Published online September 20, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 16 ***5343–5353***
doi:10.1093/nar/gki842

Gibbs sampling and helix-cap motifs

Erik Kruus[1,*], Peter Thumfort[1], Chao Tang[1,2,3] and Ned S. Wingreen[1,4]

[1]NEC Laboratories America, Inc., 4 Independence Way, Princeton, NJ 08544, USA, [2]Department of Biopharmaceutical Sciences and [3]Department of Biochemistry and Biophysics, California Institute for Quantitative Biomedical Research, UCSF Box 2540, University of California San Francisco, San Francisco, CA 94143-2540, USA and [4]Department of Molecular Biology, Princeton University, Princeton, NJ 08544-1014, USA

Received April 13, 2005; Revised August 8, 2005; Accepted August 30, 2005

ABSTRACT

Protein backbones have characteristic secondary structures, including α-helices and β-sheets. Which structure is adopted locally is strongly biased by the local amino acid sequence of the protein. Accurate (probabilistic) mappings from sequence to structure are valuable for both secondary-structure prediction and protein design. For the case of α-helix caps, we test whether the information content of the sequence–structure mapping can be self-consistently improved by using a relaxed definition of the structure. We derive helix-cap sequence motifs using database helix assignments for proteins of known structure. These motifs are refined using Gibbs sampling in competition with a null motif. Then Gibbs sampling is repeated, allowing for frameshifts of ±1 amino acid residue, in order to find sequence motifs of higher total information content. All helix-cap motifs were found to have good generalization capability, as judged by training on a small set of non-redundant proteins and testing on a larger set. For overall prediction purposes, frameshift motifs using all training examples yielded the best results. Frameshift motifs using a fraction of all training examples performed best in terms of true positives among top predictions. However, motifs without frameshifts also performed well, despite a roughly one-third lower total information content.

INTRODUCTION

Two secondary structures, α-helices and β-sheets, can be used to classify the secondary structure of ∼60% of the backbone residues in folded proteins. Individual amino acid residues are known to have different propensities to form one or the other of these structures. The ends or 'caps' of α-helices in natural proteins are also known to have specific amino acid propensities. The first systematic survey of helix caps, by Richardson and Richardson (1), identified preferred residues at specific sites of the N- and C-terminated caps of 215 α-helices. Later systematic studies of 1131 helices by Kumar and Bansal (2), and of 1316 helices by Aurora and Rose (3), reinforced the evidence for preferred residues, while Goliaei and Minuchehr (4) reported preferred neighboring residue pairs in the caps of a set of 2177 helices. A study of 8227 helices by Engel and DeGrado (5) reported position-dependent propensities throughout the length of a helix.

Residue preferences derive from the energetics of helix-cap formation. In particular, the first residues of the N-terminus have backbone N-H groups available for hydrogen bonding, while the last residues of the C-terminus have backbone C=O groups available for hydrogen bonding (see Figure 1). Satisfaction of these backbone hydrogen bonds by sidechains (1,6), as well as stabilizing hydrophobic interactions (3), and interactions with the helix dipole (1) may all contribute to amino acid preferences at the helix ends.

All these previous studies relied on specific, though somewhat different, structural criteria to identify helix caps. Ncap and Ccap, the helix-like residues at N- and C-termini of a helix segment, have commonly been assigned in terms of φ and ψ angles, hydrogen bonding, and α-carbon positions, possibly with restrictions on changes in the direction of the local helix axis. The set of examples satisfying each structural helix-cap definition gives rise to a sequence motif.

Unfortunately, the use of different structural criteria often result in different assignments of Ncap and Ccap. Indeed, different secondary-structure-assignment methods often disagree about the location of the ends of α-helices, β-strands and turns/random coils (7–14). Fourrier *et al*. (8) reported that between any two of five widely used structure assignment methods, the percentage of residues in agreement ranged

*To whom correspondence should be addressed. Tel: +1 609 951 2628; Fax: +1 609 951 2482; Email: kruus@nec-labs.com
Correspondence may also be addressed to Ned S. Wingreen. Tel: +1 609 258 8476; Fax: +1 609 258 8616; Email: wingreen@molbio.princeton.edu

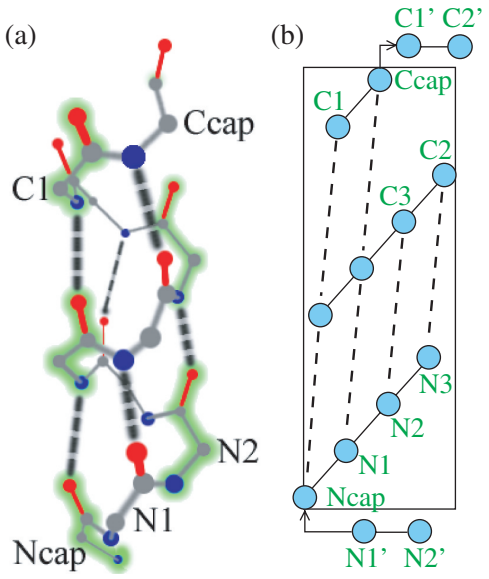

Figure 1. A figure showing helix caps, first in 3D and then in a projected helical-net representation. (**a**) Backbone of α-helix with N-terminus at bottom and C-terminus at top. According to convention, the last residue that forms a backbone hydrogen bond to the helix is denoted 'Ncap' at the N-terminus and 'Ccap' at the C-terminus. Alternate residues are outlined. (**b**) Helical-net representation of the same α-helix, showing Ncap and Ccap residues and their neighbors. Hydrogen bonds are indicated by dashed lines.

from 95% (STRIDE versus DSSP) to as low as 61% (DEFINE versus DSSP) and that much of this discrepancy was due to uncertainty about cap positions. Rost *et al*. (9) reported similar discrepancies. Some view structural uncertainty as an unavoidable consequence of the dynamic nature of proteins (9,10), while others have attempted to avoid uncertainty by defining distinct subpopulations of helices and helix caps (11–15).

Fortunately, the precise identification of Ncap and Ccap sites by rigid structural criteria is not strictly necessary for protein sequence analysis. For example, for both secondary-structure prediction and protein design, one may wish to know whether a particular sequence is likely to form a helix cap, with no need to precisely identify the Ncap and Ccap residues. In this scenario, the use of only structural criteria may result in weaker predictive power compared with a sequence-based method. We therefore decided to revisit the problem of helix-cap prediction using a 'structure-free' approach.

Specifically, starting from the structure-based helix-cap assignments in the Molecular Modeling Database (MMDB) (16), we used a Gibbs-sampling algorithm on cap sequences to self-consistently reassign Ncap and Ccap sites by shifts of up to ±1 residue. Following this methodology yields sequence motifs constructed to predict the approximate position of helix caps. In this demonstration of the algorithm, we train using only positive helix cap examples—for a global secondary structure predictor, please see the review (17). Our conclusion is that structure- and sequence-based methods are highly consistent, but that sequence-based methods, such as the one implemented here, offer some advantages for secondary-structure prediction and design.

