable burial levels, we estimate relevant mutual information measures that are compared with actual predictions obtained from a Naive Bayesian Classifier (NBC) and a Hidden Markov Model (HMM).

Results: Mutual information density for 20 amino acids and two or three burial levels were estimated to be roughly 15% of the unconditional burial entropy density. Lower estimates for the mutual information between local amino acid sequence and burial of a single residue indicated an increase in mutual information with the number of burial levels up to at least five or six levels. Prediction schemes were found to efficiently extract the available burial information from local sequence. Lower estimates for the mutual information involving single burials are consistently approached by predictions from the NBC and actually surpassed by predictions from the HMM. Near-optimal prediction for the HMM is indicated by the agreement between its density of prediction information and the corresponding density of mutual information between input and output representations.

Availability: The dataset of protein structures and the prediction implementations are available at http://www.btc.unb.br/ (in 'Software').

Contact: aaraujo@unb.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 5, 2012; revised on July 31, 2012; accepted on August 13, 2012

1 INTRODUCTION

It has been a common statement in biology that amino acid sequences contain sufficient information to determine protein tertiary structures. Fulfilment of the implied possibility of structure prediction from sequence is actually considered one of the most important unsolved problems of molecular biophysics, as reviewed by different groups (Dill *et al.*, 2008; Onuchic and Wolynes, 2004; Shakhnovich, 2006). Such an intrinsically informational assertion, however, has only more recently been extensively investigated within the context of Shannon's information theory. Although informational concepts have been used in algorithms for secondary structure prediction from local sequence since the 70s (Garnier *et al.*, 1978), for example, the limit imposed on prediction by the mutual information between these two quantities was estimated only a few years ago (Crooks and Brenner, 2004). Incidentally, an informational analysis of backbone dihedral angles has also exposed the unfeasibility of tertiary structure determination from an even perfect three-state secondary structure prediction (Solis and Rackovsky, 2004). The recurrent utilization of statistical potentials in computational biology has also been interpreted explicitly in informational terms (Solis and Rackovsky, 2007). A particularly relevant example is the analysis of pairwise contact potentials, which revealed a surprisingly modest mutual information between contact partners (Cline *et al.*, 2002; Crooks *et al.*, 2004). General distance constraints have also been investigated, at least in the context of minimalist protein models (Sullivan *et al.*, 2003).

Contrasting with secondary structure, atomic burials appear to encode sufficient information for structural determination. Contrasting with pairwise contacts, they have a much better chance of being adequately estimated from sequence information. Monte Carlo simulations of geometrically realistic protein models using native burial information, as expressed by atomic distances from the molecular center, have successfully recovered the tertiary structure of small globular proteins (Pereira de Araújo *et al.*, 2008). A simple computational experiment combining Molecular Dynamics of similar models with discretized burial levels has additionally provided an upper bound for the amount of required burial information. It actually turned out to be comparable to, and therefore encodable by, the information (entropy) of local protein sequences (Pereira de Araújo and Onuchic, 2009). The observed discriminatory difference between

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

burial and secondary structure representations does not arise therefore from a trivial difference in precision. A very precise representation of all backbone dihedral angles can clearly encode tertiary structures, even using a small amount of information, or number of letters, in α-helical regions and possibly β-strands, but requiring a large, sequence-incompatible, number of letters in intervening loops. The distinction appears to be more basic and related to different types of information encoded in the two local representations. While secondary structure is a local representation of purely local structure, burials include global structural information in a local representation, as is evident from the fact that the whole tertiary structure is required for determination of burials, but not secondary structure, of any short fragment of amino acids.

The possibility of structural determination from sequence-dependent burial information, when combined to appropriate sequence-independent constraints, is consistent with the perceptible previous success in native fold recognition from the arrangement of hydrophobic and polar residues (Huang *et al.*, 1995). It has also been further supported recently by a purely analytical model which was able to recover native-like burial traces from sequence hydrophobicity information combined to simple constraints on chain connectivity and overall globular size (England, 2011). A potential approach to tertiary structure prediction could therefore involve a sequence-to-burial intermediate prediction step. It must be noted that theoretical encodability, as provided by entropy compatibility, is necessary but not sufficient to demonstrate actual encoding. The accuracy of any burial prediction from sequence must be further limited by the observed correlation between burials and sequences, as conveniently quantified by the mutual information between these two quantities. In this study, we estimate the mutual information between burials and local amino acid sequence in globular proteins. The resulting fraction of sequence entropy actually involved in burial encoding provides theoretical limits to which prediction algorithms should be compared. We additionally investigate the efficiency of simple statistical prediction schemes, namely, a Naive Bayesian Classifier (NBC) and a Hidden Markov Model (HMM), in extracting the available burial information from local sequence.

2 METHODS

In this study, we estimated probabilities from frequencies observed in a dataset of representative globular structures derived from PDBSELECT (Hobohm and Sander, 1994). From the list made available in November 2009, we selected structures determined by X-ray crystallography with resolution better than 2.5 Å and excluded chains not satisfying the globularity criterion given by the expected relation between radius of gyration and the number of residues, $R_g \leq 2.9 N_r^{1/3}$Å (Gomes *et al.*, 2007). Membrane proteins were also excluded, simply by removing PDB files containing the word 'MEMBRANE'. The resulting collection, from now on simply referred to as the databank, is composed of 1499 chains, with a total of ~263 000 residues. Statistical errors on computed probabilities and entropies were estimated, and systematic biases corrected for, by a bootstrap procedure using 50 randomly generated replicas of the databank (Crooks and Brenner, 2004; Efron and Tibshirani, 1993). In addition to the complete alphabet of 20 amino acid identities, we have also used the reduced alphabets HP and HPN. Hydrophobic and polar residues were grouped in the HP alphabet as $\mathrm{H} = \{\mathrm{A, C, F, G, I, L, M, V, W, Y}\}$ and $\mathrm{P} = \{\mathrm{D, E, H, K, N, P, Q, R, S, T}\}$, respectively. In HPN a third, 'neutral', class includes residues from both HP groups, $\mathrm{N} = \{\mathrm{A, G, H, S, T}\}$. Burials, b, were obtained from the atomic distances from the molecular center, r, of C_α or C_β atoms, normalized by the radius of gyration, R_g, or $b = r/R_g$, and grouped in approximately equiprobable burial levels, resulting in a collection of burial alphabets $\{\chi L\}$, where χ is either α or β, representing the atomic type for which burials are defined, and L is the number of burial layers. Cutoff burial values for different burial levels were obtained from the estimated burial distribution obtained by Gomes *et al.* (2007). We usually use superscripts to indicate block size and integer subscripts to indicate position within the block, with '0' representing the central block position by convention. If necessary, however, we also indicate particular alphabets as subscripts in our notation, such as $H(Q^N_{\mathrm{HP}}), h(B_{\beta 5}), I(Q^N_{20}; B^N_{\alpha 2})$.

