PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# Amino acid alphabet reduction preserves fold information contained in contact interactions in proteins

**Armando D. Solis***

Biological Sciences Department, New York City College of Technology, the City University of New York (CUNY), Brooklyn, New York 11201

**ABSTRACT**

**To reduce complexity, understand generalized rules of protein folding, and facilitate *de novo* protein design, the 20-letter amino acid alphabet is commonly reduced to a smaller alphabet by clustering amino acids based on some measure of similarity. In this work, we seek the optimal alphabet that preserves as much of the structural information found in long-range (contact) interactions among amino acids in natively-folded proteins. We employ the Information Maximization Device, based on information theory, to partition the amino acids into well-defined clusters. Numbering from 2 to 19 groups, these optimal clusters of amino acids, while generated automatically, embody well-known properties of amino acids such as hydrophobicity/polarity, charge, size, and aromaticity, and are demonstrated to maintain the discriminative power of long-range interactions with minimal loss of mutual information. Our measurements suggest that reduced alphabets (of less than 10) are able to capture virtually all of the information residing in native contacts and may be sufficient for fold recognition, as demonstrated by extensive threading tests. In an expansive survey of the literature, we observe that alphabets derived from various approaches—including those derived from physicochemical intuition, local structure considerations, and sequence alignments of remote homologs—fare consistently well in preserving contact interaction information, highlighting a convergence in the various factors thought to be relevant to the folding code. Moreover, we find that alphabets commonly used in experimental protein design are nearly optimal and are largely coherent with observations that have arisen in this work.**

Proteins 2015; 83:2198–2216.


**Key words: protein structure; amino acid sequence; contact potential; knowledge-based potential; sequence representation; threading.**

## INTRODUCTION

Because of their complexity, the molecular structure and the amino acid sequence of proteins are routinely simplified in order to make analysis tractable.[1,2] Descriptor coarse-graining and a reduction of the amino acid alphabet are frequent simplification strategies used for large-scale protein structure analysis, simulation, and prediction. The amino acid sequence, the primary determinant of molecular structure, has the ability to tolerate many changes without significantly disrupting the protein's basic fold and function, as observed plainly in the multiple sequence alignments of proteins within homologous families.[3] This is due to fundamental similarities that exist among amino acids in terms of their physical and chemical properties. The question of how to judiciously reduce the 20-letter amino acid alphabet into a smaller number of amino acid kinds or groups directly responds to this observation.

Reduced amino acid alphabets have been shown to have sufficient discriminatory power to detect homologies in sequence alignment and in fold recognition,[4] suggesting that simplified sequence representations can be a viable tool in homology detection and structure prediction efforts. Reduced amino acid alphabets are also important to the study of early life and evolution. It has been postulated[5] that primordial proteins may have been formed from a reduced number of amino acids; the question becomes which amino acids are truly essential to preserving the stability of simple early folds. As an important corollary, protein design may also benefit

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Armando D. Solis, Biological Sciences Department, New York City College of Technology, the City University of New York (CUNY), Brooklyn, New York 11201. E-mail: asolis@citytech.cuny.edu

Received 6 May 2015; Revised 4 September 2015; Accepted 4 September 2015
Published online 26 September 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24936

from knowing the minimum number of amino acids necessary to stabilize folded structures as well as from understanding the coding properties of these reduced alphabets. Indeed, simplifying the amino acid alphabet dramatically reduces computational complexity, and should facilitate the process of designing de novo proteins.[6]

Here, we advance a direct, information-based way to formulate reduced amino acid alphabets. If the information required to determine protein structure lies in the amino acid sequence, we assert that the goal of all alphabet reduction schemes should be to preserve as much of this structural information as possible. Information theory is a natural tool for transforming this question into an optimization, one that we have used to answer various questions about the folding code and knowledge-based potential functions previously. Specifically, in our continuing work, we have clearly established the fundamental link between knowledge-based "energies" and information-theoretic quantities.[7,8] Subsequently, we have demonstrated that an overwhelming advantage of using the maximization of mutual information to optimize any parameter (or any knowledge-based potential function) is that it also optimizes that parameter (or potential function) for fold recognition,[7] without explicitly employing optimization using decoys,[9] a costly computational process. Specifically, optimizing by information maximization considers only native conformations, which renders any parameter optimization faster. Mutual information maximization preserves the discriminative ability of the parameter (or potential function) to detect the native conformation, thereby recovering those characteristics that are relevant to protein folding and stability. This is the strategy we apply here.

In previous work, we employed mutual information maximization to establish useful reduced alphabets based exclusively on local backbone structure.[10] We note that the alphabets developed in that work have performed well in independent comparisons.[4,11,12] Here, we extend our investigation by exploring alphabet contraction further, this time using the information residing in the long-range pairwise contact of amino acids. Practical considerations lead us also to verify, in this work, that these information-optimized alphabets perform well in extensive fold recognition tests. We also ask whether a reduced alphabet size is sufficient to provide the same discriminatory power as the full 20-letter amino acid alphabet.

Aside from searching for the best alphabet collapse that is optimized for long-range interactions, this work also investigates auxiliary issues. An obvious question that we attempt to answer here concerns the consistency of the reduced alphabets derived in this work with those derived from other considerations. There have been a number of amino acid alphabet reduction schemes advanced in the literature,[13] many of which are included in this article for comparative analysis. Some of the earliest work involved clustering amino acids with respect to their physicochemical properties. Subsequently, a number of alphabets were developed to take into account the patterns seen in multiple sequence alignments of homologous families, and also in the relationships between sequence and structure, both local and long-range. Most significantly, using reduced alphabets, some protein design efforts have successfully redesigned sequences while preserving folds,[14–18] while others have created entirely novel sequences and folds.[19–22] Through the information-theoretic tool we have developed, we also attempt to reconcile those experimental observations that use reduced alphabet approaches to rationalizing fold design and understanding protein evolution.

In this work, we cluster amino acids into 2–19 groups automatically, which we observe to embody well-known properties of amino acids such as hydrophobicity/polarity, charge, size, and aromaticity. We are able to demonstrate that these optimal alphabets maintain the discriminative power of long-range interactions with minimal loss of discriminative power. Analysis leads us to conclude that reduced alphabets (of $<10$) are able to capture virtually all of the information residing in native contacts, and may be sufficient for fold recognition, as demonstrated by extensive threading tests. We also find, in an expansive survey of the literature, that alphabets derived from varying approaches—including those derived from physicochemical intuition, local structure considerations, and sequence alignments of remote homologs—fare consistently well in preserving contact interaction information, highlighting a convergence and a consistency in the various factors thought to be relevant to the folding code. We also explore alphabets used in experimental protein design and find them to be nearly optimal and largely consistent with other observations in this work.

## THEORY AND METHODOLOGY

The overall strategy we employ in this work is to use mutual information $I(C, S)$ between sequence $S$ and conformation $C$ as the objective function for the optimization of reduced amino acid alphabets. The question of which reduced alphabet preserves the most structural information is thus addressed directly. To specify protein conformation in this work, we use residue-pair contacts,[23,24] whose interaction potential has been used widely in structure prediction and analysis. One distinct advantage of maximizing mutual information in a parameter optimization scheme is that the corresponding potential that results from the optimized parameters also perform best in fold recognition tests.[7,8] The link between mutual information and performance suggests that latent structural information is well-preserved when mutual information is maximized.

In this section, we first describe the contact potential and its parameters, and formulate its relationship with

mutual information. We then discuss a heuristic Monte Carlo algorithm to explore the clustering space of the amino acids. Thereafter, in order to test the effectiveness of the optimized amino acid alphabet reductions, we implement an extensive threading test. Finally, we describe the data sets used in this work, both for mutual information measurements and for threading.

## Contact potential parameters

The knowledge-based pairwise contact potential is derived with the following equation[24]:

$$\varepsilon(i,j) = -kT\ln\frac{p(i,j)_N}{p(i)p(j)} \quad (1)$$

Where $k$ is the Boltzmann constant, $T$ is the absolute temperature, $i$ and $j$ are amino acids, $p(i)$ is the probability of amino acid $i$ in the universe of protein sequences, and $p(i, j)_N$ is the probability of contact between amino acids $i$ and $j$ in the universe of native protein structures (signified by the subscript $N$), and $\varepsilon$ is the contact score or energy. The quasi-chemical approximation was employed as reference state because it has been shown to achieve the highest mutual information compared to other inter-residue contact reference states.[25] The same equation can be used to measure the contact score in a reduced alphabet scheme, except that in such cases $i$ and $j$ represent amino acid clusters instead of individual amino acids.

The parameters of contact used here have been optimized for mutual information in previous work.[8] In particular, contact between two residues was defined to exist when any of their heavy atoms come to within 4.5 Angstroms of one other. Contacts occurring between amino acids 5 residues away or less are not considered in the statistics. This sequence separation should eliminate any bias caused by contacting residues at adjacent turns of the alpha helix.[8,63] We recognize that work on long-range distance-dependent potential functions[64,65] have identified sequence separation of at least 9–11 as the ideal definition, to guard against chain connectivity artifacts. But we note that a fundamental difference exists in how contact and distance-dependent potentials are applied—in this case, heavy atom pairs must approach 4.5 Å to be recognized as a contact, while distance-dependent interactions can be defined to still occur at a greater distance (for example, 10 Å). Nevertheless, in a separate simulation, we checked the robustness of our results by applying a 10-residue sequence separation cutoff, and we are pleased to report that same optimal reduced amino acid alphabets are achieved regardless.

## Mutual information and the information maximization device

Mutual information $I(C, S)$ between sequence $S$ and conformation $C$ can be measured using the following equation[8]:

$$I(C,S) = \sum_{i,j} p(i,j)\ln\frac{p(i,j)_N}{p(i)p(j)} = -\frac{1}{kT}\langle\varepsilon(C|S)\rangle = \langle\Delta E_N\rangle \quad (2)$$

where the summation runs through all contact pairs ($i$, $j$), $\varepsilon$ is the contact "energy," $<\Delta E_N>$ is the mean contact energy of native conformations. This connection between mutual information and knowledge-based potentials allows us to optimize potential parameters by information maximization.

The use of the full 20-letter alphabet yields only one mutual information measurement. This is computed by gathering contact statistics from the nonredundant data set to construct probability estimates in Eq. (2), and then proceeding to apply the summation across all possible residue pairs. The equations above can also be used to compute mutual information for reduced amino acid alphabets. There are many ways to collapse the 20 amino acids into a smaller number of clusters, each configuration carrying a certain amount of mutual information. In order to reduce the alphabet size to $n$, the 20 amino acids are grouped together to form $n$ distinct clusters. The mutual information can be computed for a particular clustering by considering $i$ and $j$ as indices for the $n$ clusters (instead of indices for the 20 amino acids for the full alphabet), and implementing the summation in Eq. (2) by making a change in the indices: $i = \{1 \text{ to } n\}$ and $j = \{1 \text{ to } n\}$, where $n$ is the reduced alphabet size. This straightforward calculation, via the Information Maximization Device,[26] yields a mutual information quantity, which measures the amount of structural information preserved by that particular clustering.

## Monte carlo optimization

If each amino acid clustering yields a different value for mutual information, the goal of maximizing mutual information involves a search across various ways of clustering the 20 amino acids into $n$ clusters. We have devised an efficient Monte Carlo algorithm similar to what we used in previous work.[27] We implemented a Monte Carlo optimization via the following steps:

An initial clustering of the 20 amino acids into $n$ clusters is randomly generated. The associated mutual information is computed via Eq. (2), using statistics collated from a nonredundant database.

