PROTEINS: Structure, Function, and Bioinformatics 63:986–995 (2006)

Accuracy of Sequence Alignment and Fold Assessment Using Reduced Amino Acid Alphabets

Francisco Melo[1]* and Marc A. Marti-Renom[2]

[1]*Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile*

[2]*Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Mission Bay, University of California, San Francisco, California*

ABSTRACT **Reduced or simplified amino acid alphabets group the 20 naturally occurring amino acids into a smaller number of representative protein residues. To date, several reduced amino acid alphabets have been proposed, which have been derived and optimized by a variety of methods. The resulting reduced amino acid alphabets have been applied to pattern recognition, generation of consensus sequences from multiple alignments, protein folding, and protein structure prediction. In this work, amino acid substitution matrices and statistical potentials were derived based on several reduced amino acid alphabets and their performance assessed in a large benchmark for the tasks of sequence alignment and fold assessment of protein structure models, using as a reference frame the standard alphabet of 20 amino acids. The results showed that a large reduction in the total number of residue types does not necessarily translate into a significant loss of discriminative power for sequence alignment and fold assessment. Therefore, some definitions of a few residue types are able to encode most of the relevant sequence/structure information that is present in the 20 standard amino acids. Based on these results, we suggest that the use of reduced amino acid alphabets may allow to increasing the accuracy of current substitution matrices and statistical potentials for the prediction of protein structure of remote homologs. Proteins 2006; 63:986–995.** © 2006 Wiley-Liss, Inc.

Key words: statistical potentials; sequence alignment; fold recognition; reduced amino acid alphabets

INTRODUCTION

The configuration of a protein chain is determined by its primary sequence, which is a linear and asymmetric polymer made of combinations of the 20 naturally occurring amino acids. Protein structure and function are determined by their sequences and the surrounding environment. Therefore, the complex sequence of amino acids of a protein encodes for its diversity and specificity. Because different amino acid types share similar physicochemical properties and can be naturally substituted between protein sequences of the same family,[1] there have been several attempts to reduce the naturally occurring amino acid alphabet.[2–9] The problem lies on finding the proper grouping of amino acids that retains most of the information necessary for the integrity of the structure and function of proteins. During the past years, several theoretical works have suggested that the minimum number of amino acid types needed to encode for native proteins is less than 20.[2,4,10–12] In addition to this, some experimental works have demonstrated that engineered proteins with a reduced alphabet are able to preserve their fold and function. The Baker group, determined a stable and properly folded protein domain using a significantly reduced amino acid alphabet.[4] In this work, Baker and coworkers demonstrated that the SH3 fold could be preserved despite of the substantial reduction from 20 to 5 amino acid types for most of residues of the protein sequence. In a similar work, Stroud and coworkers designed a four-helix bundle protein of 108 residues long with a reduced alphabet of seven amino acid types.[2]

The most simplistic reduction of the amino acid alphabet described to date consists on the definition of two residue types, which is known as the hydrophobic-polar or HP model.[13,14] Other reductions have been proposed, by using a variety of different approaches to derive reduced amino acid alphabet representations. These approaches included genetic code mutations and optimization of hydrophobicity scales,[3] clustering of amino acids based on the pairwise similarity of distance-dependent energy terms from statistical potentials[5] and from amino acid substitution matrices,[6] minimization of mismatches between standard and reduced amino acid substitution matrices,[7] maximization of secondary structure prediction ability,[8] and local backbone conformation clustering of amino acids.[9]

Grant sponsor: FONDECYT; Grant number: 1051112; Grant Sponsor: Fundación Andes; Grant number: 13600/4; Grant sponsor: DIPUC; Grant Number: 2004/01PF.

*Correspondence to: Francisco Melo, Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Alameda 340, Santiago, Chile. E-mail: fmelo@bio.puc.cl

Received 12 April 2005; Revised 22 July 2005; Accepted 29 September 2005

Published online 27 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20881

© 2006 WILEY-LISS, INC.