N-block entropies for residue identities, $H(Q^N)$, and burials, $H(B^N)$, were computed according to Shannon's basic equation

$$H(X^N) = -\sum_{x^N} p(x^N) \log_2 p(x^N),$$

where the sum is over all blocks of N adjacent letters, x^N, either identities or burials, and probabilities are estimated from corresponding frequencies in the databank. A linear dependence of the estimated entropy on block size in the range $m < N < m'$,

$$H(X^N) = Nh(X) + E_X. \qquad (1)$$

is consistent with a Markovian process of order m, where $h(X)$ is the entropy density and E_X is the N-independent excess entropy, which indicates the uncertainty resolved by local correlations. Deviation from linearity for $N < m$ arises from these local correlations between letters while for $N > m'$ frequencies in the databank become poor estimates for actual probabilities and the estimated entropy converges to an alphabet-independent value that depends on the overall size of the databank, a situation we refer to as 'saturation'. Estimates for $h(X)$ and E_X can therefore be obtained from the observed dependence of $H(X^N)$ on N if the order of the underlying Markov process is sufficiently small and the dataset is sufficiently large so that $m \ll m'$ and the linear region can be clearly identified.

For the mutual information between blocks of identities and burials, Q^N and B^N, a limiting linear behavior is also expected, or

$$I(Q^N; B^N) = H(Q^N) - H(Q^N|B^N) = Ni(Q; B) + E_{Q;B}. \qquad (2)$$

and an estimate for the corresponding mutual information density, $i(Q; B)$, a quantity of much interest that imposes an upper limit on any possible prediction of the local sequence of burials from the local sequence of identities, could again be obtained from N-block entropy estimates. In this case, however, because the number of different blocks increases more sharply with block size, saturation should occur at a much shorter block length. We use therefore an approximation,

$$i(Q; B) \approx \lim_{N \to \infty} I(Q_0; B^N) \equiv I(Q_0; B^\infty). \qquad (3)$$

that is valid when the letters in one of the sequences are statistically independent both unconditionally and conditionally to the other sequence, as it turns out to be the case for identities with respect to burials. The density of mutual information is estimated accordingly by extrapolation of the dependence on N of $I(Q_0; B^N)$, the mutual information between N-blocks of burials, B^N, and the identity of the central residue in the block, Q_0,

$$I(Q_0; B^N) = H(Q_0) - H(Q_0|B^N). \qquad (4)$$

where $H(Q_0)$ is the single identity entropy, obtained with probabilities estimated directly from corresponding frequencies, and $H(Q_0|B^N)$ is the conditional entropy of central residue identity conditional to burial block. This procedure was used by Crooks and Brenner (2004) to estimate the mutual information density between sequences of amino acid residues and corresponding sequences of secondary structure assignments.

Underlying conditional probabilities were obtained from corresponding frequencies, or $p(Q_0|B^N) = n(Q_0, B^N)/n(B^N)$, only for the HP alphabet, since statistics turned out to be sufficient. For the other alphabets conditional probabilities were estimated as

$$p(Q_0|B^N) = \frac{n(Q_0, B^N) + (20 \times p(Q_0|B_0))}{n(B^N) + 20}, \tag{5}$$

using 20 'pseudo-counts' with prior probability $p(Q_0|B_0)$ in an attempt to minimize artifacts from low-frequency events. Due to pseudo-counts, the estimated mutual information turns out to be increasingly smaller than its actual value as N becomes large. While the actual mutual information must increase monotonically with N, its estimate will decrease for large N, providing again a simple signature of databank saturation. For the HP alphabet, pseudo-counts were not used and saturation manifests itself as an abrupt increase in estimated mutual information causing an upward inflection in the estimated curve. Data points were fitted, before saturation, to a single exponential $f(x) = a - b\exp(-x/c)$, with limiting behavior provided by adjusted parameter a, or to a symmetrically inflected sigmoid $f(x) = \frac{a}{(1-\exp(-b(x-c)))} + d$, with limiting behavior provided by $a + d$. Fitting to an asymmetric Gompertz function provided similar estimates but with larger errors, reflecting the larger number of adjustable parameters (not shown).

In addition to $I(Q_0; B^\infty)$, we are also interested in the converse quantity, $I(Q^\infty; B_0)$, since it provides a limit for the prediction of individual burial values given the local sequence of identities. Saturation might again become a problem for large alphabets of identities, in which case it is useful to consider the following lower bound:

$$\begin{aligned}\sum_{i=1}^{N} I(Q_i; B_0) &= \sum_{i=1}^{N} [H(Q_i) - H(Q_i|B_0)] \leq H(Q^N) - H(Q^N|B_0) \\ &= I(Q^N; B_0),\end{aligned} \tag{6}$$

with limiting behavior

$$I(Q^\infty; B_0)^- \equiv \lim_{N\to\infty} \sum_{i=1}^{N} I(Q_i; B_0) \leq I(Q^\infty; B_0) \tag{7}$$

Each of the N 'positional' mutual information terms between Q_i and B_0 is computed from the same number of possible combinations, independently of N. The results for the tractable HP alphabet and two burial levels, shown in the Supplementary Information, indicate that Equation (3) is indeed a good approximation while a strict inequality is expected in Equation (7).