To generate the next trial clustering, a random change is made by moving one, two, three, or four amino acids from one cluster to another simultaneously. The number

of amino acids involved in the change is randomly chosen, with a preset sampling frequency of 0.25 per possibility.

The mutual information given by this trial clustering is computed via Eq. (2) and compared to the previous mutual information value. If the new value is higher, the trial clustering is kept, and another iteration is made from the Step 2. Otherwise, the trial clustering is discarded and the old clustering is reinstated.

When no trial is accepted in Step 3 after 1,000 trials, the criterion for acceptance is loosened slightly, by accepting any trial grouping with a mutual information that is greater than a given fraction of the old value. The fraction can be tweaked depending on the complexity of the search; for typical searches done in this work, a fraction of 0.97–0.99 is adequate. Choosing to ease the acceptance criterion is done randomly and only 10% of the time.

If after another 10,000 iterations the absolute value of the maximum information gain does not increase further, the algorithm is restarted with a new random amino acid grouping.

The Monte Carlo search is terminated when the same optimized reduced alphabet is achieved by at least 100 randomly-chosen starting amino acid groupings.

It should be remarked that the procedure described above carries out a Monte Carlo optimization, rather than generating statistical ensembles, so that the search algorithm need not follow a realistic, energetically derived sampling procedure (such as a Metropolis criterion), and can be designed and altered with great flexibility. Based on extensive tests, we find the simple procedure above converges efficiently to a locally optimal clustering. Furthermore, in order to provide the opportunity for any local minima to escape traps and better explore the landscape, we recommence the Monte Carlo procedure described above from the putative optimal clustering as a starting point, but relax the move acceptance criterion further to around 0.85–0.90 range. A total of 400 distinct runs were done for each cluster number (from 2 to 19). We report that in all instances, the trajectories find themselves back to the putative clustering, giving us greater confidence in the optimality of our results. Nonetheless, it should always be remembered that there can never be a guarantee that a heuristic algorithm finds the global maximum.

### CASP10 decoy sets in threading

The practical use of optimization by mutual information maximization is that it results in an improved performance of knowledge-based potentials in fold recognition or threading.[7] Maximizing mutual information is equivalent to decreasing the energy score of native conformations, as seen in Eq. (2). Moreover, it also has the effect of increasing the mean energy score of decoy conformations. The result is the widening of the gap between the mean energy scores of the native conformation and of incorrect (decoy) conformations, resulting in sharper discrimination in fold recognition applications.[7] In this work, we confirm this by measuring the discriminative ability of the reduced alphabets to identify the correct conformation of CASP10 targets among a set of decoys. The total energy score of a particular conformation $c$ is mean energy of all $n_c$ contacts:

$$E(c) = \frac{1}{n_c}\sum_{k=1}^{n_c} \varepsilon_k(i,j) \qquad (3)$$

where $\varepsilon$ is the contact score given by Eq. (1). We formulate a way to measure the discriminative power of a particular amino acid grouping by computing the mean percentile rank $<r>$ of the energy scores $E(c_N)$ of the set of $n_p$ native conformations $\{c_N\}$ as follows:

$$\langle r \rangle = \frac{1}{n_p}\sum_{m=1}^{n_p} r_m = \frac{1}{n_p}\sum_{m=1}^{n_p} \frac{N[E(c_N) > E(c_D)]}{n_D(m)} \qquad (4)$$

where $r_m$ is the percentile rank of the $m$th native conformation in the set $\{c_N\}$, which is computed by counting $N$, the number of decoys that have an energy score $E(c_D)$ lower than $E(c_N)$, normalized by the number of decoys $n_D(m)$ of the $m$th protein in the threading test set. In essence, $<r>$ measures the expected percentile rank of the native conformation amidst a set of decoys, given the particular amino acid alphabet reduction scheme.

### Data sets: PISCES and CASP10 decoys

The nonredundant data set used to derive mutual information and potential parameters is composed of 4,641 protein chains culled from PISCES[28] with pairwise sequence identity of no >25%, resolution of at least 2.0 Angstroms, and R-factor cutoff at 0.25. The data set contains a total of 1,066,978 residues (giving an average length of 230 residues per chain), participating in a total of 4,354,181 heavy-atom contacts (an average of about 4 contacts per residue).

Threading was done using protein models submitted to the CASP10 competition (http://www.predictioncenter.org/casp10/index.cgi).[29] This data set is composed of 122 native conformations of diverse secondary structures and folds, along with 44,948 decoys (that is, the CASP10 models), or an average of 368 decoys per native conformation.

## RESULTS AND DISCUSSION

### Monte Carlo optimization

To illustrate the performance of the Monte Carlo procedure, the progress of the amino acid alphabet reduction at alphabet size 5 in the Monte Carlo optimization is plotted in Fig. 1, which plots mutual information as a

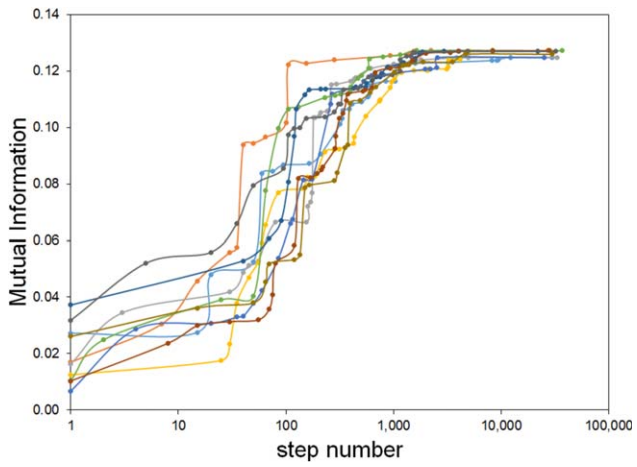


**Figure 1**

The progress of the amino acid alphabet reduction in the Monte Carlo optimization. Mutual information is plotted as a function of step number for 10 typical trajectories for an optimization at alphabet size 5. A trajectory is initiated by a randomly generated clustering of the 20 amino acids into 5 groups, and a step is made when the amino acid is randomly altered by changing the cluster membership slightly (see Methods for details). Based on 200 independent trajectories, the maximum information gain was reached 56% of the time. The searches for other alphabet sizes vary in the average length of the trajectories and the success rates, but they are qualitatively similar to the set illustrated here.

function of step number for 10 typical trajectories. (A step is made when the amino acid is randomly altered by changing the cluster membership slightly, as described in the Methodology.) An average of about 28,000 steps is needed to complete a full trajectory from an initial randomly generated clustering configuration. In our optimization procedure for alphabet size 5, we attempted a total of 200 trajectories, and out of these, the maximum information gain was reached 112 times, giving a success rate of 56%. A rapid ascent within the first 1000 steps after the initial clustering configuration is followed by a correction phase where minute increases in mutual information are achieved with small rearrangements in the amino acid grouping. A majority of local maxima were found within 6000 steps. The searches for other alphabet sizes, from 2 to 19, vary in the average length of the trajectories and the success rates, but they are qualitatively similar to Fig. 1.

Because the Monte Carlo clustering algorithm is heuristic, finding the global extremum is never guaranteed. In order to ensure confidence in the optimization result, we implemented the procedure as many times as it takes to confirm each result from at least 100 different starting amino acid cluster configurations. In the case of alphabet size of 5, we tried 200 randomly selected starting amino acid cluster configurations to meet this requirement; other alphabet sizes ranged from 120 to 350 complete trajectories. The relatively high rates of success in converging to the same maximum from different starting points gives us confidence that our Monte Carlo procedure is able to locate the optimal alphabet reductions given the alphabet size.

## Optimal amino acid partition at different alphabet sizes

The 20 amino acids were clustered into groups based on their similarities in their pairwise contact interactions, using an automatic optimization procedure that maximized mutual information. We are motivated to seek reduced alphabets that mimic the native pairwise contacts which occur in natural proteins. Mutual information was chosen as the objective function because of its success in optimizing potential functions and associated parameters in fold recognition exercises. In essence, by using mutual information, we are able to recast the question of optimal clustering of amino acids in practical terms, as the quest to find the reduced alphabets that preserve as much as possible the discriminative power of a potential function based on pairwise contact in a threading application. Specifically, we seek the best way to reduce the amino acid alphabet so as to maintain the ability of pairwise contact to identify the native conformation in a set of challenging decoy conformations.

Alphabet reduction was done from the minimum of 2 letters to a maximum of 19 letters. At one extreme, the single letter reduction is trivial, where the mutual information expected from pairwise contacts is zero (because amino acid identity is completely erased), while at the other extreme, using 20 letters also results trivially as the full expression of the complete 20 amino acid alphabet. Results of the optimization for each alphabet size are summarized in Table I, and the alphabet reduction is illustrated in Fig. 2. From the figure, it can be seen that the optimization procedure, while fully automatic, recovered recognizable patterns of clustering. The clear delineation of hydrophobic and polar/charged amino acids occurs instantly at alphabet size 2, followed by the refinement of the clustering based on other characteristics such as size, polarity, and aromaticity at higher alphabet sizes. Specifically, we find the general partitioning of side chains that are nonpolar aromatic (FWY), nonpolar aliphatic and sulfur-containing (CILMV), acidic (DE), basic (HR), small (AT), and other polar (NQS). Notably, the amino acid K segregates not with the other basic side chains but with other polar side chains. We recognize that A, despite its aliphatic side-chain, has been widely considered ambivalent, occurring frequently in the exterior as much as in the interior of folded proteins. In this clustering regime (Fig. 2), A appears to segregate with polar amino acids. We also note that two amino acids, P and Q, do not follow a clean clustering pattern, shifting from one subgroup to another as the alphabet size is increased. This is expected to happen, as this exercise

**Table I**
Mutual Information (MI) Measurements and Their Significance for Optimal Amino Acid Clustering at Different Alphabet Sizes

| Alphabet size | Contact MI (nats) | Proportion of contact MI preserved by the alphabet collapse | Mean contact MI for random clusters (nats)[a] | Standard deviation of contact MI for random clusters (nats)[a] | *Z*-score |
|---|---|---|---|---|---|
| 2 | 0.1068 | 0.751 | 0.00726 | 0.00957 | 10.40 |
| 3 | 0.1162 | 0.817 | 0.01473 | 0.01292 | 7.85 |
| 4 | 0.1223 | 0.860 | 0.02139 | 0.01503 | 6.71 |
| 5 | 0.1272 | 0.895 | 0.02918 | 0.01643 | 5.97 |
| 6 | 0.1310 | 0.921 | 0.03710 | 0.01757 | 5.34 |
| 7 | 0.1330 | 0.935 | 0.04397 | 0.01807 | 4.93 |
| 8 | 0.1347 | 0.947 | 0.05139 | 0.01845 | 4.52 |
| 9 | 0.1364 | 0.959 | 0.05917 | 0.01862 | 4.15 |
| 10 | 0.1378 | 0.969 | 0.06632 | 0.01860 | 3.84 |
| 11 | 0.1388 | 0.976 | 0.07359 | 0.01822 | 3.58 |
| 12 | 0.1395 | 0.981 | 0.08093 | 0.01803 | 3.25 |
| 13 | 0.1401 | 0.985 | 0.08885 | 0.01732 | 2.96 |
| 14 | 0.1405 | 0.988 | 0.09624 | 0.01659 | 2.67 |
| 15 | 0.1409 | 0.991 | 0.10390 | 0.01551 | 2.39 |
| 16 | 0.1413 | 0.994 | 0.11111 | 0.01438 | 2.10 |
| 17 | 0.1416 | 0.996 | 0.11903 | 0.01278 | 1.77 |
| 18 | 0.1419 | 0.998 | 0.12650 | 0.01070 | 1.44 |
| 19 | 0.1421 | 0.999 | 0.13436 | 0.00779 | 0.99 |

[a]A total of 10,000 randomly generated amino acid clustering were generated per number of clusters, from which the mean and standard deviation of their contact MI values were calculated.