TABLE I. Amino Acid Alphabets

ID	Alphabet description	Number of types	Reference
JO20	A-C-D-E-F-G-H-I-K-L-M-N-P-Q-R-S-T-V-W-Y	20	[15]
WW5	AHT-CFILMVWY-DE-GP-KNQRS	5	[7]
SR5	AEHKQRST-CFILMVWY-DN-G-P	5	[9]
MU4	AGPST-CILMV-DEHKNQR-FYW	4	[6]
MM5	AG-C-DEKNPQRST-FILMVWY-H	5	This work
RD5	100 randomly reduced alphabets	5	This work[a]

Each amino acid alphabet is described splitting the different amino acid types by a dash character. For clarity, and to simplify comparison, amino acid type clusters are sorted alphabetically based on the 20-letter code of the standard amino acid alphabet.
[a]Due to space restrictions, the 100 random alphabets are available as supplemental material.

Most of the reduced amino acid representations have been derived based on common physicochemical properties shared by different amino acids. These properties include hydrophobicity scale, size, flexibility, and common chemical groups present at the residue side chains.[1] Although these observations have a solid ground, the three-dimensional context where the residues are observed within a protein structure cannot be neglected. This is particularly true if the reduced amino acid alphabet is going to be applied to protein structure prediction.

A first objective in this work was to derive a new optimal and simplified amino acid alphabet based only on structural information. A second objective of this work was to compare the overall performance of several reduced amino acid alphabets (against the standard alphabet) in the areas that constitute a major component of current protein structure prediction methods: sequence alignment and fold assessment. To that end, based on several reduced protein amino acid alphabets of four and five residue types, we derived similarity matrices for sequence alignment and statistical potentials for fold assessment. The performance of the reduced amino acid alphabets were independently evaluated and compared to the full standard amino acid alphabet using a large set of known protein structures.

In the Methods section, we begin by describing the reduced amino acid alphabets used in this work; next, we describe the procedure used to generate a new optimally reduced amino acid alphabet of five residues; the benchmarking criteria and the reference sets used in the benchmark. In the Results, we assess the accuracy of the reduced alphabets in sequence alignment as well as fold assessment of protein structure models. Finally, we conclude by discussing the varying performance of the reduced amino acid alphabets for sequence alignments and fold assessment.

METHODS

Amino Acid Alphabets

In addition to the standard amino acid alphabet of 20 residue types[15] (JO20), several reduced alphabets were also used in this work (Table I). These alphabets can be categorized in three different groups: (1) optimally reduced alphabets previously generated by others and available in the literature, (2) an optimally reduced alphabet generated in this work by a genetic algorithm-based optimization, and (3) a set of randomly reduced alphabets to be used as a low-bound reference assessment frame. The single criterion adopted to select reduced alphabets from the literature was that they should have four or five residue types, because experimental work indicates that optimally reduced alphabets of about five residue types encode a large fraction of the original information that is present in the standard alphabet.[4]

Three previously described reduced amino acid alphabets of about five residue types were considered in this work. These included two amino acid alphabets of five residue types by Wang and Wang[7] (WW5) and by Solis and Rackovsky[9] (SR5), and a four amino acid alphabet by Murphy and coworkers[6] (MU4). The WW5 alphabet was optimized from the knowledge-based contact potential of Miyazawa and Jernigan[16] by minimizing the mismatches occurring between reduced matrices and the original Miyazawa and Jernigan matrix.[7] The SR5 alphabet was generated by minimizing the structural information loss (or maximizing the information gain) of collapsed alphabets in a large set of 4-mer peptides extracted from known protein sequences.[9] The MU4 alphabet was derived by maximizing the correlation between pairs of amino acids, based on the values of the BLOSUM50[17] similarity matrix.[6]

In addition to these existing alphabets, a new optimal alphabet of five residue types was generated and tested (MM5). This alphabet was derived by optimization with a genetic algorithm using structural information from remote homolog proteins (below).

Finally, 100 randomly reduced alphabets of five residue types (RD5) were generated as a lower bound reference frame. A detailed description of these random alphabets is available as supplemental material (http://protein.bio.puc.cl/protein-alphabets.html).