In order to compare our mutual information estimates with actual predictions, we implemented two simple statistical schemes for predicting discrete atomic burial levels from amino acid sequence in globular proteins: a NBC and a HMM. Both methods are supervised learning algorithms, i.e. they employ a learning step, in which they gather data from a training set to generate some statistical model, followed by a prediction step, in which they use the model to predict new data. We have used the same dataset of structures as for the informational analysis, now randomly divided in training and testing subsets. Statistical errors and biases were again estimated by bootstrapping resampling with 50 replicas. While the NBC estimates the probability for different burial levels of a given residue simply from a local 'window' of identities in the primary sequence, neglecting most correlations between adjacent residues, the HMM considers explicitly the correlations between 'fragments' of hidden variables, including burials, which are modeled as producing the observed primary sequence. Both algorithms are described in detail in the Supplementary Information, as well as the procedures to obtain the corresponding prediction information, I_p, and prediction information densities, i_p, to be compared with the mutual information estimates $I(Q^\infty; B)^-$ and $i(Q; B)$, respectively.

3 RESULTS

Figure 1 illustrates the statistical behavior of local sequences of C_β burials, as determined from central distances normalized by radius of gyration. N-block entropy is shown as a function of block size N. Different curves correspond to different alphabets, ranging from two to five equally probable burial levels. Deviation from linearity for large N results from saturation of the databank as all curves converge to the same alphabet-independent saturated limit behavior. Deviation from linearity for small N and, more perceptively, a positive intercept with the ordinate axis reflect the expected local correlations between adjacent burial levels. These results suggest a low-order markovicity, with m not higher than 2 or 3. Analogous results for C_α burials, shown in the Supplementary Information, indicate a qualitatively similar behavior. For identities, on the other hand, as also shown in the Supplementary Information, it is apparent that $H(Q^N)$ increases linearly from the origin for all alphabets, being consistent with zero-order markovicity, $m = 0$, or equivalently, statistical independence between amino acid identities along the sequence. Accordingly, as shown in Table 1, residue entropy density $h(Q)$ is very close to the single letter entropy, $H(Q^1)$, increasing from essentially 1 for HP sequences, $h(Q_{\mathrm{HP}}) \approx H(Q^1_{\mathrm{HP}}) \approx 1$ bit/residue, to $h(Q_{20}) \approx H(Q^1_{20}) \approx 4.18$ bits/residue for 20 amino acid letters while mutual information between adjacent identities is close to zero. Entropy densities of correlated burials, however, are significantly lower than corresponding single burial entropies, with a positive mutual information between adjacent burials, such as $h(B_{\alpha 2}) \approx 0.62 < H(B_{\alpha 2}) \approx 1$ bit/residue and $I(B_{;}B_{i+1}) \approx 0.34$ bit for two C_α burial levels. C_β burials consistently display larger entropy densities, such as $h(B_{\beta 2}) \approx 0.73$ and $h(B_{\beta 3}) \approx 1.1$ bits/residue for two and three burial levels, respectively, to be compared with $h(B_{\alpha 2}) \approx 0.62$ and $h(B_{\alpha 3}) \approx 0.95$ bit/residue for C_α burials.

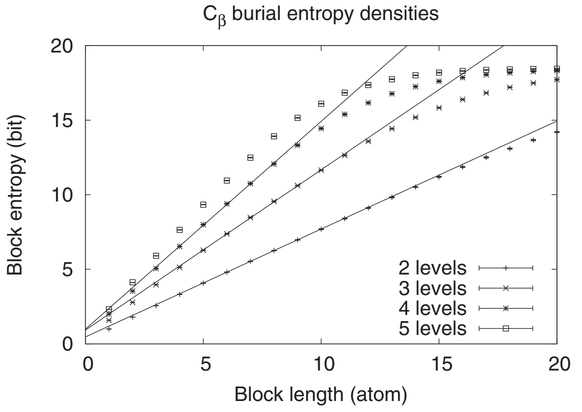

Fig. 1. N-block sequence entropy estimates as a function of block size N for different alphabets of C_β burial levels. Both the entropy density (inclination) and excess entropy (intersect with the ordinates) are obtained from straight lines fitted to the linear region, which is clearly identified for $L = 2$ and $L = 3$. Deviation from linearity for small N is indicative of local correlations while deviation at large N is due to databank saturation. Analogous results for amino acid identities and C_α burials are shown in the Supplementary Information

Table 1. Single sequence analysis

	***H*(*X*)**	$I(X_i; X_{i+1})$	***h*(*X*)**	E_X
HP	1.00000(9)	0.00072(8)	0.9969(2)	0.0096(7)
HPN	1.5806(4)	0.0009(1)	1.5734(7)	0.018(3)
20	4.185(2)	0.005(7)	4.176(4)	0.010(5)
$\alpha 2$	0.99974(6)	0.342(3)	0.619(1)	0.513(9)
$\alpha 3$	1.5796(3)	0.574(3)	0.933(4)	0.91(2)
$\beta 2$	0.9988(1)	0.211(3)	0.724(2)	0.46(2)
$\beta 3$	1.5804(2)	0.377(4)	1.075(6)	0.92(4)

Letter entropy, $H(X)$, and mutual information between adjacent letters, $I(X_i; X_{i+1})$, are in bits. Entropy density $h(X)$, in bits/letter, and corresponding excess entropy E_X, in bits, were obtained from data fits shown in Figure 1 or in the Supplementary Information. Each line corresponds to a different alphabet of amino acid identities or burials, as indicated in the first column. Error in the last significant digit is shown in parentheses.