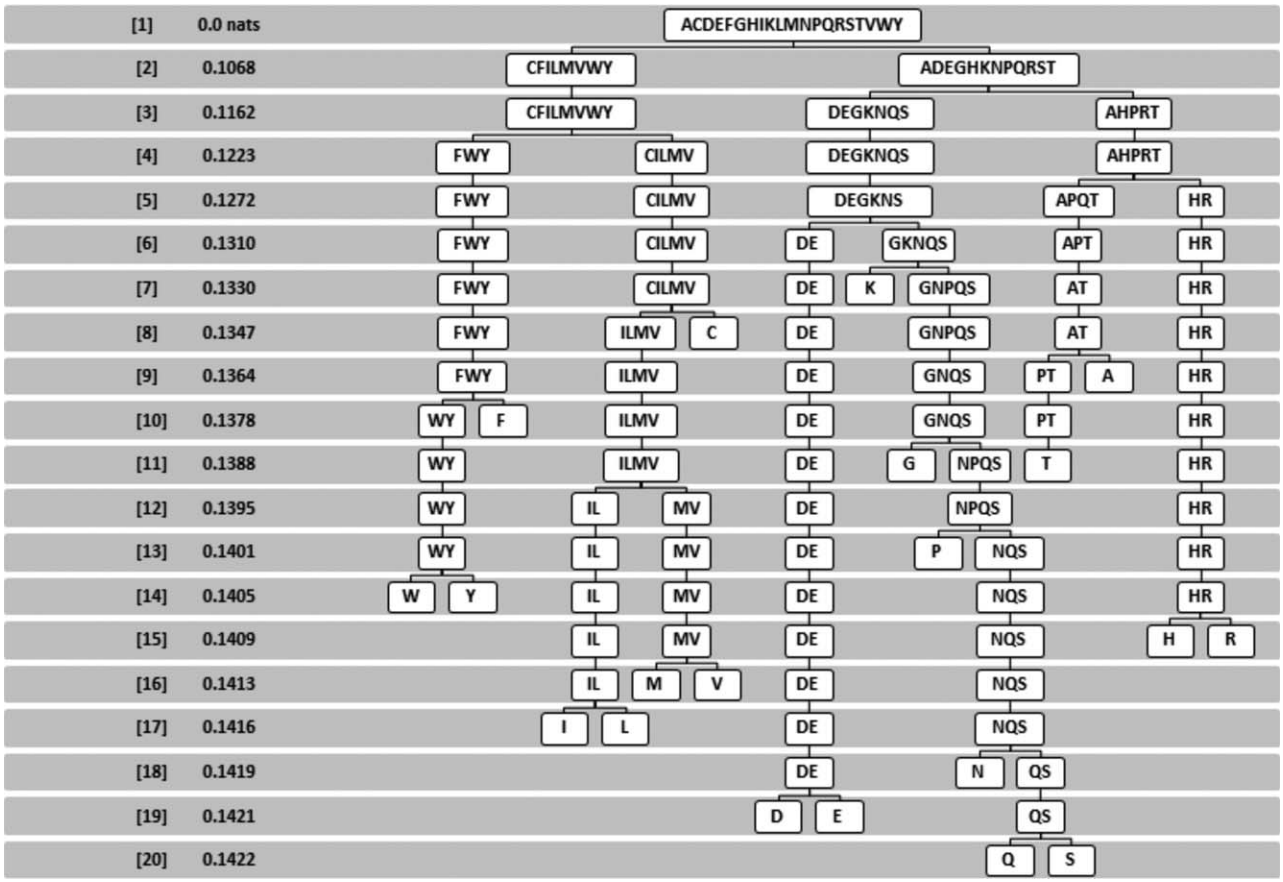


**Figure 2**
Optimal alphabet reduction using the Information Maximization Device. Numerical results of the optimization for each alphabet size are summarized in Table I. It can be seen that the optimization procedure, while fully automatic, recovered recognizable patterns of amino acid clustering.

attempts to collapse multidimensional interaction information into a simple, one-dimensional clustering scheme, and that clustering at each alphabet size is entirely independent from those at other sizes. It is remarkable that the automatic procedure actually produces clean partition patterns for the other 18 amino acids, each of which can be tracked from top to bottom (in Fig. 2) as following a conventional series of partitions. This suggests that a fundamental coding pattern exists, and that this pattern largely reducible to a simple alphabet reduction scheme.

The mutual information for each optimal clustering is shown in Table I for all alphabet sizes. Given that the contact mutual information when using the full 20-letter alphabet is 0.1422 nats, we can examine the proportion of this contact information that is preserved by any given reduced alphabet (shown in column 3 of Table I). We observe that the binary hydrophobic-polar (H/P) partition, at alphabet size 2, recovers a large proportion of contact information, at around 75%. This measurement confirms the long-standing view that a substantial amount of the tertiary structure information encoded in the amino acid sequence resides in the hydrophobicity of amino acids. This is consistent with successful attempts[16,18,20,22,30,31] to design folding sequences based principally on the H/P patterns observed in similar topologies that occur in nature. We also observe that much of the contact mutual information is preserved at significantly low alphabet sizes. Specifically, at alphabet size 5, nearly 90% of the mutual information is preserved; at 10, it's 97%. Consequently, information gains diminish when expanding to higher alphabet sizes.

We ask how significant these mutual information measurements are for the optimal partitions. To quantify significance, we generated 10,000 random amino acids groupings for every alphabet size and measured each grouping's mutual information. From this simulation, the means and the standard deviations were computed, with which the *Z*-score for the optimal partitions were measured. The *Z*-score values, shown in the last column of Table I, demonstrate the clear significance of the mutual information contained in the optimal reduced alphabets, particularly at low alphabet sizes. The optimal clustering that appears to be most significant (as measured by the *Z*-score) occurs at alphabet size 2, with the H/P partition. At moderate levels of alphabet sizes, the Z-scores, albeit lower than for the H/P partition, nonetheless signify a significant signal. Only at alphabet sizes greater than 14 does the Z-score wane to more modest levels, hinting at marginal significance of the optimal alphabets around the interval.

## Fold recognition by CASP10 threading

To confirm whether the increase in mutual information in going to higher alphabet sizes results in a concomitant improvement in discrimination, we subjected various reduced alphabets to extensive threading tests using a data set composed of 122 targets and thousands of decoys extracted from CASP10.[29] The reduced alphabets were used to derive the contact potential, as defined by Eq. (1), and the contact energy for all conformations in the threading test set were computed using Eq. (3). Comparison of the contact energy of the native conformation to the spectrum of contact energies given by its associated decoys produces a percentile rank *r*, and averaging this rank across 122 targets yields the mean percentile rank $<r>$, given by Eq. (4).

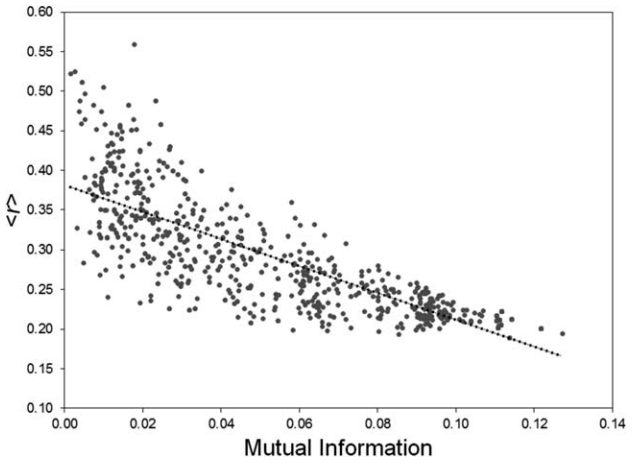


**Figure 3**

Relationship between mutual information and threading discrimination. A total of 500 amino acid groupings for reduced alphabet size 5 were generated randomly, and their mutual information $I(C, S)$ values were measured. Each randomly generated alphabet was then used to formulate a contact potential, which was subjected to CASP10 threading. The resulting $<r>$, the mean percentile rank of the native conformation amidst a spectrum of decoys, is plotted against $I(C, S)$ here. A significant correlation between $I(C, S)$ and $<r>$ can be observed in the plot, confirming the utility of the Information Maximization Device in preserving the discriminatory power of knowledge-based potentials that arise from parameter optimization.

First, we ask how mutual information affects discriminative power in alphabet reduction. We generate random amino acid groupings for a given alphabet size. For each randomly generated grouping, a mutual information can be computed, after which the contact potential can be generated anew and used in the threading test of all 122 CASP10 targets, resulting in a particular measurement for $<r>$. Generating 500 random groupings for alphabet size 5 across the spectrum of mutual information yields Fig. 3. The benefit of using $I(C, S)$ as an objective function is demonstrated here by its inverse relationship with discriminative power (at $R^2$ coefficient of 0.64), confirming that the effect of optimizing any parameter used in knowledge-based potentials via $I(C, S)$ is a general increase in discrimination in fold recognition.

Second, we ask whether the increased mutual information from lower to higher alphabet sizes results in significant improvements in fold discrimination. Results of the

**Table II**
CASP10 Threading Measurements and Their Significance for Optimal Amino Acid Clustering at Different Alphabet Sizes

| Alphabet size | $<r>$ of optimal reduced amino acid alphabet | Number of randomly generated reduced alphabet[a] | Mean $<r>$ | Standard deviation of $<r>$ | *Z*-score |
|---|---|---|---|---|---|
| 2 | 0.21292 | 500 | 0.41956 | 0.08938 | 2.31 |
| 3 | 0.20152 | 500 | 0.37801 | 0.08628 | 2.05 |
| 4 | 0.19749 | 500 | 0.34889 | 0.08035 | 1.88 |
| 5 | 0.19439 | 500 | 0.32612 | 0.07333 | 1.80 |
| 6 | 0.18870 | 500 | 0.30338 | 0.06705 | 1.71 |
| 7 | 0.18683 | 500 | 0.28964 | 0.05789 | 1.78 |
| 8 | 0.18270 | 500 | 0.27101 | 0.05290 | 1.67 |
| 9 | 0.17985 | 500 | 0.25644 | 0.04735 | 1.62 |
| 10 | 0.17986 | 500 | 0.24853 | 0.04353 | 1.58 |
| 11 | 0.18461 | 500 | 0.23677 | 0.03688 | 1.41 |
| 12 | 0.18473 | 500 | 0.22763 | 0.03186 | 1.35 |
| 13 | 0.18300 | 500 | 0.22140 | 0.02844 | 1.35 |
| 14 | 0.18143 | 500 | 0.22140 | 0.02844 | 1.35 |
| 15 | 0.18103 | 500 | 0.21333 | 0.02615 | 1.22 |
| 16 | 0.17966 | 500 | 0.20795 | 0.02324 | 1.16 |
| 17 | 0.18143 | 500 | 0.20274 | 0.01976 | 1.17 |
| 18 | 0.18007 | 500 | 0.19617 | 0.01653 | 0.89 |
| 19 | 0.18005 | 500 | 0.18484 | 0.00839 | 0.57 |

[a]A large number of randomly generated amino acid clustering were generated per alphabet size, which were used to do extensive threading using 122 targets in CASP10. Each threading test, using a given random alphabet reduction, generated a percentile rank of the native conformation amidst the energy scores of CASP10 decoys. From the series of threading tests, the mean and standard deviation of their contact mutual information values were calculated. The *Z*-score represents the significance of the threading discrimination provided by the optimal amino acid clustering for each number of clusters, by asking if the same kind of discrimination can be achieved randomly.

threading tests for the optimized reduced alphabets at each alphabet size (specified in Fig. 2) are summarized in Table II. The plot between $<r>$ and mutual information $I(C, S)$, included as Fig. 4, showing an inverse correlation with an $R^2$ coefficient of 0.95, behaves as expected at low alphabet sizes. We note that at alphabet sizes above 9, nearly 96% of the mutual information due to contact interactions is preserved. It appears that the minute increases in mutual information $I(C, S)$ in going to higher alphabet sizes do not bring about noticeable improvement in discrimination.