METHODS

Protein datasets

Training datasets of varying size were provided by the non-redundant PDB (18) sequences according to an NCBI file (http://www.ncbi.nih.gov/Structure/VAST/nrpdb.html file nrpdb.032002), which arranges non-redundant PDB protein chains according to BLAST similarity scores. A representative set of dissimilar protein sequences is successively subdivided to form a tree structure. At the lowest level of this tree, one representative PDB entry was provided for each of 10 888 distinct protein sequence. The top level of this tree uses a value of 10^{-7} for the BLAST *p*-value as a similarity cutoff. This top level provided a representative set of 2414 protein sequences (~8600 helices). We report results using these representative sequences with equal weights. All calculations were repeated using the full set of non-identical protein sequences as well (with example weights adjusted for the tree fan-out) with only slight changes in the results. For the representative set, scoring and assigning length 7 (heptamer) motifs required ~455 000 windows into the sequences, while the full set required ~2 million sequence windows. Our converged motifs are typically the result of 24–48 h of run time on a 933 MHz PC.

Secondary-structure assignments

We considered three secondary structure states: helix 'H', sheet 'S' and other (turn 'T' or random coil '.'). We used the MMDB secondary structure assignments rather than the original PDB assignments because the MMDB assignments seemed more conservative. For example, the PDB assignments included helices too short to support an intra-helix hydrogen bond. The MMDB helix assignment uses the Protein Knowledge Base programs and database of (19) (ftp://ftp.ncbi.nih.gov/pub/pkb/). Similar to the algorithm of Richardson and Richardson (1), these helix assignments are based on C_α distance constraints. For MMDB helices, C_α–C_α distance matrix elements for eight consecutive residues had to be within 0.2 Å of those of a canonical helix (see Supplementary Data for the distance matrix of a canonical helix). As such, we are working with helices whose axes are close to linear over a span of eight residues. Caps used for our training were required to have at least four helix residues flanked by at least 2 turn/random coil assignments. This resulted in 8621 training examples for Ncaps from the representative set of sequences and 39 843 from the full set, and approximately equal numbers of Ccaps.

We used the MMDB structural definition of helix caps as a starting point in the search for strong helix-cap sequence motifs. We also used the MMDB cap assignments to judge how well our sequence motifs predict the approximately correct (±1 residue) location of helix caps. In principle, any reasonable structural definition of helix caps (e.g. the DSSP or PDB one) could equally well have been used as both a starting point and a comparison standard for our sequence-based algorithm of identifying motifs.

A reduced number of secondary structure classes were defined using regular expressions. These classes were used to see the structural features that were actually predicted by sequence motifs, and to evaluate ROC (receiver operator characteristic) scores and TPF (true positive fraction) scores, as

Table 1. Performance of length 7 MMDB Ncap and Ccap motifs, $p_i^{all}(r)$, measured by ROC scores (×100) for a range of secondary-structure classes

Motif	Turn	Hloop	Sloop	N2′	N1′	Ncap	N1	Helix	C1	Ccap	C1′	C2′	SheetN2′	SheetN1′	Nsheet	SheetN1	Sheet	SheetC1	Csheet	SheetC1′	SheetC2′
Ncap	51	50	55	55	**60**	**76**	**65**	52	36	36	40	48	57	**62**	59	56	42	36	44	48	50
Ccap	42	57	49	44	41	38	39	**62**	**75**	**81**	**67**	56	42	39	36	34	48	56	55	49	43
Null	50	51	49	50	50	50	50	50	50	50	50	50	50	50	50	50	49	50	50	50	50

Motifs were directly determined from the 8621 MMDB Ncaps and 8735 Ccaps, without using Gibbs sampling. In testing, the Ncap motif, Ccap motif and null motif competed for sequences in each class; we used strict MMDB testing, i.e. no sequence frameshifts were allowed. ROC scores of 0.6 or more are in bold typeface.

described below. Each residue was uniquely assigned to 1 of the 21 classes (Table 1). Following (3), Cn' (Nn') denotes a residue n outside the C (N) terminus of a helix segment, while Cn (Nn) denotes a helix residue n away from a cap position. A helix- or sheet-cap residue was a first (Ncap, Nsheet) or last (Ccap, Csheet) helix or sheet residue in a contiguous segment of such residues. Hloop and Sloop denote short turn sequences between helix or sheet assignments and serve to avoid ambiguous assignment of a turn/random coil residue as both a Cn' of a preceding structure and an Nn' of the following one.

Gibbs sampling

The general approach to the local multiple-alignment problem is to assign regions in each input sequence to 'motifs' in such a way that the set of assignments maximizes a function. In one approach, a set of N assignments defines probabilities $p_i(r)$ for sequence elements r to occur at each position i, and the maximized function is the information content

$$\mathrm{IC} = N \sum_{i=1}^{L} \sum_{r \in \Sigma} p_i(r) \log \frac{p_i(r)}{\pi_i(r)}. \qquad \mathbf{1}$$

In our case, the motif length L describes a contiguous sequence of residues r from an alphabet Σ of 20 amino acids. Here, $\pi_i(r)$ is an a priori probability for residue r given position i. As in the Gibbs sampler of (20) we used fixed priors, usually a site-independent $\pi(r)$. Gibbs sampling is the preferred algorithm to efficiently converge on a high-quality solution to the local multiple-alignment problem of maximizing Equation 1 (21). The output of the Gibbs sampler is a statistical model, or motif, which is the most informative one describing a set of input sequences.

We implemented a version of Gibbs sampling in C++ following the basic algorithm in (22), in which further mathematical details may be found. Our cap motifs are weight matrices of $p_i(r)$ values. Sequence windows not assigned to a motif are handled by assigning them to the 'null motif' whose probabilities $p(r)$ are equal to the prior probability $\pi(r)$. Our prior probabilities were derived from the amino acid distribution over the set of input sequences from which caps were selected.

First-order motifs

For a given sequence window $\{R_i\}_{i=1..L}$, the first-order probability is $P^{(1)} = \prod_{i=1}^{L} p_i(R_i)$ and the score for one of our motifs $p_i(r)$ is

$$\mathrm{Score} = \frac{\prod_{i=1}^{L} p_i(R_i)}{\prod_{i=1}^{L} \pi(R_i)}. \qquad \mathbf{2}$$

Gibbs sampling operates by assigning sections of input sequences (windows) to competing motifs probabilistically, giving motifs with higher scores a higher chance to be assigned a particular sequence window. As soon as the assignment is made, that motif is updated, so that its subsequent scores better reflect its current set of assigned sequence windows. The algorithm iterates, continually reassigning sequence windows to motifs and allowing the motifs to change. Generally, the motifs converge to comprise stable subsets of sequence windows, each of which 'matches' a single motif. By default, all motifs compete on equal footing; however, for some of our results we artificially favored certain motifs in order to control the total number of assigned sequence windows (23).