The dependence on N of the estimates for mutual information between N-blocks of burials and central residue identities, $I(Q_0; B^N)$, is shown in Figure 2 for two and three levels of C_β burials. Analogous results for C_α burials are shown in the Supplementary Information. Mutual information density, $i(Q; B) \approx I(Q_0; B^\infty)$, was obtained by extrapolation from exponential or sigmoidal fits to the points before saturation, as indicated by solid lines and shown in Table 2. Mutual information density is always larger for C_β burials when compared with C_α burials with the same alphabet combination, such as $i(Q_{20}; B_{\alpha 2}) \approx 0.09 < i(Q_{20}; B_{\beta 2}) \approx 1.13$ bits/residue. As could be anticipated, it tends to increase with alphabet size either of amino acid identities or burials such as, in the case of C_β atoms, from $i(Q_{HP}; B_{\beta 2}) \approx 0.07$ bit/residue for the HP alphabet and $L = 2$ burial layers, to $i(Q_{20}; B_{\beta 3}) \approx 0.18$ bit/residue, for 20 amino acid letters and $L = 3$ layers. Databank saturation prevented reliable density estimates for $L > 3$.

Positional mutual information values, $I(Q_i; B_0)$, are shown in Figure 3a for 20 amino acid letters and different numbers of burial levels of C_β atoms. Positional mutual information is essentially 0 for burial and identity pairs separated by more than 15 residues. We therefore use the sum $\sum_{i=1}^{N} I(Q_i; B_0)$ with $N = 31$ as a reasonable approximation of $I(Q^\infty; B_0)^- \equiv \sum_{i=1}^{\infty} I(Q_i; B_0)$ which, as indicated in the Supplementary Information, is expected to be a lower bound for $I(Q^\infty; B_0)$. We were also able to explore the effect of many burial levels on $I(Q^\infty; B_0)^-$. As shown in Figure 3b, $I(Q^\infty; B_0)^-$ for C_β increases significantly from two layers to five layers, approximately from 0.13 to 0.18 bit, but only slightly for additional layers with asymptotic limit close to 0.2 bit. Qualitatively similar results were obtained for C_α atoms but mutual information between single burials and local sequence tends again to be smaller in this case when compared with C_β atoms, although the difference is smaller than for mutual information density, as also seen in Table 2. We also show for comparison in the same table the mutual information between single letters, $I(Q; B)$.

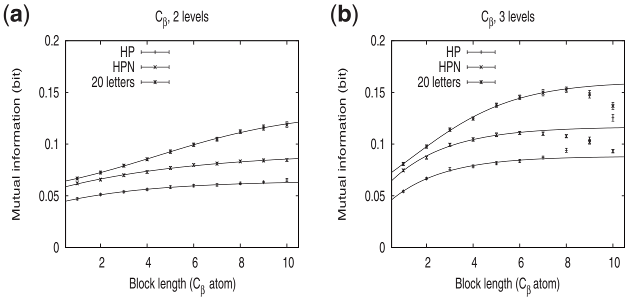

Fig. 2. Estimates for the mutual information, $I(Q_0; B^N)$, between a single central amino acid identity, Q_0, and N-blocks of burials, B^N, as a function of block size N, for two (**a**) and three (**b**) levels of C_β burials. Different sets of points correspond to different alphabets of amino acid identities. Lines represent exponential or sigmoidal fits to the data before saturation from which limiting values $i(Q; B) \approx I(Q_0; B^\infty)$ are obtained. Saturation for $L = 2$ occurs at $N \approx 11$ and is not perceived in the displayed range, while for $L = 3$ it occurs $N \approx 8$, as observed in (b). Analogous results for C_α burials are shown in the Supplementary Information

Table 2. Inter-sequence analysis

L		***I*(*Q*; *B*)**		***i*(*Q*; *B*)**		$I(Q^\infty; B_0)^-$	
		C_α	C_β	C_α	C_β	C_α	C_β
2	HP	0.0297(6)	0.0472(9)	0.050(3)	0.068(2)	0.059(3)	0.070(3)
	HPN	0.0420(9)	0.062(1)	0.068(3)	0.092(2)	0.089(7)	0.100(4)
	20	0.046(1)	0.067(1)	0.091(7)	0.13(1)	0.113(5)	0.124(5)
3	HP	0.0357(9)	0.054(1)	0.066(4)	0.088(2)	0.075(4)	0.086(4)
	HPN	0.051(1)	0.075(1)	0.091(4)	0.117(3)	0.114(5)	0.125(5)
	20	0.0570(9)	0.081(2)	0.130(6)	0.176(6)	0.149(7)	0.159(6)

Mutual information between single letters, $I(Q; B)$ in bits, mutual information density, $i(Q; B)$ in bits/pair, as obtained in Figure 2 and Supplementary Information, and the lower estimate for the mutual information between single burial and local sequence of identities, $I(Q^\infty; B_0)^-$, as obtained in Figure 3, for C_α and C_β atoms are shown for different combinations of identity alphabet and number of burial layers, as indicated in the first two columns. Error in the last significant digit is shown in parentheses.