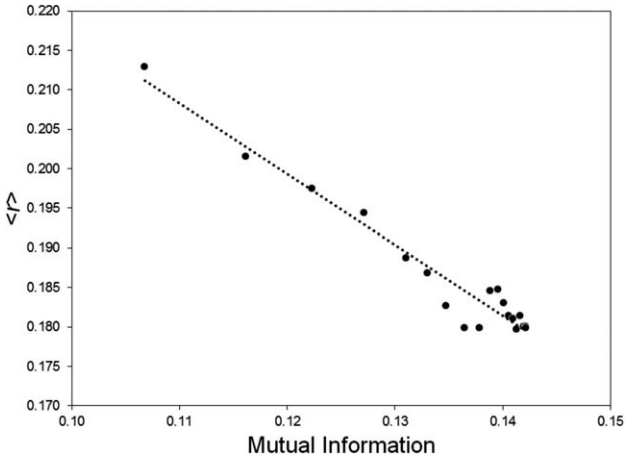


**Figure 4**
CASP10 decoy threading using optimal reduced alphabet from Size 2–20. Numerical results of the threading tests for the optimized reduced alphabets at each alphabet size are summarized in Table II. The plot between $<r>$ and mutual information $I(C, S)$, shown here, confirms the expected correlation at low alphabet sizes. At alphabet sizes above 9, where nearly 96% of the mutual information due to contact interactions is preserved, further increases in mutual information $I(C, S)$ in going to higher alphabet sizes do not bring about any significant improvement in $<r>$ discrimination.

To further investigate the behavior of discrimination in this region, the *Z*-score was computed, for each alphabet size, using the distribution of $<r>$ values of potentials arising from large sets of randomly generated amino acid clustering. These *Z*-score values, given in Table II, show that potentials arising from low alphabet sizes have significant discriminatory power, while at high alphabet sizes, the advantage of using information-optimized alphabets over other alphabets becomes less significant (if 5% *P* values, with the corresponding one-tailed *Z*-score at 1.645, is used). It appears that above alphabet size 9, due to an increased number of clusters, there is a higher probability that differently coding amino acids will occur in separate groups anyway when the 20 amino acids are partitioned randomly. This suggests that there may be a minimum number of subgroups that can encapsulate the coding behavior of amino acids, at least as far as long-range interactions are concerned. Coupled with the observation that alphabet sizes above 9 show similar levels of discriminatory power (Fig. 4), the observation that the optimized alphabet crosses the threshold of significance at around 8 (Table II) suggests that informatically, alphabet sizes around 8–9 appears to be quite sufficient to capture the discriminatory power of tertiary contact interactions contained in the 20-letter alphabet.

**Table III**
Amino Acid Alphabets Based on Amino Acid Physicochemical Properties. [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

| Reference | Alpha. Size | Amino acid alphabet | Contact MI (nats) | *Z*-score for contact MI | Proportion of MI preserved[a] | <*r*> in CASP10 threading[b] | Approach/notes |
|---|---|---|---|---|---|---|---|
| Mahler, 1968[32] | 8 | AGILV-CM-DE-FWY-HKR-NQ-P-ST | 0.1299 | 4.26 | 0.964 | 0.1840 | Clustering based on amino acid chemistry, from Stephenson (2013)[13] |
| Lehninger, 1970[33] | 4 | AFILMPVW-CGNQSTY-DE-HKR | 0.1012 | 5.31 | 0.828 | 0.1991 | Clustering based on amino acid chemistry, from Stephenson (2013) |
| Dickerson 1983[34] | 3 | ACGPSTWY-DEHKNQR-FILMV | 0.0911 | 5.91 | 0.784 | 0.1998 | Clustering based on amino acid chemistry, from Stephenson (2013) |
| Taylor 1986[35] | 10 | ACGS-DE-FWY-HKR-ILV-M-N-P-Q-T | 0.1253 | 3.17 | 0.909 | 0.1982 | Clustering based on amino acid chemistry |
| Thomas 1996[36] | 2 | ACFILMVWY-DEGHKNPQRST | 0.0859 | 8.21 | 0.804 | 0.2144 | Method of clustering not explicitly indicated, but clustering likely based on amino acid chemistry |
| | 3 | AFILMVWY-C-DEGHKNPQRST | 0.0873 | 5.62 | 0.751 | 0.2017 | |
| | 4 | AFILMVWY-C-DEKR-GHNPQST | 0.0994 | 5.19 | 0.813 | 0.1965 | |
| | 5 | AFILMVWY-C-DE-GHNPQST-KR | 0.0973 | 4.15 | 0.765 | 0.1940 | |
| | 6 | A-C-DE-FILMVWY-GHNPQST-KR | 0.1223 | 4.85 | 0.934 | 0.1862 | |
| | 7 | A-C-DE-FILMVWY-G-HNPQST-KR | 0.1245 | 4.46 | 0.936 | 0.1908 | |
| | 8 | A-C-DE-FILMV-G-HNPQST-KR-WY | 0.1300 | 4.26 | 0.965 | 0.1916 | |
| | 9 | A-C-DE-FILMV-G-HNQST-KR-P-WY | 0.1303 | 3.82 | 0.955 | 0.1919 | |
| | 10 | A-C-DE-FILMV-G-HNQ-KR-P-ST-WY | 0.1307 | 3.46 | 0.948 | 0.1914 | |

[a]Reduced alphabets that preserve at least 75% of the contact MI are indicated in orange and green; those that preserve at least 90% are indicated in green.
[b]Reduced alphabets that perform as well or better than the optimized alphabets (Fig. 2) are indicated in red.

## Amino acid alphabets in the literature derived through various approaches

We assembled 114 reduced alphabets (of alphabet sizes 10 and below) from the literature, and classified them into four general categories: (A) alphabets constructed based on physicochemical properties of amino acid side chains; (B) alphabets derived from multiple sequence alignments of structural homologs; (C) alphabets based on similarities in local structure or backbone coding; and (D) other alphabets derived from long range interaction considerations. For each of the 114 alphabets, the contact mutual information was computed, along with its associated *Z*-score (based on statistics in Table I), as well as the proportion of mutual information preserved (based on the maximal mutual information for the given alphabet size). In order to measure their relative effectiveness, the reduced alphabets were also used to derive contact potentials, which were then subjected to threading using the same CASP10 decoy sets. The resulting mean percentile rank <*r*> and the associated *Z*-score were also computed (based on statistics in Table II). The reduced alphabets and their associated measurements are summarized in Tables III–VI.

To gain better insight into the general effectiveness of these reduced alphabets in preserving contact information and in generating potential functions for fold recognition, some relevant statistics were computed and summarized in Table VII. First, we ask how many of these alphabets actually capture significant amounts of contact interaction information. We observe that more than a third of all alphabets (39%) retain at least 90% of the mutual information residing in contact structure (those indicated in green), and more than 4 out of 5 retain at least 75% (indicated in green and yellow). We then looked at the mean proportion of mutual information in all the alphabets, and found that the alphabets examined here contain, on average, 81% of the information residing in contact structure. It is indeed remarkable that a significant proportion of long range structure information is preserved in these alphabets regardless of the clustering approach. Moreover, there are a number of distinct alphabets that encode significant amounts of contact structure information, some approaching maximum levels, suggesting that there is some degeneracy in encoding long-range structure in the full 20-letter alphabet.

We then ask how well each of the four general approaches perform in preserving contact information. We find that alphabets derived from homologous sequence alignments appear to embody long-range interactions fairly well, at more than 89% on average. This indicates that substitution patterns in homologous families, summarized in those alphabets, factor in similarity in long-range interactions principally. We also find that alphabets derived from physicochemical considerations, often using intuition to cluster the amino acids, perform rather well, at 86% on average. This observation confirms that early protein scientists were correct to consider notions of hydrophobicity, aromaticity, size, and polarity as the key properties that influence the folding propensities of amino acids. Alphabets from these two approaches experience the least degradation in <*r*> from

**Table IV**
Amino Acid Alphabets Based on Sequence/Structure Alignments and Other Sequence-Based Considerations. [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