Second-order motifs

To test the predictive power of more general motifs, we extended the algorithm to support arbitrary motif models. In particular, our program also supports second-order motifs [site–site–residue–residue probabilities $p_{ij}(rs)$] and second-order priors, similar to the treatment of (24). In a second-order motif, for sites i and j, and residues r and s, the score for a sequence is based on the probability given by Equation (3),

$$P^{(2)} = \frac{\prod_{1 \leqslant i < j \leqslant L} p_{ij}(rs)}{\prod_{i=1}^{L} p_i(r)^{L-2}} \qquad \mathbf{3}$$

$$= \underbrace{\prod_{i=1}^{L} p_i(r)}_{\text{First order model}} \cdot \prod_{1 \leqslant i < j \leqslant L} \alpha_{ij}(rs), \qquad \mathbf{4}$$

where

$$\alpha_{ij}(rs) = \frac{p_{ij}(rs)}{p_i(r) p_j(s)}. \qquad \mathbf{5}$$

All quantities in $P^{(2)}$ may be evaluated by counting site–site–residue–residue occurrences. Equation 4 illustrates how a second-order motif can be viewed as a first-order motif modified by second-order corrections $\alpha_{ij}(rs)$. Values of $\alpha_{ij}(rs)$ above (below) 1.0 signal particular residue pairs occurring more (less) frequently than the underlying first-order motif would predict.

Z-scores for site–site–residue–residue pairs were calculated as follows. Given a set of examples defining a motif, the expected number of counts $e_{ij}(rs)$ is the product of a prior probability (see below) and the total number of examples, with an assumed standard deviation of $\sqrt{e_{ij}(rs)}$. The Z-score is the number of standard deviations away from the expected number of counts. For example, if we observe $o_{ij}(rs)$ counts, then the Z-score is $\left[o_{ij}(rs) - e_{ij}(rs)\right] / \sqrt{e_{ij}(rs)}$. Z-scores for second-order motifs may be calculated using

the underlying first-order model $p_i(r)$ as a prior. The highest Z-scores signal the most significant deviations from the first-order motif. Chi-squared estimates for various motifs were calculated as sums of squares of Z-scores.

Motif scores for Gibbs sampling are $P^{(2)}$ values divided by prior probabilities. As in Equation 2, we used the site-independent prior probabilities $\pi(r)$ derived from the global amino acid distribution. Results defined using priors from helical regions and second-order priors, and details about our implementation, are available in Supplementary Data.

Training helix-cap motifs

Two types of motif training were implemented. In the first, we used the MMDB assignments of Ncap and Ccap positions. Gibbs sampling was used to assign each helix cap to the constantly updated cap motif or to the fixed null motif, until convergence. We refer to motifs trained in this way as Gibbs–MMDB motifs. Ncap and Ccap motifs were trained separately.

In the second type of motif training, we used the MMDB assignments only as an approximate guide to the Ncap and Ccap positions. This was carried out by assigning three possible windows to every cap: a 0, −1 and +1 'reading frame', where 0 is the MMDB assignment. Gibbs sampling was then used to determine both the frame assignment and assignment to the cap motif or null motif in a self-consistent manner. We call motifs trained in this fashion frameshift motifs. Adding the flexibility of frameshifts can only result in stronger sequence motifs. The price paid is that we are allowed to deviate from the original structural assignment.

For both Gibbs–MMDB motifs and frameshift motifs, we separately trained motifs of lengths 7, 9, 11 and 13 residues. For training Gibbs–MMDB motifs, windows were centered at the MMDB Ncap/Ccap positions, while ±1 shifts were allowed in frameshift-motif training. For all trained motifs, we verified convergence from multiple motif initializations.

Testing helix-cap motifs

While training used sequence windows only at cap regions, all sequence windows were used for testing. To quantify the ability of our motifs to predict helix caps, we used the method of ROC (25). To generate an ROC curve, one first ranks all the scores of a classifier from highest to lowest. As the rank number increases (decreasing score), one plots the cumulative number of true positives versus the cumulative number of false positives. Axes are renormalized to [0,1] and the integral under the ROC curve is the ROC score. The ROC score for a perfect predictor is 1.0, and the ROC score for a random predictor is 0.50.

Since we trained both Ncap and Ccap motifs, when testing we typically ran an Ncap motif, a Ccap motif and the null motif simultaneously, assigning a given sequence window to the motif with the highest score given by Equation 2.

Similar to motif training, two types of motif testing were implemented. In the first type of testing, sequences were assigned to the top-scoring motif, and the MMDB assignments were taken to be the true helix caps for purposes of determining true positives. We refer to this as MMDB testing. In the second type of testing, regions that were not helix caps were treated the same as above. However, MMDB helix caps were scored against each competing motif three times, for frameshifts 0, −1 and +1. The highest of these scores was used to assign the sequence to a motif; assignments frameshifted by 0, −1 or +1 from the MMDB assignment were all considered to be true positives. We refer to this as frameshift testing. Frameshift testing allows us to test how well our sequence-based frameshift motifs predict proximity to a structure-based helix-cap assignment.

In addition to generating ROC scores, we also evaluated the fraction of true positives among the top 500 test examples, TPF_{500}. While the ROC score reflects the global predictive power of a motif, the TPF_{500} reflects the success of the motif in correctly predicting those sequences most likely to be helix caps.

RESULTS

MMDB-motif performance

We first report a straightforward approach, without Gibbs sampling, using the statistics of all MMDB helix caps in the representative set of sequences to define Ccap and Ncap motifs, i.e. probabilities for residue r at helix-cap position i, $p_i^{\mathrm{all}}(r)$. We refer to these structurally derived motifs as MMDB motifs. We tested the ability of MMDB motifs to identify their corresponding helix caps, and, as a control, to identify other secondary-structure elements as well. In the test procedure, the Ccap motif, the Ncap motif and the null motif competed for sequences from various secondary-structure classes (according to MMDB annotation). Each sequence window was assigned to the motif with the highest score given by Equation 2. This left a set of sequence windows with various MMDB secondary-structure annotations assigned to each of the three motifs. The predictive ability of each motif for each secondary-structure class is shown in Table 1 as ROC scores (see Methods). Reassuringly, the MMDB motifs' ROC scores were highest for the cap each motif was trained to predict.

Secondary-structure classes with ROC scores <0.5 indicate that the cap motif discriminates against such secondary structures. Conversely, high ROC scores, other than for the trained class, indicate secondary structures that may confuse the cap motif. Table 1 shows that the Ncap motif is sometimes confused by SheetN1′ sheet-cap sequences (TTTSS secondary structure), and that the Ccap motif can be confused by helix regions. For all secondary-structure classes, throughout all our testing, the null motif always had ROC scores of 0.50 ± 0.02.

Interestingly, both MMDB motifs fared well as predictors of cap positions close (±1) to the positions assigned by the MMDB. In other words, the MMDB motif is already a reasonable predictor of proximity to a structurally defined helix cap. We will find frameshift sequence motifs to be even better predictors of such proximity.

Table 2 shows ROC scores for sequence motifs of different lengths. In all cases, ROC scores increased as motif length changed from 7 to 13 residues. Again, offsets of ±1 residue from the MMDB helix-cap assignments also have significant ROC scores.