Generation of a Reduced Alphabet by Optimization with Genetic Algorithms

A genetic algorithm (GA) is a search technique to solve optimization problems. In a GA, abstract representations of candidate solutions (called individuals) encoded in

artificial chromosomes (strings of numbers in a computer), evolve toward optimal solutions in a given number of iterations (generations). At each iteration of the algorithm, a fitness that measures how well the individual solves the particular problem is calculated. Chromosomes evolve by using several types of operators of natural evolution, which include inheritance, mutation, selection, and recombination. The initial population consists of random individuals that evolve to the next generation by applying the GA operators and by biasing the selection of individuals by their calculated fitness. After a new generation is created, the fitness of all the population is evaluated and a set of its fittest individuals will contribute to the next generation/iteration of the algorithm.[18]

As a training set for the GA optimization, a total of 180 structural pairwise alignments were selected from the SCOP database[19,20] based on the following criteria: the pairwise sequence identity between the structurally aligned sequences was below 20% and the two aligned structures had more than 100 aligned residues with a global RMSD within 2.5 Å after optimal superposition. We then removed all gaps (i.e., insertions and deletions) from the alignments and concatenated them into a single pairwise alignment that resulted in a total of 25,140 pairs of structurally aligned residues. This long pairwise alignment constituted the training alignment for the genetic algorithm optimizer. Additionally, the residues in the training alignment were shuffled, generating random alignments of the same composition as the training alignment.

A genetic algorithm with chromosomes encoding for reduced amino acid alphabets was trained to maximize the difference between the pairwise sequence identities occurring in the training and the random alignments. The genetic algorithm had a constant population of 1000 chromosomes. Each chromosome was constituted of 20 genes arbitrarily representing a standard amino acid residue. Given a chromosome, the training and random alignments were translated from the standard 20 amino acid alphabet to the reduced alphabet that it encoded. Then, the sequence identity in the training and random alignments was calculated and its difference used as a fitness function of the chromosome adaptation. The genetic algorithm used elitism in each cycle, the evaluation was carried out using linear normalization, and the Roulette wheel parent selection method was adopted for the parent selection process previous to each reproductive step.[18] The reproduction process was carried out using double point crossing over and a rate of 1% chance per chromosome for the occurrence of mutation events after reproduction. This optimization protocol led to a five-residue amino acid alphabet with a fitness score of 23.8% (58.1% sequence identity in the training alignment minus a 34.3% sequence identity in the random alignment) after 100,000 iterations. The algorithm was executed for 100 independent runs, converging to the same final grouping of amino acids in the optimal chromosome for most of the cases. A list containing the alternative near-optimal alphabets along with their fitness scores is available as supplemental material (http://protein.bio.puc.cl/protein-alphabets.html).

Calculation of Reduced Amino Acid Substitution Matrices

A series of amino acid substitution matrices based on each of the reduced amino acid alphabets were built to evaluate their accuracy in sequence alignment. The mathematical derivation of the reduced amino acid substitution matrices was carried out from the Johnson and Overington (JO) frequency matrix[15] as follows: first, let i and j be the indexes of two standard amino acids ranging from 1 to 20. Let i' and j' be the indexes of a reduced amino acid alphabet ranging from 1 to R, where R is the total number of amino acid types in the reduced alphabet. Thus, any amino acid type index i' and j' will be constituted by a set of one or more indexes from the standard alphabet, such that the union of the i' or j' sets will generate the total set of indexes i or j, respectively. For each reduced alphabet, the frequency matrix F was calculated as follows:

$$F(i',j') = \sum_{i=1}^{20} \sum_{j=1}^{20} \mathrm{JO}(i,j) \;\; \forall\, i,j \in i',j'$$

where JO(i,j) represents the observed frequency in the JO substitution matrix for the amino acid pair ij.