| Reference | Alpha. size | Amino acid alphabet | Contact MI (nats) | *Z*-score for contact MI | Proport-ion of MI preserved[a] | <*r*> in CASP10 threading[b] | Approach/notes |
|---|---|---|---|---|---|---|---|
| Dayhoff, 1978[3] | 6 | AGPST-C-DENQ-FWY-HKR-ILMV | 0.1241 | 4.95 | **0.948** | 0.1974 | First comprehensive substitution study; amino acid clusters based on chemistry are confirmed by PAM matrix |
| Landes, 1994[37] | 10 | AST-C-DEN-FY-G-H-ILMV-KQR-P-W | 0.1284 | 3.34 | **0.932** | 0.1895 | Based on Risler (1988),62 a quantitative comparison of substitution matrices, and confirmed in FASTP and SCAN sequence alignment |
| Murphy, 2000[38] | 2 | ACFGILMPSTVWY-DEHKNQR | 0.0360 | 3.00 | 0.337 | 0.2531 | Reduction based on correlations observed from the BLOSUM50 matrix |
| | 4 | AGPST-CILMV-DEHKNQR-FWY | 0.1160 | 6.29 | **0.948** | 0.2143 | |
| | 8 | AG-CILMV-DENQ-FWY-H-KR-P-ST | 0.1256 | 4.02 | **0.932** | 0.1998 | |
| | 10 | A-C-DENQ-FWY-G-H-ILMV-KR-P-ST | 0.1299 | 3.42 | **0.942** | 0.1910 | |
| Prlic, 2000[39] (HSDM) | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | **1.000** | **0.2129** | Derived from HSDM based on superimpositions of dissimilar sequences of similar structure |
| | 3 | ADEGHKNPQRST-CFILMVY-W | 0.1087 | 7.27 | **0.936** | 0.2127 | |
| | 4 | ADEGHKNQRST-CFILMVY-P-W | 0.1090 | 5.83 | **0.891** | 0.2109 | |
| | 5 | ADEGKNQRST-CFILMVY-H-P-W | 0.1114 | 5.01 | **0.876** | 0.2042 | |
| | 6 | ADEGKNQRST-C-FILMVY-H-P-W | 0.1131 | 4.32 | **0.863** | 0.2013 | |
| | 7 | ADEKNQRST-C-FILMVY-G-H-P-W | 0.1144 | 3.90 | **0.860** | 0.2057 | |
| | 8 | AEKQRST-C-DN-FILMVY-G-H-P-W | 0.1167 | 3.54 | **0.866** | 0.1985 | |
| Prlic, 2000[39] (SDM) | 4 | ADEGKNPQRST-C-FILMVWY-H | 0.1108 | 5.95 | **0.906** | 0.2025 | Derived from SDM based on superimpositions of dissimilar sequences of similar structure |
| | 5 | ADEGKNPQRST-C-FILMVY-H-W | 0.1127 | 5.08 | **0.886** | 0.2030 | |
| | 6 | ADEGKNQRST-C-FILMVY-H-P-W | 0.1131 | 4.32 | **0.863** | 0.2013 | |
| | 7 | ADEKNQRST-C-FILMVY-G-H-P-W | 0.1144 | 3.89 | **0.860** | 0.2057 | |
| | 8 | ADEKNQRST-C-FILMV-G-H-P-W-Y | 0.1176 | 3.59 | **0.873** | 0.2071 | |
| Rogov, 2001[40] | 9 | ADNST-C-EKQR-FILVY-G-H-M-P-W | 0.1154 | 3.02 | **0.846** | 0.2072 | Based on an analysis of local sequence patterns, specifically the relative distribution of surrounding residues in the primary sequence |
| Cannata, 2002[41] | 2 | ACDEGHKNPQRST-FILMVWY | 0.1035 | 10.05 | **0.969** | 0.2184 | Partitions based on BLOSUM40 matrix |
| | 3 | ADEGHKNPQRST-C-FILMVWY | 0.1085 | 7.26 | **0.934** | 0.2098 | |
| | 4 | ADEGHKNPQRST-C-FILMVY-W | 0.1104 | 5.92 | **0.903** | 0.2100 | |
| | 5 | AGPST-C-DEHKNQR-FILMVY-W | 0.1134 | 5.13 | **0.891** | 0.2085 | |
| | 6 | AGPST-C-DEKNQR-FILMVY-H-W | 0.1160 | 4.49 | **0.885** | 0.1982 | |
| | 7 | ADGNST-C-EKQR-FILMVY-H-P-W | 0.1141 | 3.88 | **0.858** | 0.2043 | |
| | 8 | ADGNST-C-EKQR-FY-H-ILMV-P-W | 0.1186 | 3.64 | **0.880** | 0.2076 | |
| | 9 | AGST-C-DN-EKQR-FY-H-ILMV-P-W | 0.1220 | 3.37 | **0.894** | 0.1968 | |
| | 10 | AGST-C-DN-EQ-FY-H-ILMV-KR-P-W | 0.1265 | 3.24 | **0.918** | 0.1901 | |
| Kosiol, 2004[42] | 5 | ADEGHKNPQRST-C-FY-ILMV-W | 0.1147 | 5.20 | **0.902** | 0.2126 | Partition based on PAM primarily |
| Fan, 2003[43] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | **1.000** | **0.2129** | Based on sequence alignment of remote homologs, with 10 as the ideal alphabet |
| | 5 | APST-CFWY-DEHKNQR-G-ILMV | 0.1173 | 5.36 | **0.922** | 0.2115 | |
| | 8 | AST-CFWY-DEQ-G-HN-ILMV-KR-P | 0.1254 | 4.01 | **0.931** | 0.1955 | |
| | 10 | AST-C-DEQ-FWY-G-HN-IV-KR-LM-P | 0.1280 | 3.31 | **0.929** | 0.1958 | |

**Table IV**
(Continued)

| Reference | Alpha. size | Amino acid alphabet | Contact MI (nats) | *Z*-score for contact MI | Proport-ion of MI preserved[a] | <*r*> in CASP10 threading[b] | Approach/notes |
|---|---|---|---|---|---|---|---|
| Li, 2003[44] | 10 | AC-DE-FWY-G-HN-IV-KQR-LM-P-ST | 0.1270 | 3.26 | 0.921 | 0.1958 | Based on BLOSUM, confirmed by BLAST alignment of SCOP-classified proteins |
| Edgar, 2004[45] (SE-B) | 6 | AST-CP-DEHKNQR-FWY-G-ILMV | 0.1166 | 4.53 | 0.890 | 0.2134 | Based on the information compression of substitution matrices |
| | 8 | AST-C-DHN-EKQR-FWY-G-ILMV-P | 0.1204 | 3.74 | 0.894 | 0.2119 | |
| | 10 | AST-C-DN-EQ-FY-G-HP-ILMV-KR-W | 0.1274 | 3.28 | 0.924 | 0.1923 | |
| Edgar, 2004[45] (SE-V) | 10 | AST-C-DEN-FY-G-H-ILMV-KQR-P-W | 0.1284 | 3.34 | 0.932 | 0.1895 | Based on the information compression of substitution matrices |
| Wrabl, 2005[46] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | 1.000 | 0.2129 | Based on multiple sequence alignments found in the BLOCKS database |
| | 3 | ACGPST-DEHKNQR-FILMVWY | 0.1068 | 7.13 | 0.919 | 0.2149 | |
| | 4 | ACGST-DEHKNQR-FPWY-ILMV | 0.0944 | 4.86 | 0.772 | 0.2127 | |
| | 5 | ACST-DEKNQR-FHWY-GP-ILMV | 0.1117 | 5.02 | 0.878 | 0.2020 | |
| | 6 | ACST-DEKNQR-FHWY-G-ILMV-P | 0.1127 | 4.30 | 0.856 | 0.2023 | |
| | 7 | ACST-DEN-FWY-G-HKQR-ILMV-P | 0.1228 | 4.36 | 0.923 | 0.2003 | |
| | 8 | AST-C-DEN-FWY-G-HKQR-ILMV-P | 0.1268 | 4.09 | 0.941 | 0.1975 | |
| | 9 | AST-C-DEN-FWY-G-H-ILMV-KQR-P | 0.1281 | 3.70 | 0.939 | 0.1913 | |
| | 10 | AST-C-DEN-FY-G-H-ILMV-KQR-P-W | 0.1284 | 3.34 | 0.932 | 0.1895 | |
| Melo, 2006[4] | 5 | AG-C-DEKNPQRST-FILMVWY-H | 0.1137 | 5.14 | 0.894 | 0.2051 | Based on substitutions observed in structural alignments of remote homologs from SCOP |

[a]Reduced alphabets that preserve at least 75% of the contact MI are indicated in orange and green; those that preserve at least 90% are indicated in green.
[b]Reduced alphabets that perform as well or better than the optimized alphabets (Fig. 2) are indicated in red.

**Table V**
Amino Acid Alphabets Based on Local Structure (Protein Blocks) Considerations. [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

| Reference | Alpha. size | Amino acid alphabet | Contact MI (nats) | *Z*-score for contact MI | Proportion of MI preserved[a] | $<r>$ in CASP10 threading[b] | Approach/notes |
|---|---|---|---|---|---|---|---|
| Solis, 2000[10] (DSSP) | 2 | ACEFHIKLMQRVWY-DGNPST | 0.0378 | 3.19 | 0.354 | 0.2565 | Based on the effect of local sequence on DSSP-defined secondary structure |
| | 3 | AEHKQR-CFILMVWY-DGNPST | 0.1077 | 7.20 | **0.927** | 0.2103 | |
| | 4 | AEHKQR-CFILMVWY-DNST-GP | 0.1083 | 5.78 | **0.885** | 0.2127 | |
| | 5 | AEHKQR-CST-DN-FILMVWY-GP | 0.1081 | 4.80 | **0.849** | 0.2136 | |
| | 6 | AEKQR-CHST-DN-FIV-GP-LMWY | 0.1096 | 4.13 | **0.837** | 0.2111 | |
| | 7 | ACH-DN-EKQR-FIV-GP-LMWY-ST | 0.1127 | 3.81 | **0.848** | 0.2002 | |
| | 8 | AM-CF-DNS-EKQR-GP-HT-IV-LWY | 0.1136 | 3.37 | **0.844** | 0.2053 | |
| | 9 | AMW-CY-DNS-EKQR-F-GP-HT-IV-L | 0.1104 | 2.75 | **0.809** | 0.1968 | |
| | 10 | AM-C-DNS-EKQR-F-GP-HT-IV-LY-W | 0.1178 | 2.77 | **0.855** | 0.2020 | |
| Solis, 2000[10] (GBMR) | 2 | ACDEFHIKLMNQRSTVWY-GP | 0.0125 | 0.54 | 0.117 | 0.4975 | Based on the effect of local sequence on virtual alpha carbon backbone |
| | 3 | ACDEFHIKLMNQRSTVWY-G-P | 0.0134 | 0.11 | 0.115 | 0.5026 | |
| | 4 | ADEKNQRST-CFHILMVWY-G-P | 0.1063 | 5.65 | **0.869** | 0.2063 | |
| | 5 | AEHKQRST-CFILMVWY-DN-G-P | 0.1111 | 4.98 | **0.873** | 0.2094 | |
| | 6 | AEFHIKLMQRVWY-CT-DN-G-P-S | 0.0403 | 0.68 | 0.308 | 0.2675 | |
| | 7 | AEFIKLMQRVWY-CH-DN-G-P-S-T | 0.0425 | −0.08 | 0.320 | 0.2574 | |
| | 8 | AEFIKLMQRVWY-CH-D-G-N-P-S-T | 0.0434 | −0.43 | 0.322 | 0.2513 | |
| | 9 | AEFIKLMQRVWY-C-D-G-H-N-P-S-T | 0.0448 | −0.77 | 0.328 | 0.2462 | |
| | 10 | AEFIKLMQRVW-C-D-G-H-N-P-S-T-Y | 0.0516 | −0.20 | 0.374 | 0.2635 | |
| Andersen, 2004[47] | 2 | AFILMVWY-CDEGHKNPQRST | 0.0822 | 7.83 | **0.769** | 0.2197 | Based on the ability of the reduced alphabet to preserve secondary structure prediction |
| | 3 | AFILMVWY-CDEHKNQRST-GP | 0.0835 | 5.32 | 0.719 | 0.2043 | |
| | 4 | ALM-CDEHKNQRST-FIVWY-GP | 0.0947 | 4.87 | **0.774** | 0.2016 | |
| | 5 | ALM-CDHNST-EKQR-FIVWY-GP | 0.0951 | 4.01 | 0.747 | 0.2029 | |
| | 6 | ALM-CHT-DNS-EKQR-FIVWY-GP | 0.1017 | 3.68 | **0.751** | 0.1922 | |
| | 7 | A-CHT-DNS-EKQR-FIVWY-GP-LM | 0.1172 | 4.05 | **0.881** | 0.2055 | |
| | 8 | A-CHT-DNS-EKQR-FIVWY-G-LM-P | 0.1182 | 3.62 | **0.877** | 0.2049 | |
| | 9 | A-CHT-DNS-EKQR-FWY-G-IV-LM-P | 0.1233 | 3.44 | **0.904** | 0.2055 | |
| | 10 | A-CH-DNS-EKQR-FWY-G-IV-LM-P-T | 0.1249 | 3.15 | **0.906** | 0.2011 | |
| Etchebest 2007[48] | 5 | AEKLMQR-CDHNST-FIVWY-G-P | 0.0732 | 2.68 | 0.575 | 0.2528 | Based on structural equivalencies in local protein blocks (structural alphabet) |
| | 9 | ALM-CT-DN-EKQR-FWY-G-HS-IV-P | 0.1070 | 2.57 | **0.784** | 0.1956 | |

[a]Reduced alphabets that preserve at least 75% of the contact MI are indicated in orange and green; those that preserve at least 90% are indicated in green.
[b]Reduced alphabets that perform as well or better than the optimized alphabets (Fig. 2) are indicated in red.

optimal—that is, those derived using side-chain chemical properties increase $<r>$ by 0.41% while those derived from sequence alignments increase $<r>$ by 1.39% (Column 7 of Table VII). These are modest increases in $<r>$ from optimal, consistent with the high mutual information retention rates by both approaches.