The performance of two-layer C_β burial prediction is summarized in Figure 4. Analogous results for C_α burials are shown in the Supplementary Information. Prediction accuracy, A (a,b), and prediction information, I_p (c,d), as determined by Equations (S9) and (S10) of the Supplementary Information, are plotted as a function of window size for the NBC (a,c) and as a function of fragment size for the HMM (b,d). In addition to the complete alphabet of 20 amino acids, tests were also performed using the HP and HPN-reduced alphabets. For the NBC, we report results for the simpler variation provided by Equation (S4) of the Supplementary Information, NBC1 (non-shaded symbols), and also for the variation using positional probabilities conditional to central residue identity, as provided by Equation (S5) of the Supplementary Information, NBC2 (shaded symbols). Both accuracy and information increase significantly as the window grows from one to nine residues, but not perceptibly for longer windows. Overall performance is higher for C_β than for C_α atoms. For 20 amino acids, accuracy increases from ~61% to above 65% for C_α atoms and from around 63% to above 66% for C_β. These few percentage points in accuracy improvement actually correspond to around 100% increase in prediction information, from around 4 to above 10 centibits and

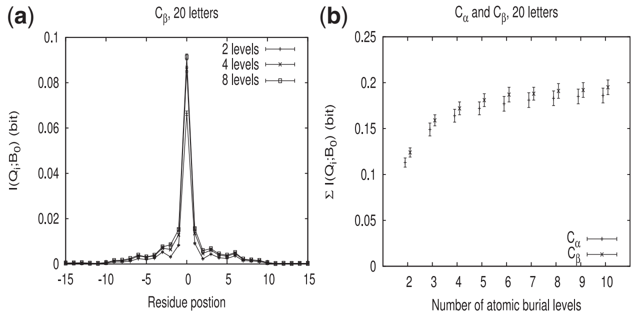

Fig. 3. Positional mutual information $I(Q_i;\, B_0)$ between amino acid identity at position i, Q_i, within the N-block of identities Q^N, and central C_β burial, B_0, for 20 amino acid letters and various numbers of burial levels (**a**) and limiting behavior for the sum of positional mutual information terms, obtained with fixed block size $N = 31$, as a function of the number of burial levels for C_α and C_β atoms (**b**)

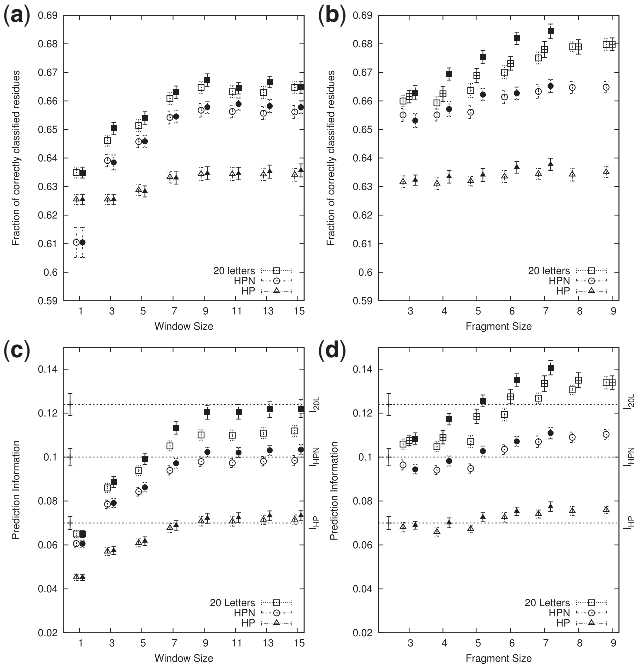

Fig. 4. Prediction accuracy A (**a**,**b**) and prediction information I_p (**c**,**d**) for two levels of C_β burials with different identity alphabets. Plots in the first column (a,c) show results for NBC predictions; the second column (b,d) refers to the HMM results. The NBC method is bounded, within error, to the limits established by corresponding $I(Q^\infty;\, B_0)^-$ estimates (dotted horizontal lines), while the same limits are surpassed by the HMM method (d). In all plots, unshaded symbols represent the simplest version of each algorithm (NBC1 or HMM with nothing but burial levels encoded into the hidden variables) and shaded symbols represent improved versions (NBC2 or HMM with secondary structures). For HMM, half-shaded symbols represent the version that used side-chain orientations

from around 6 to above 11 centibits for C_α and C_β, respectively, for NBC1. Further improvement provided by NBC2, although hardly perceptible in the accuracy measure, is consistently observed for prediction information, accounting for more than 1 centibit of additional information for 20 amino acids while sampling error is of the order of millibits. For the HP and HPN alphabets, both NBC1 and NBC2 predictions agree, within sampling error, with the corresponding lower limits provided by $I(Q^\infty;\, B_0)^-$ while for 20 amino acids this is the case for NBC2.

For the HMM, tested fragment lengths ranged from 3 to 9, but some configurations could not be tested due to hardware constraints related to computer memory usage with many hidden variables. It is clear in the plots of Figure 4b and d that the fragment length has a direct correlation with the quality of results for HMM prediction, especially when the full 20-letter alphabet is used to represent amino acid sequences. The connections between burial levels and secondary structures (shaded symbols) and between burial levels and two possible side chain orientations (obtained from the comparison between C_β and C_α burials and represented as half-shaded symbols) were also investigated by incorporating the corresponding hidden variables into the HMM states. Both approaches were successful in improving the prediction of burial levels, and the usage of secondary structures was slightly more effective than that of side-chain orientations. Incidentally, it was found that not only the prediction accuracy of burial levels but also that of secondary structures is improved when both features are considered together (data not shown). Our most accurate results for burial prediction were around 67.5 and 68.5% of correctly classified residues, respectively, for C_α and C_β. Corresponding prediction information values of ~0.13 and 0.14 bit are higher than the lower limits provided by $I(Q^\infty;\, B_0)^-$, as was consistently observed for the HMM algorithm, particularly with the configurations that employed additional descriptors to the hidden variables and fragment sizes of at least six to seven residues. As with the NBC, prediction of C_β was generally better than that of C_α.