We checked the ability of sequence motifs to generalize to unseen examples by testing on sequences from the full dataset.

Table 2. Performance of various length MMDB Ncap and Ccap motifs, measured by ROC scores (×100)

Motif length (residues)	Ncap motif			Ccap motif		
	N1′	Ncap	N1	C1	Ccap	C1′
7	60	76	65	75	81	67
9	65	78	69	77	83	71
11	68	80	72	78	84	73
13	70	82	75	79	85	75

MMDB motifs were defined, without Gibbs sampling, using all the helix caps in the representative set of sequences, as in Table 1.

ROC scores for testing on the full dataset were essentially indistinguishable (±1%) from those in Table 2.

Next, we used Gibbs sampling to train helix-cap motifs in competition with a null motif. This approach assigns to the null motif all residue sequences more likely to be drawn from the background distribution than from the cap motif, and these sequences do not contribute to the definition of the cap motif. In unbiased competition with the null motif, the Ncap motif used 52% of training examples to define itself while the Ccap motif used 59%. For the Gibbs–MMDB motifs trained in this way, the general trends were identical to those in Table 2; however, ROC scores, which measure overall predictive success, were reduced across the board by 1–6%.

While including the null motif in the training slightly reduced the overall ROC scores, these Gibbs–MMDB motifs yielded more correct predictions among the highest-scoring sequences. For the untrained MMDB motifs of Table 1, the 500 top-scoring heptamers from the representative set included 35% (40%) correctly predicted Ncaps (Ccaps). For the Gibbs–MMDB motifs, these success rates were increased to 42% (47%). Henceforth, all motifs we consider are generating by Gibbs sampling using a null motif during training.

Frameshift motif performance

We noted in Tables 1 and 2 that the MMDB motifs were often good predictors for helix caps shifted by ±1 residue from the MMDB assignment. This suggests that some of the MMDB structure-based assignments might be improved, from the point of view of consistency with sequence motifs, by a frameshift of ±1 residue. We therefore used Gibbs sampling to train frameshift Ncap and Ccap motifs (see Methods). Since each motif was trained in competition with the null motif, the fraction of helix-cap examples used to define each motif was variable [controlled by an expected prior fraction and 'belief in prior' parameters (23)].

Figure 2 shows that about half the Ncaps and two-thirds of the Ccaps have been shifted from the MMDB assignment. This may be a reflection of the higher backbone entropy of the Ccap, as shown in Brasseur's analysis of the allowed ϕ, ψ regions in Ramachandran plots of the DSSP assignments (26). For all numbers of defining examples Ncaps remained preferentially at the positions assigned in the MMDB; however, Ccaps were most often frameshifted by −1 (toward the helix interior) compared with the MMDB assignments.

In Figure 3, we compare the performance of frameshift motifs with Gibbs–MMDB motifs (no frameshift) as a function of the number of examples used to define the motifs. For all motifs, we used frameshift testing, i.e. a cap assignment within ±1 of the MMDB assignment was considered correct.

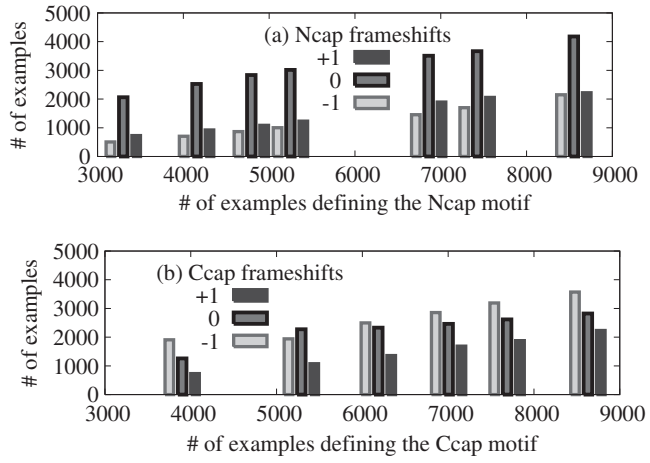

Figure 2. Frameshift distributions as a function of the total number of examples selected to define a frameshift motif: (**a**) Ncap motif, (**b**) Ccap motif.

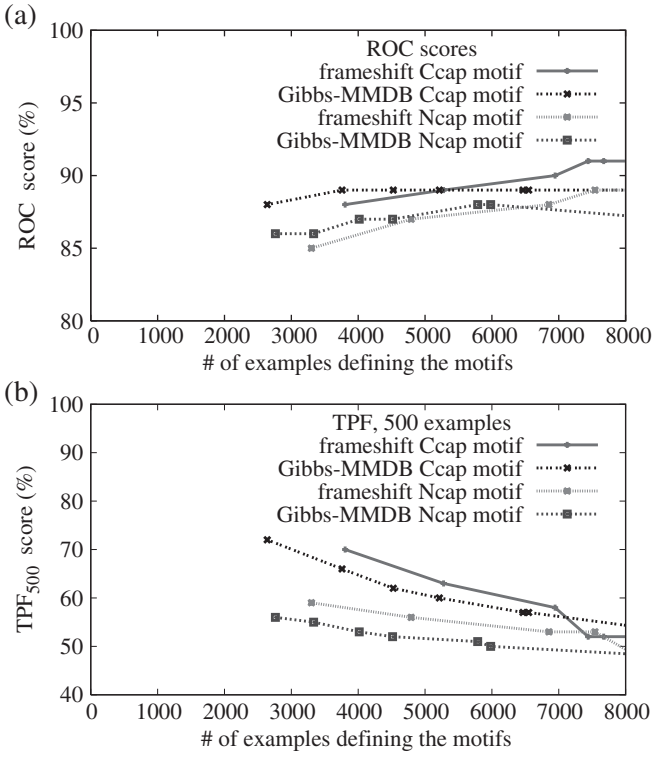

Figure 3. (**a**) ROC and (**b**) TPF_{500} scores as a function of the number of training examples selected by Gibbs sampling to define each length 7 motif. The number of examples selected was controlled by varying the expected prior fractions of the competing motifs. For all motifs, frameshift testing was employed, i.e. assignments within ±1 residue of MMDB cap assignments were considered correct.