One surprise is the relative underperformance of alphabets formulated using long-range considerations. We should note that a subgroup, namely those derived from various reductions of the Miyazawa-Jernigan (MJ) contact potentials,[60] perform much better than those derived through other means. In fact, 5 reduced alphabets perform as well as the optimized alphabets derived in this work, as measured by CASP10 threading tests. Their superior $<r>$ values are indicated in red in Table VI. It appears that alphabets generated from the MJ potential, a well-established and thoroughly vetted energy function, are at least as effective in preserving contact information as our optimal alphabets.

We observe that the alphabets derived from local structure/backbone considerations perform least effectively in capturing contact information, both in terms of the level of mutual information preserved and the performance of contact potentials derived using those alphabets. It appears that some locally-relevant structural information is not fully embodied in contact structure information, and vice versa. For instance, the distinctive nature of G and P, which allow for the greatest and least flexibility of the backbone respectively, are highlighted in many of the locally determined alphabets, which feature their early separation from the rest of the amino acids, factors that do not appear to be critical in reduced alphabets obtained via other approaches. This hints at some frustration in the way the local and long-range folding codes are embodied in the amino acid sequence. Despite these disparate emphases, however, roughly 2/3 of long-range information is preserved in these alphabets on average. Part of the reason is that secondary structure (in Solis DSSP) implicitly embodies long range information in the necessary interactions between secondary structure elements, particularly beta sheets. On the other hand, the truly locally based alphabet (Solis GBMR) shows the frustration acutely. That particular approach[10] considered only the effect of the local sequence on the

**Table VI**
Amino Acid Alphabets Based on Long Range Interaction Considerations. [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

| Reference | Alpha. size | Amino acid alphabet | Contact MI (nats) | *Z*-score for contact MI | Proportion of MI preserved[a] | $\langle r \rangle$ in CASP10 threading[b] | Approach/notes |
|---|---|---|---|---|---|---|---|
| Crippen, 1991[49] | 4 | AV-CFILM-DKPQTW-EGHNRSY | 0.0699 | 3.23 | 0.571 | 0.2090 | Based on pairwise residue contacts |
| Maiorov, 1992[50] | 7 | ACFILM-DEQR-G-HSV-KNT-P-WY | 0.0844 | 2.24 | 0.634 | 0.1931 | Based on pairwise residue contacts at separation of 3-4 residues, using Crippen (1991) |
| Maiorov 1992[50] | 7 | AV-CFILM-DENQ-G-HSTWY-KR-P | 0.0954 | 2.85 | 0.717 | 0.1941 | Based on pairwise residue contacts at separation of 5 or more residues, using Crippen (1991) |
| Wang, 1999[51] | 5 | AHT-CFILMVWY-DE-GP-KNQRS | 0.1184 | 5.43 | 0.930 | 0.1912 | Based on a reduction of the Miyazawa-Jernigan (MJ) potential |
| Cieplak, 2001[52] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.4 | 1.000 | 0.2129 | Based on a reduction of the MJ potential |
| | 5 | AH-CMVWY-DEGNPQRST-FIL-K | 0.1129 | 5.10 | 0.888 | 0.1943 | |
| Liu, 2002[53] | 5 | ACW-DE-FILMV-GHNPQSTY-KR | 0.0932 | 3.30 | 0.733 | 0.1973 | Based on a reduction of the MJ potential |
| Esteve, 2004[54] | 5 | AGHNPQST-CFILMVWY-DE-K-R | 0.1215 | 5.62 | 0.956 | 0.1909 | Based on a reduction of the MJ potential |
| | 7 | AGNPQST-C-DE-FILMVWY-H-K-R | 0.1257 | 4.52 | 0.945 | 0.1844 | |
| Bacardit, 2009[12] | 2 | ACFGHILMVWY-DEKNPQRST | 0.0584 | 5.34 | 0.547 | 0.2370 | Based on pattern similarities in contact number |
| | 3 | ACFILMVW-DEKNPQR-GHST | 0.0868 | 5.58 | 0.747 | 0.2018 | |
| | 4 | AFHTY-CILMV-DEKPQ-GNRSW | 0.0795 | 3.87 | 0.650 | 0.2058 | |
| | 5 | AIS-CHLV-DEPQY-FGMW-KNRT | 0.0413 | 0.74 | 0.325 | 0.2112 | |
| Bacardit, 2009[12] | 2 | ACFILMVWY-DEGHKNPQRST | 0.0859 | 8.21 | 0.804 | 0.2144 | Based on pattern similarities in solvent accessibility |
| | 3 | AGHNPST-CFILMVWY-DEKQR | 0.1095 | 7.33 | 0.942 | 0.2055 | |
| | 4 | ACHIMT-DEKPQ-FLVWY-GNRS | 0.0959 | 4.96 | 0.784 | 0.1938 | |
| | 5 | ACIY-DESW-FLMV-GHNT-KPQR | 0.0805 | 3.13 | 0.633 | 0.1962 | |
| Pape, 2010[55] | 2 | AEIKLMQRV-CDFGHNPSTWY | 0.0035 | −0.39 | 0.333 | 0.3850 | Based on distance-dependent pairwise interactions, maximum distance at 8 Angstroms |
| | 4 | AEKQR-CFGPWY-DHNST-ILMV | 0.0715 | 3.34 | 0.585 | 0.2106 | |
| | 7 | AEQ-CGP-DN-FWY-HST-ILMV-KR | 0.1179 | 4.09 | 0.886 | 0.2059 | |
| Pape, 2010[55] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | 1.000 | 0.2129 | Based on distance-dependent pairwise interactions, maximum distance at 50 Angstroms |
| | 3 | AGHNPST-CFILMVWY-DEKQR | 0.1095 | 7.33 | 0.942 | 0.2055 | |
| | 5 | AG-CIV-DEKQR-FLMWY-HNPST | 0.1138 | 5.15 | 0.894 | 0.2127 | |
| | 8 | AG-CIV-DE-FWY-HST-KQR-LM-NP | 0.1262 | 4.05 | 0.937 | 0.2001 | |

[a]Reduced alphabets that preserve at least 75% of the contact MI are indicated in orange and green; those that preserve at least 90% are indicated in green.
[b]Reduced alphabets that perform as well or better than the optimized alphabets (Fig. 2) are indicated in red.

virtual alpha carbon backbone. We note, however, that two reduced alphabets, sizes 4 and 5 in Solis GBMR, contain substantial contact information (around 87%) while also embodying maximal local structure information. Not coincidentally, these two alphabets perform very well in independent evaluations. In particular: GBMR-4 was found by one study[11] to outperform all other reduced alphabets in one of three metrics (recall); GBMR-5 was found by another study[4] to outperform other reduced alphabets in RSMD for alignment accuracy; and GBMR-5 was found by another study[12] to be comparable in performance to their optimized alphabets.

We ask what the general commonalities are in the alphabet reduction schemes from the literature that successfully preserve contact structure information. A simple analysis was implemented to probe which amino acids consistently cluster together, and which subgroups always exist apart from one another. We counted the number of times each pair of amino acids occurred within the same cluster in those reduced alphabets in the literature found to preserve at least 75% of contact structure mutual information. (The frequency matrix is included as Supporting Information) A clear consensus emerges from this analysis. The most consistent subgroups are: {ILMV}, {FWY}, {DN}, {EKQR}, and {AST}. The amino acids within each subgroup are clustered together in the majority of the highly informative reduced alphabets found in literature. Moreover, none of the amino acids in the first two subgroups are ever found clustered together with any of the amino acids in the last three subgroups. (We should note here that when we re-implemented this clustering procedure with 80%, 85%,

**Table VII**
Summary of All Alphabets Examined

| Alphabet reduction approach | Total number of alphabets | Alphabets with %MI ≥ 90% | Alphabets with %MI ≥ 75% | Alphabets with %MI < 75% | Mean %MI preserved | Difference in <r> from optimal | Alphabets with ≤ <r> of optimal[a] | NOTES |
|---|---|---|---|---|---|---|---|---|
| Physicochemical considerations | 13 | 8 (62%) | 13 (100%) | 0 (0%) | 0.868 | 0.41% | 4 (33%) | Clustering based on similarities in the physicochemistry of amino acid side chains |
| Sequence alignments and sequence-structure alignments | 48 | 26 (54%) | 47 (98%) | 1 (2%) | 0.896 | 1.39% | 3 (6%) | Based primarily on alignments of homologous proteins of low sequence identity |
| Backbone/local structure/protein blocks | 29 | 3 (10%) | 18 (62%) | 11 (38%) | 0.674 | 4.65% | 0 (0%) | Includes reductions based on secondary structure, which may contain nonlocal information |
| Long range interactions | 24 | 8 (33%) | 13 (54%) | 11 (46%) | 0.766 | 2.32% | 7 (29%) | A subset of alphabets derived from Miyazawa-Jernigan potentials performs well |
| Totals | 114 | 45 (39%) | 91 (80%) | 23 (20%) | 0.809 | 2.26% | 14 (12%) | |

[a]The alphabets that appear to perform better than our optimal alphabets are within the standard error, and so it cannot be said that the differences are significant.

and 90% as the cutoff mutual information level, we found that the amino acids cluster together the same way.) These groups once again highlight the prominence of hydrophobicity, aromaticity, charge/polarity, and size in determining the interchangeability of amino acids in the context of long-range interactions. Consequently, an alphabet reduction scheme that follows this generalized recipe appears to be assured of preserving a significant level of contact mutual information residing in sequence. The remaining 4 amino acids, CGHP, appear to distribute themselves into different subgroups depending on the metric used to cluster the amino acids. Interestingly, these amino acids also possess unique properties, like disulfide bridge formation (C) and extremes in backbone flexibility (GP), and may be deployed in the sequence for fold-specific and function-specific roles.

## Simplified amino acid alphabet experiments

A number of studies have been successful in designing novel protein sequences that exhibit cooperative folding and functional properties. An effective strategy used in these studies to overcome sequence complexity is to employ reduced amino acid alphabets in formulating these folding sequences. Here, we ask how well the alphabets and sequence design rules used in these experimental design studies actually preserve structural information encoded in amino acid contacts in folded structures.