Since the HMM algorithm works with relative probabilities of fragments of burial levels, it is meaningful to estimate the *density* of prediction information, i_p, according to Equation (S12) of the Supplementary Information, i.e. the amount of new prediction information discovered for each new residue once the previous burials have already been established. Figure 5 shows $h_N(B|B(Q))$ (a) Equation (S14) of the Supplementary Information, for the various HMM prediction schemes for C_β burials, as well as corresponding values of $h_N(B)$, Equation (S13) of the Supplementary Information, computed from block entropies shown in Figure 1. The difference between these quantities is the estimate for the prediction information density, i_p, Equation (S12) of the Supplementary Information, which is shown in (b). Our results can be compared with the corresponding estimates for the mutual information density between sequences and burials, $i(B;\, Q)$, from Table 2, also displayed in (b) as dotted horizontal lines, which should act as effective upper limits on prediction quality. Analogous results for C_α burials are shown in the Supplementary Information. Since i_p for $N \geq 7$ agrees within sampling error with $i(B;\, Q)$, it is suggested that our best overall results for two burial levels are extracting virtually all of the burial information that is available in local sequences.

Figure 6 compares the prediction information achieved when the NBC and HMM methods are applied to predict discrete C_β burials into more than two layers. Analogous results for C_α are shown in the Supplementary Information. In all cases, it is clear that the quality of prediction is improved when the number of

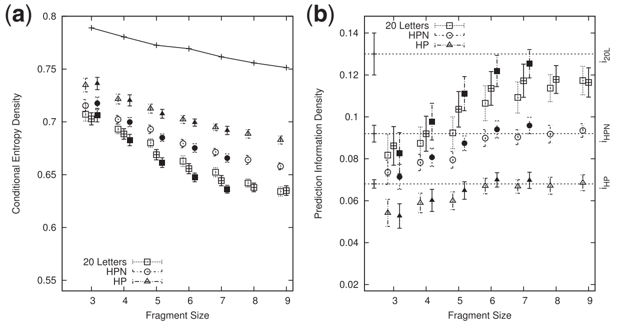

Fig. 5. For HMM results, the *density* of prediction information, i_p, can be calculated as the difference between an N-dependent estimate for the entropy density of burial levels, $h_N(B)$, Equation (S13) of the Supplementary Information (shown as a solid line in **a**), and an analogous estimate for the entropy density conditional to prediction, $h_N(B|B(Q))$, Equation (S14) of the Supplementary Information (shown as points in a). Resulting differences are plotted in (**b**) in comparison to the upper limit provided by the observed existing mutual information density between burials and sequences, $i(B; Q)$ (horizontal dashed lines). The results for C_β predictions are shown here. Analogous results for C_α predictions are shown in the Supplementary Information. Point symbols are encoded similarly to Figure 4

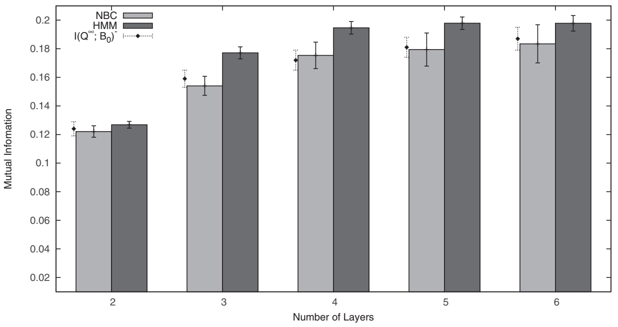

Fig. 6. As the number of discrete burial layers increases, the quality of prediction, as measured by the prediction information, I_p, also improves, at least up to four to five layers. The results are shown for NBC2 and HMM C_β predictions. Analogous results for C_α are shown in the Supplementary Information. Window size of 15 and fragment size of 7 were used for NBC and HMM, respectively. HMM predictions were performed with no additional descriptors to the hidden variables. Dotted error bars represent the estimated lower bounds for the mutual information between single burial and sequence of identities, $I(Q^\infty; B_0)^-$

layers is increased up to a number of 4. The rise in quality for five or six layers, however, is less significant, suggesting an upper limit for the number of layers into which it is useful to split a protein for burial-level prediction. As already observed for two burial layers, prediction information tends to be larger for C_β when compared with C_α atoms. Furthermore, $I(Q^\infty; B_0)^-$ values are also approached by NBC and surpassed by HMM predictions.

4 DISCUSSION

In this study, we estimate by extrapolation, neglecting long range correlations, the mutual information density between local sequence of amino acid identities and corresponding burials, $i(Q; B) \approx I(Q_0; B^\infty)$. It must be noted that the underlying probability distributions, estimated from local block statistics, are much simpler than distributions of whole amino acid sequences and tertiary structures. In particular, they are consistent with markovicity and a linear dependence of entropy, and mutual information, on block length, as shown in Figure 1. Meaningful densities of entropy and mutual information can be estimated for this simplified statistical scheme with different reduced alphabets. Additionally, and most importantly, resulting estimates for $i(Q; B)$ provide upper limits for the quality of prediction associating local sequences of burials and identities, a clearly attemptable task with established learning algorithms. Prediction of single burial values from local sequence, on the other hand, should be limited simply by the mutual information between local sequence and single burial, $I(Q^\infty; B_0)$, which is difficult to estimate for 20 amino acid letters due to databank saturation. We provide therefore a lower bound, $I(Q^\infty; B_0)^- < I(Q^\infty; B_0)$, further neglecting local correlations between amino acid identities conditional to single central burial. For the tractable HP alphabet, the difference between $I(Q^\infty; B_0)$ and $I(Q^\infty; B_0)^-$ is a single centibit, as shown in the Supplementary Information.