Second, the frequency matrix was converted into a probability matrix P by:

$$P(i',j') = \frac{F(i',j')}{\sum_{j'=1}^{R} F(i',j')}$$

where R is the total number of residue types. Third, the probability matrix P was converted to the odds matrix O by:

$$O(i',j') = \frac{P(i',j')}{\left[\dfrac{\sum_{i'=1}^{R} F(i',j')}{\sum_{i'=1}^{R}\sum_{j'=1}^{R} F(i',j')}\right]}$$

Fourth, a scaled log-odds matrix L was calculated:

$$L(i',j') = 100 \times \log_{10}[O(i',j')]$$

Fifth, the scaled log-odds matrix was converted to a similarity matrix S' by the following procedure: the minimum value of the matrix was selected, the sign of it changed, and added to each element of the matrix. Thus, the matrix values were converted to the range [0 . . .N], where N is a positive value. Finally, the reduced similarity matrix S' was converted back to the 20 × 20 standard residue similarity matrix S by applying an inverse procedure as the one described above in the first step of the matrix calculation process:

$$S(i,j) = S'(i',j') \;\; \forall\, i,j \in i',j'$$

TABLE II. Optimal Gap Penalties for the Amino Acid Substitution Matrices

ID	Initiation gap penalty	Extension gap penalty
JO20	−280	−50
WW5	−80	−25
SR5	−180	0
MU4	−180	0
MM5	−140	0
RD5†	−80	0

[a]In the case of randomly reduced alphabets (RD5), the most observed initiation and extension gap penalties are shown. The exact penalty values for each random alphabet are available as supplemental material.

We also generated a scaled and converted standard amino acid substitution matrix from the JO original frequency matrix.

Sequence Alignment

The ALIGN command in the MODELLER program[21] was used to align two sequences by global dynamic programming algorithm.[22] All default parameters were used, and only the substitution matrix and the optimal gap penalties were changed for each run. The optimal gap initiation and extension penalties for each substitution matrix were empirically calculated by maximizing the average structural overlap for the training set of 118 pairwise structure alignments (below). The optimization was carried out in a grid of values spanning from −400 to 0 in steps of 20, and from −100 to 0 in steps of 5 for the initiation and extension gap penalties, respectively. The optimal gap penalties for each amino acid substitution matrix based on the reduced alphabets, including the most observed values for the 100 matrices based on the randomly reduced alphabets, are listed in Table II. Individual optimal gap penalties for each of the 100 random matrices are available as supplemental material at: http://protein.bio.puc.cl/protein-alphabets.html.

Statistical Potentials for Fold Assessment

Distance-dependent statistical potentials based on each amino acid alphabet were calculated to evaluate their accuracy in the assessment of comparative protein structure models. A total of 532 nonredundant protein chains from the Protein Data Bank (PDB)[23] were used to derive the statistical potentials. This set excluded any protein structure with duplicated or missing atoms. All proteins in the set shared 25% or less sequence identity among them and had a resolution higher than 3.0 Å.[24] The statistical potentials were calculated as previously described by Sippl,[25] but using the optimal parameters obtained in our recent work.[24] The following expression was used to calculate the potentials:

$$\Delta E^{ij}(l) = RT \ln[1 + M_{ij}\sigma] - RT \ln\left[1 + M_{ij}\sigma \frac{f^{ij}(l)}{f^{xx}(l)}\right]$$

where M_{ij} is the number of observations for the atomic pair ij and corresponds to:

$$M_{ij} = \sum_{l=1}^{n} f(i,j,l)$$

where σ is the weight given to each observation. We have used σ = 1/50 as proposed by Sippl,[25] such that on 50 observations $f^{ij}(l)$ and $f^{xx}(l)$ have the same weight for the calculation of $\Delta E^{ij}(l)$. The term $f^{ij}(l)$ is the relative frequency of occurrence of the atomic pair ij in the class of distance l and corresponds to:

$$f^{ij}(l) = \frac{f(i,j,l)}{M_{ij}}$$

where $f^{xx}(l)$ is the relative frequency of occurrence of all the atomic pairs in the class of distance l, and can be expressed as:

$$f^{xx}(l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} f(i,j,l)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} f(i,j,l)}$$

The temperature was set to 293 K, so that RT is equivalent to 0.582 kcal/mol.