Figure 3a shows ROC scores for the various motifs. ROC score differences between frameshift motifs and Gibbs–MMDB motifs are small. All ROC scores increase slightly with number of examples—as might be expected since ROC scores measure the global performance of the motifs on all sequences. In contrast, Figure 3b shows that frameshift motifs

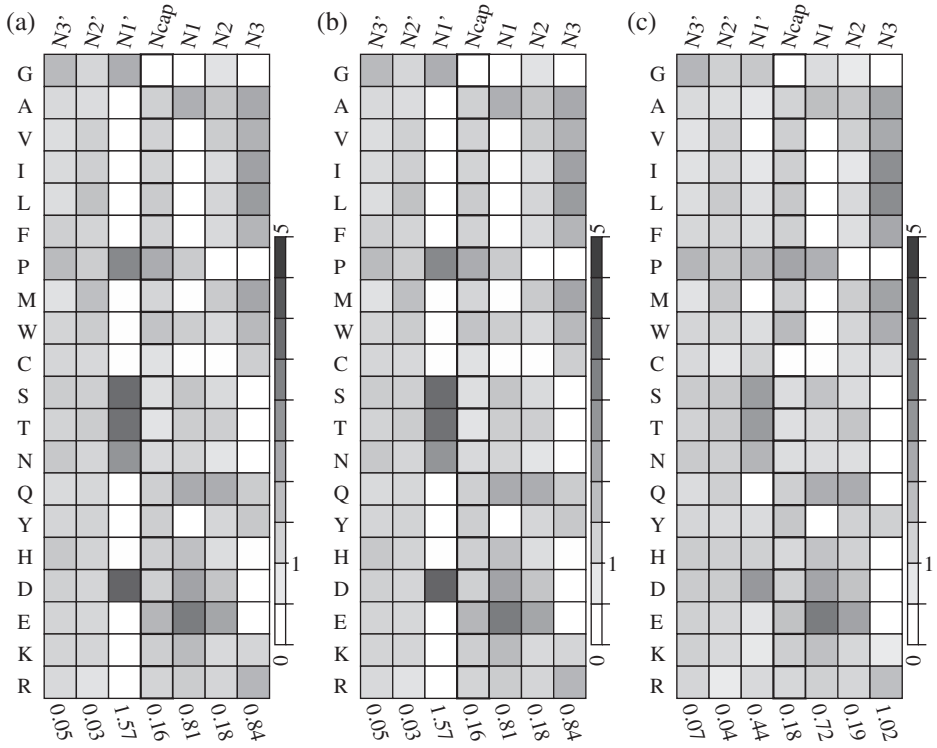

Figure 4. Examples of Ncap motifs. Shown in grayscale is p(residuelsite)/π(residue), i.e. the normalized probability for each amino acid-residue type for a 7-residue Ncap motif. π(residue) is the proportion of each residue type found in the entire representative set of proteins. Total information content per site (in bits) is shown along the bottom of the matrices. During motif training, competition with the null motif selected a self-consistent set of Ncaps to define each motif. The Gibbs–MMDB motif in (**a**) is defined by 4769 MMDB Ncaps selected from a total of 8551 via an unbiased competition with the null motif. The frameshift motifs in (**b**) and (**c**) are defined by MMDB Ncaps allowing a ±1 residue shift: in (b), the competition with the null motif was biased to select approximately the same number of Ncaps (4792) as in (a). In (c), unbiased competition with the null motif led to selection of 6832 Ncaps.

perform significantly better than their no-frameshift counterparts in correctly identifying helix caps among the 500 top-scoring sequences. All TPF_{500} scores decrease with number of examples. (TPF_{100} scores behave similarly. We report TPF_{500} scores to include more sequence diversity.) Note that Ccap ROC and TPF_{500} scores are consistently better than corresponding Ncap scores. As expected, motifs defined using small subsets of sequences do better in predicting the best positive examples (high TPF_{500}), but such highly specific motifs generalize less well (poorer ROC scores).

Examples of generated motifs

In this section, we present some comparisons between sequence motifs generated by our Gibbs-sampling algorithm and the sequence motifs derived from the MMDB structural assignments. More extensive sequence motif comparisons (as well as the motif datasets themselves) may be found in the Supplementary Data. There we also present four published sequence motifs corresponding to rigid structural criteria, as well as sequence motifs corresponding to MMDB and PDB structural assignments, and compare these with our frameshift motifs.

In Figures 4 and 5, we show examples of motifs for Ncaps and Ccaps, respectively. Panels (a) show Gibbs–MMDB motifs, i.e. with no frameshift from the MMDB assignments, while panels (b) and (c) show frameshift motifs using increasing fractions of defining examples. For both helix caps, a frameshift motif with approximately the same number of defining examples as the Gibbs–MMDB motif [(b) versus (a)] increased the total information content by ∼33%. For all motifs, the information content per site, $\Sigma_r p_i(r)\log_2[p_i(r)]/[\pi(r)]$, is small for the 2′ and 3′ turn sites. In both Figures 4 and 5, all motifs seem to display an alternation of information content as one progresses from the turn region into the helix.

In Figure 4, comparing panels (a) and (b) shows that allowing frameshifts reinforced the strongest N1 and N1′ features: the frameshift motif in Figure 4b consists almost exclusively of DSTPNG (i.e. small/polar residues) at the N1′ site.

Similarly, comparing panels (a) and (b) for the Ccaps in Figure 5 shows that allowing frameshifts (b) produced more 'forbidden' residues (white squares) at positions C3 through C1′. The strong preference for a G at C1′ in panel (a) has been lowered in panel (b) in favor of a more precise description of sites C1 and C3 in the helix portion. For example, the C3 site consists almost exclusively of the hydrophobic residues AVILFMWY.

How can we quantify the changes introduced in motifs by frameshifting? Motif probabilities reflect the energetics of cap formation. The physically important entries in a motif can be characterized by the direction $\hat{E}$ of a vector comprised only of the favorable (stabilizing) interaction energies, with all the unfavorable interaction energies set to zero, $E_i(r) = \min\{-\log[p_i(r)/\pi(r)], 0\}$. This neglect of unfavorable interaction energies, i.e. low counts, is essential because Gibbs sampling