Single sequence statistical behavior, as summarized in Table 1, is qualitatively similar to what was previously observed for secondary structure by Crooks and Brenner (2004). While amino acid identities in local sequences appear to be statistically independent, short-range correlations are detected for the one-dimensional structural descriptor, either secondary structure or burial. Correlations between burials are stronger for C_α than for C_β atoms, as evidenced by smaller entropy density and larger mutual information between adjacent letters in the first case. This observation is likely to be at least partly associated to a longer distance along the sequence between adjacent C_β when compared with C_α atoms. As shown in Table 2, local sequence appears to be more informative about C_β than C_α burials, as indicated by larger values of $I(Q^\infty; B_0)^-$ and $i(Q; B)$ in the first case. Nevertheless, the proportional contribution to mutual information from local sequence beyond single residue identity appears to be larger for C_α when compared with C_β, as suggested by larger values for $I(Q^\infty; B_0)^-/I(Q; B)$ for the backbone atom.

Our estimates for the mutual information density, $i(Q; B)$, indicate that the uncertainty about burials that is resolvable from local sequence, already considering the reduction provided by sequence-independent burial local correlations, can be as small as 9 centibits/residue, as for two levels of C_α burials, and also at least as large as 18 centibits/residue, observed for three levels of C_β burials. These values are comparable to estimates involving secondary structure (16 centibits/residue; Crooks and Brenner, 2004), and are around 15% of the corresponding burial entropy density. Estimates for $i(Q; B)$ tend to be larger than for corresponding estimates for $I(Q^\infty; B_0)^-$, particularly for C_α atoms, in which case the difference is consistently between 1 and 2 centibits. It is suggested, therefore, that a couple of centibits of extra burial information might be extracted from sequences, in this case, when local burial correlations are accounted for. The effect on C_β atoms is smaller, again indicating a milder dependence of burial behavior from the side-chain atom on adjacent residues, either through their identities or burials.

Presently investigated burial levels, defined by equiprobable layers of central distances, display some qualitative similarity with burial levels defined from accessible surface areas, as reported by Crooks *et al.* (2004). Oscillations in positional mutual information observed in Figure 3, reflecting secondary structure exposure periodicity, are also observed in analogous plots involving burials in that previous investigation, although not for identities or secondary structure assignments. Notably, however, single amino acid identities appear to be more informative about accessible surfaces than about central distances. While single residue mutual information between identity and two bins of burials reported by Crooks *et al.* (2004), is 0.15 bit, our presently estimated value for $I(Q_{20}; B_{\beta 2})$ is only 0.07 bit, or about half of the corresponding density, $i(Q_{20}; B_{\beta 2})$, as shown in Table 2. Correlations between adjacent central distances, on the other hand, appear to be larger, as shown by larger values of $I(B_i; B_{i+1})$ in Table 1 when compared with values reported in that previous investigation.

These discrepancies might be partly associated to different procedures for determination of burial levels. While levels of accessible surfaces were explicitly determined from mutual information maximization, our levels of central distances simply maximize unconditional uncertainty. It is possible, nevertheless, that intrinsic physical differences between the two measures are also involved. Although correlated in globular proteins (Pereira de Araújo *et al.*, 2008), it is apparent that accessible surface area should be affected more directly by residue hydrophobicity while being somewhat less dependent on adjacent residues. It is not presently clear how much information could be expected from actual predictions of accessible areas from local sequence, since mutual information densities have not been reported. Weaker correlations when compared with central distances, however, are indicative of a less pronounced increase in prediction information with additional local environment beyond single residue. In any case, even if eventually more predictable than central distances, it remains to be shown if accessible areas can be as efficient in tertiary structure determination.

Our prediction results indicate that most of the burial information shared by local sequences is easily captured by simple statistical prediction schemes based on HMM or, to a lesser extent, NBC. Interestingly, $I(Q^{\infty}; B_0)^{-}$ is approached by the NBC algorithm, which neglects most identity correlations conditional to single burials, and actually surpassed by the HMM algorithm, which appropriately accounts for such correlations. Furthermore, near-optimal prediction for HMM algorithms is indicated by the corresponding mutual information density approaching our present estimate for $i(Q; B)$. From the results with reduced identity alphabets, it is apparent that only about half of the burial information extractable from local sequence using all 20 amino acid letters is still extractable when the HP-reduced alphabet is used instead. The significant improvement provided by the HPN alphabet, with just a single additional letter, indicates however that judiciously chosen reduced alphabets might still be useful in actual prediction, particularly in situations in which the size of the training set might become a limiting factor. In the opposite situation, when the training set is sufficiently large, prediction could be improved by increasing the number of burial levels, as indicated by Figure 6, or by including more hidden variables in the HMM.

In any case, independently of the size of the databank, burial prediction information is unavoidably restricted within a small fraction of the unconditional burial uncertainty, as provided by the density of mutual information between identities and burials, $i(Q; B)$. Even considering the possibility of judicious partitioning of the databank, such as according to chain size or structural class, the basic situation is unlikely to change significantly. As has been previously noted (Crooks and Brenner, 2004), a small amount of mutual information between local sequence and structural descriptors, when compared with the descriptor entropy density, indicates that local structure, as reflected in secondary structure or burials, must be largely determined by non-local information. It is useful, however, to distinguish between sequence-dependent and sequence-independent non-local information. After all, a large amount of structure-determining information is provided by sequence-independent constraints, analogous to grammatical rules of human languages (Pereira de Araújo and Onuchic, 2009). The information to be obtained from sequences, corresponding in the same analogy to the actual literature codified in written texts, should actually be much smaller. The distinction between sequence-dependent and sequence independent information is already apparent locally. The uncertainty of 1 bit for two burial levels of a single C_{α} atom, for example, diminishes to 0.6 bit due to sequence-independent local information, or a reduction of 0.4 bit, while around 0.1 bit is resolvable by sequence-dependent local information. A particularly interesting possibility, from the predictor's perspective, would correspond to sufficient sequence-dependent information for tertiary structure determination being exclusively local, while non-local information would be sequence-independent.