Based on the energy obtained by the potentials described above, an energy Z-score of a model from the benchmark set of models was calculated for each of the newly derived statistical potentials. The energy Z-scores were calculated for the statistical potential energy of a model, using the mean and standard deviation of the statistical potential energy of 200 random sequences with the same composition and structure of the model as previously described.[24]

Benchmarking Criteria

The accuracy of an alignment was measured by relying on the aligned native structures extracted from the PDB.[23] First, the root-mean-square deviation (RMSD) between the corresponding Cα atoms in the two structures was calculated upon rigid-body least-squares superposition, as implemented in the SUPERPOSE command of MODELLER.[21] Second, the percentage of structurally equivalent positions was calculated as the number of the Cα atoms within a certain cutoff (i.e., 1, 2, 3, 4, and 5 Å, and their average) normalized by the length of the shorter of the two structures ("structure overlap"). Unless indicated otherwise, the structure overlap quoted is the average over all cutoffs. A statistical analysis of the differences between alignment accuracies of various methods was also performed. For this analysis, the alignment accuracy of a method was measured independently by the RMSD and the structural overlap after superimposition, both calculated as average accuracy for the 220 pairwise alignments in the testing set. The significance of the differences was computed using the Student's t-test statistics.[26]

TABLE III. Alignment Accuracy

	All testing set (220)		<40% Sequence Identity (55)	
ID	RMSD (Å)	Structure overlap (%)	RMSD (Å)	Structure overlap (%)
JO20	1.39 ± 2.45	80.0 ± 21.9	2.39 ± 3.62	70.4 ± 24.5
WW5	1.45 ± 2.46	78.9 ± 22.7	2.58 ± 3.52	66.8 ± 25.9
SR5	1.28 ± 2.14	79.7 ± 21.5	1.88 ± 2.84	69.5 ± 23.1
MU4	1.43 ± 2.46	79.2 ± 22.2	2.46 ± 3.57	69.4 ± 24.9
MM5	1.40 ± 2.40	79.1 ± 22.2	2.37 ± 3.42	67.8 ± 24.5
⟨RD5⟩	5.00 ± 6.61	73.0 ± 26.7	7.15 ± 5.57	50.5 ± 28.9

Average and standard deviation of the structure-based accuracy criteria are shown for the testing set of 220 alignments and for 55 alignments with sequence identity lower than 40%. Structure overlap represents the fraction of Cα carbons that are found at less than 3.5 Å after optimal superposition of two structures based on the pairwise sequence alignment generated. ⟨RD5⟩ represents the average performance of 100 randomly reduced alphabets of five amino acid types each.

The performance of classifiers based on the statistical potential *Z*-scores as a single feature was assessed by means of receiver operating characteristic (ROC) curves as previously described.[24] An ROC curve is obtained by plotting the false negatives fraction against the corresponding false positives fraction for all cutoffs on the energy *Z*-score. The area under the ROC curve represents the probability of incorrect classification over the whole range of cutoffs. This area is usually taken to be an important index because it provides a single measure of overall accuracy that is not dependent upon a particular feature threshold. The optimal classification threshold was also obtained for each statistical potential energy *Z*-score as the value where the highest positive prediction rate was observed.

Training and Testing Sequence Alignment Sets

Because our aim was to assess the accuracy of reduced amino acid alphabet matrices for aligning two sequences, the reference alignments were pairwise structure-based alignments. They were obtained from our comprehensive database of pairwise structure-based alignments, DBAli.[27] The alignments in DBAli were calculated by superposing all pairs of proteins of known structure in the PDB[23] with the program MAMMOTH.[28] The 38,579 chains in the PDB database (as of May, 2003) were clustered to remove redundancy. The final set of 6993 nonredundant chains did not superimpose to each other with a global RMSD smaller than 2 Å, had less than 80% of their Cα atoms within 4 Å and a difference in length larger than 30 residues. We randomly selected 400 pairwise structure-based alignments from the nonredundant set of alignments, which uniformly covered most of the spectra of sequence similarity (from 20 to 100% sequence identity). To avoid errors in the calculations of the accuracy of the alignments, all pairwise structural alignments with a chain that corresponded to an obsolete or incomplete PDB coordinates file were removed from the list of 400 pairwise alignments. The final pairwise alignments did not include regions of the chains that were not structurally superimposable. They included only those positions that were superimposed, plus insertions and deletions, removing additional parts of the chains such as long N- or C-terminal regions or additional domains. Finally, the list was randomly divided into two sets of alignments that maintained the uniformity of the sequence identity distribution. This resulted in a training set of 118 pairwise alignments that was used to optimize the gap penalties for each of the substitution matrices and a set of 220 pairwise alignments that was used in the benchmark presented here. The PDB chain identifiers, chain lengths, percentage sequence identities, RMSD for the aligned Cα atoms, average percentage of the aligned Cα atoms, and percentage of structurally equivalent residues are listed separately for the training and testing alignments in Supplementary Table I (http://protein.bio.puc.cl/protein-alphabets.html).