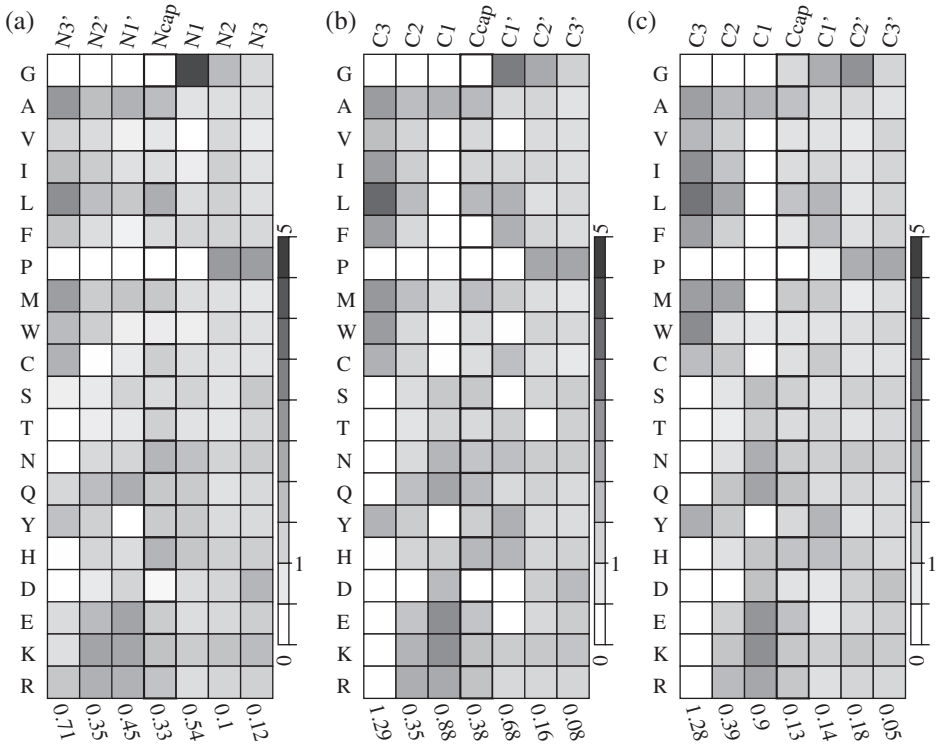

Figure 5. Examples of Ccap motifs (cf. Figure 4). The Gibbs–MMDB motif in (**a**) is defined by 5337 MMDB Ccaps selected from a total of 8628 via an unbiased competition with the null motif. The frameshift motifs in (**b**) and (**c**) are defined by MMDB Ccaps allowing a ±1 residue shift. In (b), the competition with the null motif was biased to select 5289 Ccaps, approximately the same number as in (a). In (c), unbiased competition with the null motif led to selection of 6854 Ccaps.

drops many real cap sequences from the final count. We can compare two motifs A and B by measuring the overlap between their stabilization-energy vectors, $\hat{E}_A \cdot \hat{E}_B$. In Figures 4 and 5, the overlap between the Gibbs–MMDB motif [panel (a)] and our frameshift motif [panel (b)] is 0.95 for Ncaps and 0.89 for Ccaps. To put these overlaps in perspective, random shuffling of the residues at each site of the frameshift motifs yielded overlaps of ∼0.2.

Adequacy of motif description

An important question is whether our motifs, which treat each site in a helix cap as independent, provide an adequate description of naturally occurring caps. Expectations from our first-order motifs for site–site–residue–residue pairs were compared with observed counts (see Methods). The chi-squared values for full second-order motifs were only ∼10% higher than the number of degrees of freedom, indicating that residue frequencies at particular sites are largely independent.

However, *Z*-scores did indicate a small subset of significant pairwise correlations. We obtained second-order motifs using the set of defining examples from the frameshift motifs of Figures 4b and 5b. In this analysis, high positive *Z*-scores indicate residue pairs that occur more frequently than one would expect from the first-order motif. This suggests that the residues 'cooperate' to favor helix-cap formation. Negative *Z*-scores (anticorrelations) indicate residue pairs that occur less frequently than expected. The statistically significant overrepresented site–site–residue–residue pairs are listed in Table 3. For Ncaps, 5 of the 9 correlations with *Z*-scores > 4 had a significant number of expected counts (>4). The strongest significant Ncap anticorrelation is $A_{N1}I_{N3}$, followed by proline–proline neighbors within sites N2′–N1. For Ccaps, 8 of the 23 correlations with $Z > 4$ had a significant number of expected counts. The two strongest significant Ccap anticorrelations (*Z*-scores < −3.5) involved neighboring G, P residues in sites C1′–C3′, suggesting that either G or P alone in the appropriate site is sufficient to favor Ccap formation.

Table 3. Most overrepresented site–site–residue–residue pairs, obtained by comparing second-order motifs to the expectations of the first-order frameshift motifs of Figures 4c and 5c, using the same sets of defining examples

Feature	Overrepresented site–site–residue–residue pairs			
Ncap	$T_{N1'}E_{N2}$	$I_{N3'}H_{N2'}$	$L_{Ncap}P_{N1}$	$F_{N2}N_{N3}$
	$N_{Ncap}T_{N2}$			
Ccap	$G_{C2'}V_{C3'}$	$W_{C2}E_{C1'}$	$F_{C3}S_{C1}$	$H_{C3}L_{C1}$
	$D_{C2}Y_{Ccap}$	$Y_{C3}K_{C3'}$	$G_{C2'}L_{C3'}$	$E_{C1}Y_{C1'}$

Overrepresented pairs with *Z*-score > 4 and expected to have 4 or more counts are reported.

Little overlap was found with the significant pairs identified by Goliaei and Minuchehr (4): they reported deviations of the number of neighboring residue–residue pairs in caps from expectations based on counts of the same residues occurring as neighbors anywhere within their helix database, while we addressed deviations from our first-order motif description of the caps themselves. Even if helices and caps were each well described by first-order motifs, one would still expect to find,

as they did, that nearest-neighbor frequencies in caps and helices differ. We supply the full second-order motifs and *Z*-score analyses (using various cap assignments and priors) in Supplementary Data.

In terms of predictive ability, the full second-order motifs (with or without frameshifts) showed only small improvements in ROC and TPF_{500} scores. An intermediate description in terms of a mixture of first-order motifs (22) was also investigated. Predictive ability in mixture models showed only modest improvements, and peaked or leveled off between four and six components, indicating that further components were likely overfitting the training data. Use of a global second-order prior, instead of the global amino acid frequencies, also led to only modest improvement in ROC and TPF_{500} scores.

Adequacy of training set

We investigated how using a larger, weighted training set influenced predictive ability. Motifs (of all types and of lengths 7–13) defined using the full set of sequences showed only minor improvement in predictive ability over those motifs obtained using the representative set of protein sequences, even when testing with the full dataset.

DISCUSSION

Helix-cap sequence motifs have traditionally been identified by first applying structural criteria to identify Ncap and Ccap sites (1–3). Amino acid frequencies are then calculated for these and nearby sites. However, many applications of cap motifs, e.g. secondary-structure prediction and protein design, do not require precise structural identification of the Ncap and Ccap sites. Such applications may benefit from a sequence-based approach to finding cap motifs which is less dependent on the use of rigid structural criteria. Here, we reported such an approach in which a structural definition of helix caps was used as a starting point for a sequence-alignment-based definition of Ncap and Ccap motifs. The motifs defined in this way were significantly more informative, as measured by their ability to predict helix caps, than traditional motifs based solely on structural considerations. Indeed, our sequence-based approach is foreshadowed in the original work of Richardson and Richardson (1) in which their particular structural definition of the Ncap and Ccap sites was chosen to '... give the strongest and most position-specific amino acid preferences'.

Our sequence-based approach to finding cap motifs combined two distinct novel components: frameshifts and Gibbs sampling. Frameshifts of ±1 residue were allowed with respect to the MMDB structural assignments of Ncap and Ccap sites, if these frameshifts increased the information content of motifs. In this way, caps were aligned by sequence in addition to structure. Gibbs sampling, in competition with a null motif, was employed to select a self-consistent set of sequences to define each cap motif. This selection of only the best cap sequences is justified by the collective nature of protein folding: some sequences form caps based on strong free-energy preferences, while others are forced to form caps by global folding constraints.