A class of design studies have exploited the canonical H/P patterning in amphiphilic helices and sheets in formulating folding sequences. To design for 4-helix bundles, Hecht *et al.*[30,31] used degenerate codons to generate combinatorial libraries that place DEHKNQ in polar positions and any of MLIVF in hydrophobic positions. Another design study[16,18] reduced the complexity of the chorismate mutase sequence using 9–14 amino acids set in clear binary pattern, situating NDEK in polar positions and FILM in hydrophobic positions. Both these amino acid groupings conform to the polar/hydrophobic dichotomy in all 2-letter reductions in the literature (Tables III–VI), including the classification obtained in this work (Fig. 2). Another study[21] designed a 20-residue polypeptide containing the three amino acids KIA which self-associates into a well-packed 4-helix bundle protein. The two hydrophobic amino acids (IA) and one charged amino acid (K) exist separately in all 2-letter alphabets examined in this work. Moreover, these three amino acids exist in separate clusters in our optimal clustering at cluster number 3 (Fig. 2), signifying that they may be ideal amino acid representatives of their clusters, at least as far as this type of helix packing is concerned.

For beta sheet proteins, the simple amphiphilic patterning of alternating H (any of the amino acids DEHKNQ) and P (any of the amino acids LIVF) residues produced the expected beta sheet monomers that oligomerized into amyloids.[19] In a subsequent study,[20] the

substitution of a lysine in a position normally occupied by the hydrophobic amino acids (i.e., the PHPKPHP pattern) for the edge strands in a 6-strand beta sheet protein discouraged oligomerization due to the need for lysine to be water accessible, forming instead monomeric six-stranded beta-sheet proteins. An attempt to design a mixed alpha-beta ($\alpha 3\beta 3$) protein scaffold *de novo*[57] also utilized binary H/P patterning using the amino acids AEKT for polar positions and ILV for hydrophobic positions. While this alphabet is distinct from that used by Hecht's group, the two groups also occur as distinct subgroups in the optimized 2-letter alphabet reduction derived in this work (Fig. 2). We note that in this particular design effort, A is clustered into the polar group, which a number of alphabet reduction schemes listed in Tables III–VI classify into the hydrophobic group.

These studies demonstrate the power and simplicity of designing novel sequences using the H/P sequence dichotomy for bulk positions within secondary structures, but also highlight the need for specifying sequences in key positions to ensure proper folding. Intervening turns that connect the helices have been designed to be rich in G to accommodate tight spacing, in addition to utilizing DENPS to promote the formation of turns and short loops. These amino acids were not explicitly part of the design's H/P library, but remained in the simplified sequences because they appear to be critical to inducing the intended fold.

Another set of studies focuses on asking whether a handful of specified amino acids is sufficient to preserve the fold of the protein scaffold. Work by Riddle *et al.*[14] among the first successful attempts to severely simplify sequences of extant folds, utilized a 5-letter alphabet AEGIK to reduce most sequence positions of the small beta-protein SH3 domain. Looking into the effectiveness of this reduced alphabet in preserving contact information, we modified the Monte Carlo clustering procedure to discover the optimal amino acid grouping would arise from the scenario where particular amino acids are pinned down into separate clusters. In this SH3 domain example, the clustering involved placing the amino acids in the reduced alphabet AEGIK into separate groups while allowing the other 15 amino acids to group themselves into these clusters to maximize contact mutual information. The results of this procedure for the SH3 domain redesign, as well as various design experiments, are shown in Table VIII. We observe that the optimal clustering around the alphabet AEGIK preserves 96% of the total contact information available in the optimal 5-letter alphabet (in Fig. 2). Upon closer inspection, we find that all of the amino acids that group into these 5 clusters have been substituted correctly with their respective AEGIK representative in the simplified SH3 domain sequence, with the exception of N and Q, both of which occur in the cluster containing G but were actually substituted far more often by E and K. When these two amino acids N and Q are moved into either the E cluster or the K cluster, the modified alphabets (also in Table VIII) appear to preserve 94% of the total contact mutual information, still close to optimal. The study notes that for positions where phylogenetic data suggested that one of the reduced alphabet residues might not be tolerated, additional residues were included in the sequence redesign. This included some positions across the sequence that contain any of the amino acids DNPSVWY.

In another study to reverse-engineer a beta-alpha barrel,[15] a reduction into the 7-letter alphabet AEFKLQV was feasible for 142 out of 182 positions without loss of function. To see how well this seven amino acid alphabet captures contact information, we again used the modified Monte Carlo clustering algorithm to yield the optimized clustering that corresponds to this particular partition. The optimized clustering that results from pinning down AEFKLQV into separate clusters, shown in Table VIII, recovers 98% of the maximum contact mutual information possible in alphabet size 7, a remarkably high level. We should note that, consistent with other design studies, the rest of the residue positions (40 out of 182) were found to be sensitive to mutation, in particular, at the interior of the beta barrel core, and also at some positions containing G and P which appear to act as critical secondary structure punctuations.[61,62]

Another study[17] deployed a 9-letter alphabet ADEGLPTVY to redesign 88% of the sequence positions of the 213-residue orotate phosphoribosyltransferase protein. These 9 amino acids can actually be condensed further into 6 well-delineated substitution clusters, as illustrated in Fig. 1 of their article.[17] The optimized reduced alphabet resulting from constraining the 6 amino acids into separate clusters, shown in Table VIII, recovers 91% of the maximum mutual information possible, a relatively high level. It is important to note that this is a solid example of a sequence design that uses a full alphabet reduction: substitutions follow clear rules, and are restricted by their membership within their respective clusters.

Solubility is a primary consideration in the design of nonmembrane proteins. One study[56] tested the solubility of random sequence libraries and found that those made from only five amino acids (ADEGV), which are thought to be abundant in the prebiotic environment, possess significant solubility. The optimized reduced alphabet resulting from constraining these 5 amino acids into separate clusters, listed in Table VIII, preserves 93% of the maximum contact mutual information possible. Succeeding work[58] confirms that a 12-letter reduced alphabet (ADEGIKMNRSTV) also has high solubility, and clustering the rest of the amino acids to form optimal clusters produces an alphabet reduction, also listed in Table VIII, that preserves 98% of the maximum contact mutual information. It appears that the reduced alphabets listed in Table VIII should be excellent choices for future de

**Table VIII**
Inferred Amino Acid Alphabets Arising From Experimental Work [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

| reference | Alpha. size | Amino acid alphabet* | Contact MI (nats) | *Z*-score for contact MI | Proportion of MI preserved** | <r> in CASP10 threading*** | notes |
|---|---|---|---|---|---|---|---|
| Kamtekar, 1993 & many subsequent studies (Hecht)[30] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | 1.000 | 0.2129 | Alphabet reduction that maximizes mutual information when DEHKNQ (polar) and FILMV (hydrophobic) are held in two distinct clusters |
| Riddle, 1997[14] | 5 | APT-CFILMVWY-DE-GNQS-HKR | 0.1221 | 5.66 | 0.960 | 0.1868 | Alphabet reduction that maximizes mutual information when AEGIK are held in different clusters |
| | | APT-CFILMVWY-DE-GS-HKNQR | 0.1202 | 5.54 | 0.945 | 0.1938 | As above, but N and Q moved into the K cluster |
| | | APT-CFILMVWY-DENQ-GS-HKR | 0.1191 | 5.47 | 0.936 | 0.1961 | As above, N and Q moved into the E cluster |
| Silverman, 2001[15] | 7 | A-CMV-DE-FWY-GKNPS-HQRT-IL | 0.1299 | 4.76 | 0.977 | 0.1944 | Alphabet reduction that maximizes mutual information when FVLAKEQ are held in different clusters |
| Taylor, 2001[16] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | 1.000 | 0.2129 | Alphabet reduction that maximizes mutual information when DEKN (polar) and FILM (hydrophobic) are held in two distinct clusters |
| Akanuma, 2002[17] | 6 | ACGST-DENQ-FWY-HKR-ILMV-P | 0.1198 | 4.71 | 0.915 | 0.2016 | Used 9-13-letter alphabet, which condenses into 6 clusters (Figure 1 of paper). |
| Doi, 2005[56] | 5 | AHPRT-CFILMVWY-D-E-GKNQS | 0.1186 | 5.44 | 0.932 | 0.1856 | Alphabet reduction that maximizes mutual information when AGVDE are held in different clusters |
| De la Osa, 2007[21] | 3 | AHPRT-CFILMVWY-DEGKNQS | 0.1162 | 7.85 | 1.000 | 0.2015 | Alphabet reduction that maximizes mutual information when AIK are held in different clusters. The optimized reduction here is the same as the unrestricted optimized reduction (Figure 2). |
| Jumawid, 2009[57] | 2 | ADEGHKNPQRST-CFILMVWY | 0.1068 | 10.40 | 1.000 | 0.2129 | Alphabet reduction that maximizes mutual information when AEKT (polar) and ILV (hydrophobic) are held in two distinct clusters |
| Tanaka, 2010[58] | 12 | A-CV-D-E-FMWY-G-HR-IL-K-N-PQS-T | 0.1367 | 3.09 | 0.980 | 0.1870 | Alphabet reduction that maximizes mutual information when ADEGIKMNRSTV are held in different clusters |
| Longo, 2012[59] | 10 | A-CMV-D-E-FIWY-G-HRT-KS-L-NPQ | 0.1313 | 3.49 | 0.953 | 0.1989 | Alphabet reduction that maximizes mutual information when ADEGILPSTV are held in different clusters |

*Amino acids that were used in the alphabet reduction experiments are indicated in blue.
**Reduced alphabets that preserve at least 90% are indicated in green.
***Reduced alphabets that perform as well or better than the optimized alphabets (Figure 2) are indicated in red.

novo design work, at least as far as mimicking contact information in naturally evolved folding sequences.

Finally, the question of alphabet reduction is also relevant to studies on the evolution of early abiotic proteins, when the primordial environment is thought to have contained only a subset of the current set of 20 amino acids. From a synthesis of current data,[59] a consensus has emerged as to the minimum folding set that could have ushered in proteogenesis, namely the 10-letter reduced alphabet ADEGILPSTV. Two observations with this set are: (1) no basic amino acids (HKR) are included, which made early proteins negatively charged, and (2) there are no aromatics (FWY), amino acids that are known to encourage stable hydrophobic cores in evolved proteins. Again, we subjected this 10-letter reduced alphabet to the modified clustering procedure, to ask how the other amino acids cluster around these 10 amino acids. The optimal clustering, in Table VIII, reveals that more than 95% of the contact mutual information can potentially be preserved by such an alphabet reduction. This means if the substitution rules specified by the reduced alphabets in Table VIII are followed, many folds can, in principle, be generated by just using these 10 amino acids, at least as far as generalized contact structure is concerned. As a demonstration of the viability of this prebiotic alphabet, a beta-trefoil fold sequence was successfully reformulated to include only 12–13 amino acids, while increasing the proportion of the 10-letter prebiotic amino acids up to 80%.[5] These proteins have an acidic *pI* and appear to require high salt concentration for cooperative folding, conditions thought to have existed in prebiotic environments.

## CONCLUDING REMARKS

In this work, we have designed a fully automatic amino acid alphabet reduction algorithm that considers as the objective function the mutual information residing in amino acid contacts in native folds. This algorithm has generated an optimal clustering of the 20 amino acids into smaller cluster numbers (from 2 to 19, shown in Fig. 2). As would be expected, the clustering pattern recovers amino acid properties such as hydrophobicity, charge, size, and aromaticity. Through extensive threading tests, we confirm the well-established link between mutual information of a particular descriptor and performance in fold recognition of the potential function

that uses that descriptor. We show that, on average, contact potentials that utilize the optimized alphabets perform best in fold recognition, highlighting their potential utility in coarse-grained fold recognition and structure prediction efforts.