Testing Set of Comparative Models

To benchmark the accuracy of the reduced amino acid alphabets in protein structure fold assessment, we used a set consisting in 800 3D models divided in 400 correct and 400 incorrect models.[24,29] All correct models had a proper fold assignment and were built based on a relatively accurate sequence/structure alignment. Incorrect models either were built using a wrong fold or built based on the correct fold, but containing a large fraction of misalignments.

RESULTS

In an attempt to obtain optimally reduced representations of the standard amino acid alphabet that were exclusively based on structural information, we developed a new genetic algorithm (see Methods). The optimal solution evolved by the genetic algorithm consisted of a reduced alphabet of five types of amino acids. This new reduced alphabet, along with three other reduced alphabets published elsewhere, the standard alphabet of 20 residues and 100 randomly reduced alphabets were considered and assessed (Table I).

First, we proceeded with testing the accuracy of the alignments produced when using the amino acid substitution matrices based on the different reduced alphabets. The testing was carried out with a benchmark set of 220 pairwise structural alignments. The benchmarking set covered the whole range of sequence identity where pairwise alignments above 40% sequence identity may have

been trivial to reproduce. To assess the accuracy of the matrices with more difficult alignment pairs, we also calculated the average RMSD for 55 alignments in the testing set with sequence identities below 40%. The alignments obtained with the substitution matrix based on the standard alphabet (JO20) resulted in an average 1.4 Å RMSD after rigid superimposition of the two structures (2.4 Å RMSD for the difficult alignments). Similar average results were obtained with the other four nonrandom reduced alphabet matrices. In particular, the SR5 matrix produced the best alignments with 1.3 and 1.9 Å RMSD averages for all and difficult alignments, respectively (Table III). However, the differences observed between the JO20 and the reduced matrices for the RMSD accuracy measure were not statistically significant at a 95% confidence level in the Student's *t*-test [Fig. 1(A)]. The JO20 and the reduced alphabets matrices overperformed the random matrices (⟨RD5⟩) with statistical significance as demonstrated by the Student's *t*-test analysis [Fig. 1(A)]. The JO20 standard matrix outperformed with statistical significance all the reduced matrices (except matrix SR5) by ~1% for the structural overlap measure. All other matrices are statistically more accurate than the random matrix but cannot be distinguished among them [Fig. 1(B)]. The difference between MM5 and JO20 is only 0.95% structural overlap. However, the fact that the differences are favorable to the JO20 for most of the 220 pairwise alignments (171 of the alignments have equal or higher structural overlap and only 49 of them have a lower structural overlap) makes the difference in structural overlap statistically significant.

The discriminative power of statistical potentials based on reduced amino acid alphabets was also tested. In this benchmark, we assessed how much three-dimensional information of native folds is retained when the number of amino acids is reduced to five or four residue types. The benchmark consisted in the evaluation of fold assessment accuracy using a representative set of correct and incorrect comparative protein structure models spanning a large range of sequence length. The statistical potential energy *Z*-score was the measure used to separate correct from incorrect protein models. The statistical potential based on the standard alphabet (JO20) shows the best classification performance, irrespective of model size (Table IV). The performance was evaluated as the total positive predictive rate (percentage of correctly predicted cases) using the optimal threshold for classification in a large set of correct and incorrect protein models (800 models in total). The complete set was separated into four subsets based on the sizes of the protein models, which are shown in terms of total number of amino acids. Each subset contains a total of 100 correct and 100 incorrect models.