A large amount of sequence-independent non-local structural information is actually inferred from the small expected total number of protein shapes, Ω_s, which has been estimated by different groups to be in the order of several thousands (Chotia, 1992; Govindarajan *et al.*, 1999; Koonin *et al.*, 2002; Zhang and DeLisi, 1998). If Ω_s is assumed to be 10 000, for example, the corresponding entropy would be limited from above by $\log_2 \Omega_s$ and could not be more than around 13 bits per structure, or only 0.05 bit/residue for a putative typical length of 260 residues (0.1 bit/residue for 130 residues). This would be the uncertainty about whole structures, and therefore burials, to be resolved from sequence. The large remaining single burial uncertainty, e.g. $\approx (1 - 005 = 095)$ bits/residue for two C_{α} burial levels, must therefore be resolvable by sequence-independent information, both local (≈ 0.4 bits/residue, as discussed above) and non-local (≈ 0.55 bits/residue, as a consequence). Note that even if the total effective number of structures turns out to be larger or smaller by up to two orders of magnitude, the estimated amount of sequence-dependent structural information could not change by more than a couple of centibits/residue. It is interesting that an independent argument, based the thermodynamic stability of globular proteins, provided a compatible entropy estimate, $\approx$ 10–30 bits per macromolecule (Crooks *et al.*, 2004).

This small amount of sequence-dependent information (literature), when compared with the large amount of sequence-independent constraints (grammar), is an unavoidable consequence of a modest total number of structures when compared with possible sequences. It is also clearly consistent with the sound elusiveness of possible solutions for the problem of *ab initio* protein structure prediction, contrasting to significant success in homology modeling. Note that the entropy of whole amino acid sequences must indeed be much larger than structural entropy since many sequences fold to each single structure (Koehl and Levitt, 2002; Larson *et al*., 2002), although smaller, and less trivial, than estimated from local statistics. Long-range sequence correlations have been detected (Pande *et al*., 1994) and must produce deviations from markovicity, contributing not only to reduce the entropy but also to destroy its linear dependence on chain length. Crucially, in any case, the presently reported small information for burial predictions can still turn out to be sufficient for structural determination when combined to appropriate sequence-independent constraints.

5 CONCLUSION

Knowledge about atomic burial levels has been previously shown to be both sufficient for structural determination of small globular proteins and entropically compatible with amino acid sequences. Our present results, however, indicate that only a fraction around 15%, at least for C_α and C_β atoms, of burial uncertainty is resolvable by local amino acid sequence. On the bright side, most of this sequence-dependent burial information is easily extractable by simple prediction schemes, such as the presently implemented NBC and HMM. Most importantly, these predictions provide parameters for future folding simulations completely independent of knowledge about the native structure. The possibility of structural prediction of globular proteins from amino acid sequence using atomic burials as informational intermediates, including a possible combined improvement of sequence-independent constraints and burial prediction schemes, can now be investigated directly.

Funding: This research was supported by the Conselho Nacional de Pesquisa (CNPq), grant 478121/2011-3. J.R.R. received a graduate stipend from the Cordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES). M.G.V.L. received a graduate stipend from CNPq. D.C.F. and P.H.A. received undergraduate research stipends (IC) from CNPq. A.F.P.A. received a research stipend (PQ) from CNPq.

Conflict of Interest: none declared.

REFERENCES

Chotia,C. (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Cline,M. *et al*. (2002) Information-theoretic dissection of pairwise contact potentials. *Proteins*, **49**, 7–14.

Crooks,G.E. and Brenner,S.E. (2004) Protein structure prediction: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603–1611.

Crooks,G. *et al*. (2004) Measurements of protein sequence–structure correlations. *Proteins*, **57**, 804–810.

Dill,K.A. *et al*. (2008) The protein folding problem. *Annu. Rev. Biophys*., **37**, 289–316.

Efron,B. and Tibshirani,R. (1993) *An Introduction to the Bootstrap Monographs on Statistics and Applied Probability*. Chapman & Hall, New York.

England,J. (2011) Allostery in protein domains reflects a balance of steric and hydrophobic effects. *Structure*, **19**, 967–975.

Garnier,J. *et al*. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol*., **120**, 97–120.

Gomes,A.L.C. *et al*. (2007) Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. *Proteins*, **66**, 304–320.

Govindarajan,S. *et al*. (1999) Estimating the total number of protein folds. *Proteins*, **35**, 408–414.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci*., **3**, 522–524.

Huang,E.S. *et al*. (1995) Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol*., **252**, 709–720.

Koehl,P. and Levitt,M. (2002) Protein topology and stability define the space of allowed sequences. *Proc. Natl Acad. Sci. USA*, **99**, 1280–1285.

Koonin,E.V. *et al*. (2002) The structure of protein universe and genome evolution. *Nature*, **420**, 218–223.

Larson,S.M. *et al*. (2002) Throughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci*., **11**, 2804–2813.

Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol*., **14**, 70–75.

Pande,V.S. *et al*. (1994) Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proc. Natl Acad. Sci. USA*, **91**, 12972–12975.

Pereira de Araújo,A.F. and Onuchic,J.N. (2009) A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proc. Natl Acad. Sci. USA*, **106**, 19001–19004.

Pereira de Araújo,A.F. *et al*. (2008) Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. *Proteins*, **70**, 971–983.

Shakhnovich,E. (2006) Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev*., **106**, 1559–1588.

Solis,A. and Rackovsky,S. (2004) On the use of secondary structure in protein structure prediction: a bioinformatic analysis. *Polymer*, **45**, 525–546.

Solis,A.D. and Rackovsky,S. (2007) Property-based sequence representations do not adequately encode local protein folding information. *Proteins*, **67**, 785–788.

Sullivan,D.C. *et al*. (2003) Information content of molecular structures. *Biophys. J*., **85**, 174–190.

Zhang,C. and DeLisi,C. (1998) Estimating the total number of protein folds. *J. Mol. Biol*., **284**, 1301–1305.