We found that a significant amount of mutual information (around 75%) is preserved by the most rudimentary 2-letter reduction, consistent with diverse studies that point to the primacy of the H/P alphabet in determining folds and designing sequences *de novo*. We also found that much of the contact mutual information is preserved at significantly low alphabet sizes (the 5-letter alphabet captures around 90%, while a 9-letter alphabet nearly 96%), and that diminishing increments of information are achieved when expanding to higher alphabet sizes. We demonstrated, through the fold recognition tests, that virtually all of the effective contact structure information can be captured by a reduced alphabet of fewer than 10. The finding that contact-based discrimination can be achieved by a reduced alphabet means that a smaller number of amino acids can effectively mimic the essential tertiary interactions that stabilize native folds, a hypothesis that has broad implications for de novo protein design and also for questions of proteogenesis and protein evolution.

We examined numerous attempts in the literature to cluster amino acids in terms of their ability to preserve contact information. Remarkably, we found that a good proportion of these clustering schemes recover significant amounts of mutual information (Table VII), irrespective of the metric used to derive the clusters, suggesting that there is degeneracy in encoding long-range structure in sequence. It is no surprise that reduced alphabets derived from multiple sequence alignments and sequence-structure alignments preserve contact information at high levels, owing to the observation that well-established substitution patterns in homologous families manifest the same properties that appear to be important in tertiary structure. It is notable that alphabets derived from intuition that consider only physicochemical properties also preserve contact information effectively, suggesting a fundamental simplicity in the design of natural sequences. Integrating the clustering patterns of high-performing reduced alphabets reveals that high levels of contact information are preserved by any reduction built from the following subgroups {ILMV}, {FWY}, {DN}, {EKQR}, and {AST}, with none of the amino acids in the first two subgroups clustering with any of the amino acids in the last three subgroups.

Approaches based on local considerations seem to have the least success in capturing contact information, pointing to some frustration between the local and long-range folding codes. An exception is the reduced alphabet derived exclusively from the influence of local sequence on the protein backbone at alphabet sizes 4 and 5 (Table V, Solis 2000, GBMR),[10] which appear to be consistent with long-range structure considerations. It is no accident that these two alphabets have been singled out in independent comparative studies.[4,11,12] The early partitions of G and P due to their unusual structural properties are hallmarks of local-structure-based alphabet reduction schemes, even though they cluster unremarkably with other polar amino acids in this work (Fig. 2). Moreover, other amino acids that punctuate secondary structures and induce turns appear prominently in locally derived alphabets. Indeed, experimental sequence design studies, including those examined here, have identified positions that are sensitive to mutation and in some cases have explicitly specified G-rich linker regions that induce the intended fold, suggesting that an integrated local/long-range approach to alphabet reduction is the ideal direction for future work.

A primary challenge in *de novo* protein sequence design is the vastness of sequence space. Experimental evidence shows that alphabet reduction is a viable strategy to reduce sequence complexity. A directed design approach, which takes into account general topological properties of the target protein, has led to some notable successes, for instance the H/P redesign of particular proteins[16,18–20,30,31] that take advantage of simple periodicity rules as observed in amphiphilic alpha helices and beta strands. Another design method is the random sequence approach, greatly simplified by reduced alphabets that effectively reduce combinatorial complexity. To this end, Doi *et al.*[56] have demonstrated that using a handful of amino acids can produce proteins of sufficient solubility, a necessary condition for functional proteins. We find that the alphabets identified in a number of sequence design efforts are consistent with the results of our work. Alphabet reduction is also relevant to questions about the evolution of primordial proteins, as a limited number of abiotic amino acids may have existed during proteogenesis. To understand sequence design, both naturally in evolution and artificially in the lab, it may be a practical objective to uncover topology-specific folding rules. This may involve reducing most "bulk" positions with highly informative alphabets that preserve generalized tertiary interactions such as those derived in this work, and identifying determinant positions that appear essential to the particular fold, whether they are locally critical or tertiary stabilizing.

## REFERENCES

1. Hills RD, Lu L, Voth GA. Multiscale coarse-graining of the protein energy landscape. PLoS Comput Biol 2010;6:e1000827
2. Ghavami A, van der Giessen E, Onck PR. Coarse-grained potentials for local interactions in unfolded proteins. J Chem Theory Comput 2013;9:432–440.
3. Dayhoff MO, Schwartz RM. A model of evolutionary change in proteins. Dayhoff MO, editor. In: Atlas of protein sequence and structure. Washington, DC: Natl. Biomed. Res. Found; 1978. pp 345–352.

4. Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Prot Struct Funct Bioinform 2006;63:986–995.
5. Longo LM, Lee J, Blaber M. Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. Proc Natl Acad Sci USA 2013;110:2135–2139.
6. Floudas CA, Fung HK, McAllister SR, Mönnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: a review. Chem Eng Sci 2006;61:966–988.
7. Solis AD, Rackovsky S. Improvement of statistical potentials and threading score functions using information maximization. Prot Struct Funct Bioinform 2006;62:892–908.
8. Solis AD, Rackovsky S. Information and discrimination in pairwise contact potentials. Prot Struct Funct Bioinform 2008;71:1071–1087.
9. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.
10. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Prot Struct Funct Genet 2000;38:149–164.
11. Peterson EL, Kondev J, Theriot JA, Phillips R. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics 2009;26:1356–1362.
12. Bacardit J, Stout M, Hirst JD, Valencia A, Smith RE, Krasnogor N. Automated alphabet reduction for protein datasets. BMC Bioinform 2009;10:6.doi:10. 1186/1471-2105-10-6.
13. Stephenson JD, Freeland SJ. Unearthing the root of amino acid similarity. J Mol Evol 2013;77:159–169.
14. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Molec Biol 1997;4:805–809.
15. Silverman JA, Balakrishnan R, Harbury PB. Reverse engineering the (β/α) 8 barrel fold. Proc Natl Acad Sci USA 2001; 98:3092–3097.
16. Taylor SV, Walter KU, Kast P, Hilvert D. Searching sequence space for protein catalysts. Proc Natl Acad Sci USA 2001;98:10596–10601.
17. Akanuma S, Kigawa T, Yokoyama S. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA 2002;99:13549–13553.
18. Walter KU, Vamvaca K, Hilvert D. An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem 2005;280:37742–37746.
19. West MW, Wang W, Patterson J, Mancias JD, Beasley JR, Hecht MH. De novo amyloid proteins from designed combinatorial libraries. Proc Natl Acad Sci USA 1999;96:11211–11216.
20. Wang W, Hecht MH. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric β-sheet proteins. Proc Natl Acad Sci USA 2002;99:2760–2765.
21. de la Osa JL, Bateman DA, Ho S, González C, Chakrabartty A, Laurents DV. Getting specificity from simplicity in putative proteins from the prebiotic Earth. Proc Natl Acad Sci USA 2007;104:14941–14946.
22. Patel SC, Bradley LH, Jinadasa SP, Hecht MH. Cofactor binding and enzymatic activity in an unevolved superfamily of de novo designed 4-helix bundle proteins. Prot Sci 2009;18:1388–1400.
23. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
24. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Prot Struct Funct Genet 2000;38:3–16.
25. Solis AD, Rackovsky S. Information-theoretic analysis of the reference state in contact potentials used for protein structure prediction. Prot Struct Funct Bioinform 2010;78:1382–1397.
26. Solis AD. Deriving high-resolution protein backbone structure propensities from all crystal data using the information maximization device. PLoS ONE 2014;9:e94334doi:10. 1371/journal.pone.0094334.
27. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Prot Struct Funct Genet 2002;48:463–486.
28. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
29. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round X. Prot Struct Funct Bioinform 2014;82:1–6
30. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. Protein design by binary patterning of polar and nonpolar amino acids. Science 1993;262:1680–1685.
31. Go A, Kim S, Baum J, Hecht MH. Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles. Prot Sci 2008;17:821–832.
32. Mahler HR, Cordes EH. Biological chemistry. New York: Harper and Row; 1966.
33. Lehninger AL. Biochemistry. New York: Worth and Co.; 1970.
34. Dickerson RE, Geis I. Hemoglobin: structure, function, evolution, and pathology. Menlo Park: Benjamin/Cummings; 1983.
35. Taylor WR. The classification of amino acid conservation. J Theor Biol 1986;119:205–218.
36. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93: 11628–11633.
37. Landès C, Risler JL. Fast databank searching with a reduced amino-acid alphabet. Computer Applications in the Biosciences: CABIOS 1994;10:453–454.
38. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Prot Eng 2000;13:149–152.
39. Prlić A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Prot Eng 2000; 13:545–550.
40. Rogov SI, Nekrasov AN. A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences. Prot Eng 2001;14:459–463.
41. Cannata N, Toppo S, Romualdi C, Valle G. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. Bioinformatics 2002; 18:1102–1108.
42. Kosiol C, Goldman N, Buttimore NH. A new criterion and method for amino acid classification. J Theor Biol 2004;228:97–106.
43. Fan K, Wang W. What is the minimum number of letters required to fold a protein? J Mol Biol 2003; 328:921–926.
44. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Prot Eng 2003; 16:323–330.
45. Edgar RC. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Nucl Acids Res 2004;32:380–385.
46. Wrabl JO, Grishin NV. Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization. Prot Struct Funct Bioinform 2005;61:523–534.
47. Andersen CA, Brunak S. Representation of protein-sequence information by amino acid subalphabets. AI magazine 2004; 25:97
48. Etchebest C, Benros C, Bornot A, Camproux AC, De Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 2007;36:1059–1069.
49. Crippen GM. Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 1991;30: 4232–4237.
50. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
51. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Molec Biol 1999;6:1033–1038.
52. Cieplak M, Holter NS, Maritan A, Banavar JR. Amino acid classes and the protein folding problem. J Chem Phys 2001;114:1420–1423.
53. Liu X, Liu D, Qi J, Zheng WM. Simplified amino acid alphabets based on deviation of conditional probability from random background. Phys Rev E 2002;66:2. 021906
54. Esteve JG, Falceto F. A general clustering approach with application to the Miyazawa–Jernigan potentials for amino acids. Prot Struct Funct Bioinform 2004;55:999–1004.

55. Pape S, Hoffgaard F, Hamacher K. Distance-dependent classification of amino acids by information theory. Prot Struct Funct Bioinform 2010;78:2322–2328.
56. Doi N, Kakukawa K, Oishi Y, Yanagawa H. High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. Prot Eng Des Sel 2005;18:279–284.
57. Jumawid MT, Takahashi T, Yamazaki T, Ashigai H, Mihara H. Selection and structural analysis of de novo proteins from an α3β3 genetic library. Prot Sci 2009;18:384–398.
58. Tanaka J, Takashima H, Yanagawa H. Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Prot Sci 2010;19:786–795.
59. Longo LM, Blaber M. Protein design at the interface of the prebiotic and biotic worlds. Arch Biochem Biophys 2012;526:16–21.
60. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Prot Struct Funct Genet 1999;34:49–68.
61. Gunasekaran K, Nagarajaram HA, Ramakrishnan C, Balaram P. Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals. J Mol Biol 1998;275:917–932.
62. Risler JL, Delorme MO, Delacroix H, Henaut A. Amino acid substitutions in structurally related proteins a pattern recognition approach: determination of a new and efficient scoring matrix. J Mol Biol 1988;204:1019–1029.
63. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG, Haussler D. Information-theoretic dissection of pairwise contact potentials. Prot Struct Funct Bioinform 2002;49:7–14.
64. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
65. Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 1998;277:1141–1152.