The optimal fraction of sequences used to define a motif depends on the intended application. From Figure 3 it is clear that using more sequences gives better overall predictions (ROC scores) while using fewer sequences gives better top predictions (TPF_{500} scores). For protein design, where the objective is to identify a set of highly stable capping sequences, motifs defined by fewer sequences will perform best. Specifically, for a competition among multiple secondary-structure motifs, it is most important that each motif's top predictions (i.e. those most likely to be awarded to that particular motif) are correct. Therefore, helix-cap motifs defined by a restricted set of sequences are likely to be well suited for use within full secondary-structure predictors. Moreover, some of the ambiguities found with our motifs are also likely to be cleared up by the inclusion of motifs for additional secondary-structure classes. For example, the confusion of the MMDB Ncap motif by SheetN1′ sequences and of the MMDB Ccap motif by helix sequences (Table 1) might be eliminated by competition with an Nsheet motif and a helix motif, respectively.

While competition among motifs may improve cap predictions within the context of entire proteins, our intent was to identify sequences which tend to form caps independent of context. We therefore focused on short motifs, in order to separate cap-forming tendency from helix-forming tendency. As shown in Table 2, longer motifs resulted in higher ROC scores; however, the improvement beyond seven residues is due almost entirely to information from the helical end of the motif, i.e. the longer motifs work better by identifying helices, not doing a better job identifying helix caps.

Comparison to previously described motifs

In Figure 6 we display, for ease of reference, frameshift motifs alongside structurally defined motifs from several sources, including motifs derived from PDB assignments. Bear in mind that our ±1 frameshift motifs constitute sharpened versions of their original MMDB structural assignments, and as such should not be used to judge the 'quality' of any particular geometric definition. More complete side-by-side comparisons of motifs [PDB, MMDB and cap motifs from articles (1–4)] may be found in Supplementary Data.

Our frameshift Ncap motif bears considerable resemblance to the sequence motif corresponding to Kumar and Bansal's structurally defined Ncaps (2) (DSSP cap assignments, adjusted so that O_i N_{i+4} distances were $\leqslant$3.5Å and with changes in the local helix axis restricted to <20°). Both motifs capture the well-known STNDGP preference at the first non-helical residue (as do the MMDB and PDB assignments, although not as strongly).

For the Ccap, our frameshift motif resembles a sharpened version of the original MMDB or PDB assignments as well as the motif of (4) (based on DSSP assignments). All favor G at C1′ with P favored at C2′. Richardson and Richardson (1), using fewer sequences, found Ccap probabilities that resemble our C1′ probabilities.

Aurora and Rose (3) proposed 6–7 structural motifs, with hydrophobicity patterns, for each helix cap. Our single motifs may correspond to linear combinations of their patterns. For our mixture models (see Supplementary Data), the strongest concurrence with Aurora and Rose was for a motif placing G at the C1′ position. However, instead of a separate proline-only motif, we found two motifs with

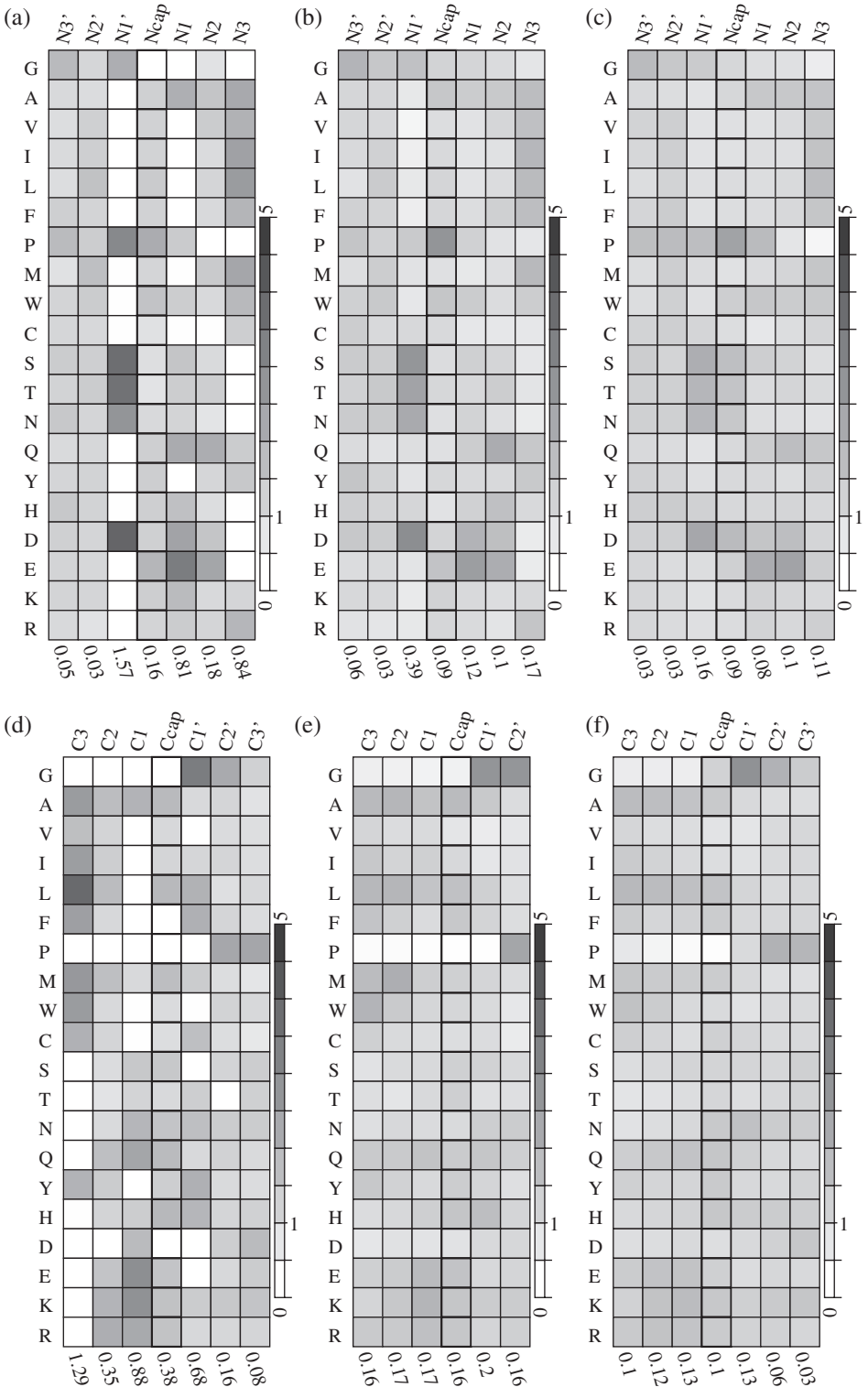

Figure 6. Comparison of frameshift sequence motifs (**a** and **d**) alongside sequence motifs based on rigid structural criteria. Ncap and Ccap are final helix-like residues. The global amino acid prior was used for all figures. (**b** and **e**) are derived from data in references (2,4) (details are in Supplementary Data). The final column (**c** and **f**) shows motifs derived from PDB cap assignments by selecting subsequences whose secondary structure consisted of four helix assignments adjacent to three non-helix assignments.

PGND at the Ncap. These two motifs were distinguished by different residue probabilities within the helix. The best-performing Ncap component had ST at N1′ with EDQ at N1 or N2.

Structures of top-scoring sequences

What do the structures of our top-scoring sequences really look like? We viewed hundreds of the top-scoring sequences

for length 7 motifs. Our true positives, by definition, are near helix terminations as annotated by the MMDB. Interestingly, for many false positives we found a single hydrogen bond between two residues in a loop-like structure. In these cases, our motifs confused isolated loops with helix terminations. Indeed, within the top-scoring Ncaps, >90% had at least one helix-like loop, even though the MMDB TPF_{500} values suggested that only 60% form real Ncaps. This means that many of our 'false positives' may in fact have strong cap-forming tendencies. Wider motifs, or a global analysis (27) (particularly one including an explicit helix predictor), might help correctly identify these helix-like loops.

In our approach, all structural caps were treated equally. However, many of the reported capping structures involve turn residues folded over to supply H-bonds or stabilizing hydrophobic interactions. Such caps are structurally unambiguous: frameshifts of ±1 are structurally 'far'. It might be useful to avoid frameshifts for structurally unambiguous caps. Beyond this, a next step might be to simultaneously and self-consistently cluster by both sequence and structure, leading to joint motifs as in (3).

CONCLUSION

Currently, the recommended approach to applications which require predicting approximate cap positions is still to use one of the many excellent secondary structure predictors trained by strict structural criteria or concensus approaches (17). However, if applications require only approximate cap positions, then in principle one should train a machine-learning method from the outset to allow laxity in geometric criteria and gain predictive ability from sequence information. As a demonstration of this principle, we presented a method of obtaining self-consistent sequence motifs by allowing frameshifts of ±1 residue from an original set of structural assignments. The method was applied to structurally based MMDB helix-cap assignments. Allowing frameshifts significantly increased the information content of both Ncap and Ccap motifs (Equation 1). The frameshift motifs yielded small but consistent improvements (in ROC, TPF and generalization ability) over their unshifted, structurally defined counterparts. The small increase of predictive power despite the large increase in information content indicates that our frameshift motifs are sharper but otherwise similar to the MMDB structure-based motifs. High scoring sequences derived from frameshift motifs may be useful in protein design.

Motif mixture models and second-order motifs, including all site–site–residue–residue correlations, gave only minor predictive improvements over single, first-order motifs. This means that a single first-order motif, i.e. one that treats each site in a helix cap as independent, provides a good physical model for the tendency of sequences to form helix caps.

In practice, a global analysis of secondary structure, possibly including the use of frameshift motifs for other secondary-structural classes, will likely improve the performance of the helix-cap motifs obtained here.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the referees for a number of observations and suggestions which helped them improve the manuscript. This work was partially supported by NSF Grant No. DMR-0313129. Funding to pay the Open Access publication charges for this article was provided by NEC Laboratories America, Inc.

Conflict of interest statement. None declared.

REFERENCES

1. Richardson,J.S. and Richardson,D.C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science*, **240**, 1648–1652.
2. Kumar,S. and Bansal,M. (1998) Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins. *Proteins*, **31**, 460–476.
3. Aurora,R. and Rose,G.D. (1998) Helix capping. *Protein Sci.*, **7**, 21–38.
4. Goliaei,B. and Minuchehr,Z. (2003) Exceptional pairs of amino acid neighbors in alpha helices. *FEBS Lett.*, **537**, 121–127.
5. Engel,D.E. and DeGrado,W.F. (2004) Amino acid propensities are position-dependent throughout the length of alpha-helices. *J. Mol. Biol.*, **337**, 1195–1204.
6. Presta,L.G. and Rose,G.D. (1988) Helix signals in proteins. *Science*, **240**, 1632–1641.
7. Colloc'h,N., Etchebest,C., Thoreau,E., Henrissat,B. and Mornon,J.P. (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.*, **6**, 377–382.
8. Fourrier,L., Benros,C. and de Brevern,A. (2004) Use of a structural alphabet for analysis of short loops connecting repetitive structures. *Bioinformatics*, **5**, 58–71.
9. Andersen,C.A., Palmer,A.G., Brunak,S. and Rost,B. (2002) Continuum secondary structure captures protein flexibility. *Structure*, **10**, 175–184.
10. Wang,Y. and Jardetzky,O. (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci.*, **11**, 852–861.
11. Pal,L., Dasgupta,B. and Chakrabarti,P. (2005) 3(10)-helix adjoining alpha-helix and beta-strand: sequence and structural features and their conservation. *Biopolymers*, **78**, 147–162.
12. Kumar,S. and Bansal,M. (1998) Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophys. J.*, **75**, 1935–1944.
13. Dasgupta,B., Lipika,P., Basu,G. and Chakrabarti,P. (2004) Exapanded turn conformations: Characterization and sequence-structure correspondence in α-turns with implications in helix folding. *Proteins*, **55**, 305–315.
14. Thomas,A., Benhabiles,N., Meurisse,R., Ngwabije,R. and Brasseur,R. (2001) Pex, analytical tools for pdb files. II. H-Pex: noncanonical H-bonds in alpha-helices. *Proteins*, **43**, 37–44.
15. Bansal,M., Kumar,S. and Velavan,R. (2000) HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.*, **17**, 811–819.
16. Wang,Y., Anderson,J.B., Chen,J., Geer,L.Y., He,S., Hurwitz,D.I., Liebert,C.A., Madej,T., Marchler,G.H., Marchler-Bauer,A. *et al.* (2002) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **30**, 1249–1252.
17. Ginalski,K., Grishin,N.V., Godzik,A. and Rychlewski,L. (2005) Practical lessons from protein structure prediction. *Nucleic Acids Res.*, **33**, 1874–1891.
18. Berman,H., Battistuz,T., Bhat,T., Bluhm,W., Bourne,P., Burkhardt,K., Feng,Z., Gilliland,G., Iype,L., Jain,S. *et al.* (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **D58**, 899–907.
19. Bryant,S.H. (1989) PKB: a program system and data base for analysis of protein structure. *Proteins*, **5**, 233–247.
20. Hughes,J., Estep,P., Tavazoie,S. and Church,G. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
21. Akutsu,T., Arimura,H. and Shimozono,S. (2000) On approximation algorithms for local multiple alignment. *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, Japan, pp. 1–7.

22. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
23. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*., **4**, 3836–3845.
24. Moore,G.L. and Maranas,C.D. (2003) Identifying residue–residue clashes in protein hybrids by using second-order mean-field approach. *Proc. Natl Acad. Sci. USA*, **100**, 5091–5096.
25. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*., **29**, 2994–3005.
26. Ho,B.K., Thomas,A. and Brasseur,R. (2003) Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. *Protein Sci*., **12**, 2508–2522.
27. Schmidler,S.C., Liu,J.S. and Brutlag,D.L. (2000) Bayesian segmentation of protein secondary structure. *J. Comput. Biol*., **7**, 233–248.