The positive prediction rate of reduced potentials for the large models is only between 1 to 3.5% points lower than the prediction accuracy of the JO20 potential. The major differences in assessment performance are observed for the very small models. It must be noted that, in general, those reduced alphabets derived from structural features

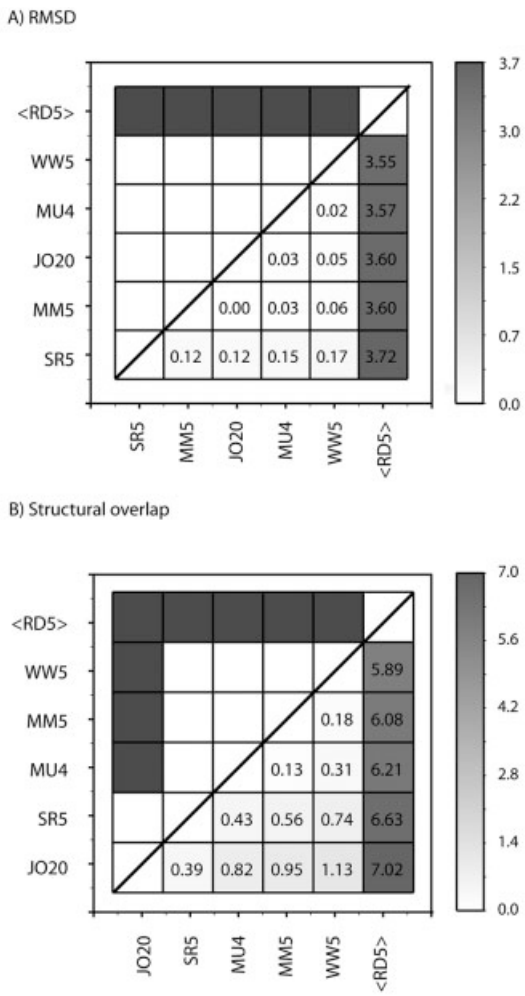

Fig. 1. Sequence alignment accuracies. Statistical significance of the differences in the alignment accuracies of the tested matrices. Upper diagonal: gray and white squares indicate pairs of matrices whose performance are and are not significantly different at a confidence level of 95%, respectively. Lower diagonal: the intensity of gray indicates the degree of the average difference between the corresponding matrices. (**A**) The accuracy of a matrix is measured as the average RMSD after rigid superimposition. (**B**) The accuracy of a matrix is measured as the average structural overlap after rigid superimposition. ⟨RD5⟩ represents the average performance of 100 randomly reduced alphabets of five amino acid types each.

of proteins (i.e., SR5 and MM5) lead to statistical potentials with better performances than other reduced alphabets. On the other hand, the potentials based on random alphabets exhibit a poor performance, approximating what would be expected from a random classification process.

When the performance of the statistical potentials was assessed over the whole range of possible classification thresholds (i.e., by means of receiver operating characteristic or ROC curves), a clearer picture emerges (Fig. 2). The JO20 potential had the highest specificity and sensitivity among all potentials. In other words, it was the best classifier irrespective of the chosen threshold. The MM5 statistical potential is the second best classifier for any

TABLE IV. Fold Assessment Accuracy

ID	Very small (0–50 AAs)	Small (50–100 AAs)	Medium (100–200 AAs)	Large (>200 AAs)	All
JO20	90.0	93.0	98.0	100.0	92.3
WW5	79.0	88.5	96.0	97.5	88.3
SR5	81.5	88.5	96.5	98.5	88.6
MU4	77.0	86.0	95.5	99.0	86.1
MM5	83.0	90.0	94.5	98.0	89.5
⟨RD5⟩	52.2 ± 1.4	57.2 ± 1.9	54.1 ± 2.0	52.1 ± 1.2	52.0 ± 0.9

The accuracy of fold assessment of statistical potentials based on the reduced amino acid alphabets was evaluated as the total positive predictive rate (percentage of correctly predicted cases) using the optimal threshold for classification in a large set of correct and incorrect protein models. The complete set was separated in four subsets based on the sizes of protein models. ⟨RD5⟩ represents the average performance and standard deviation of 100 randomly reduced alphabets of five amino acid types each.

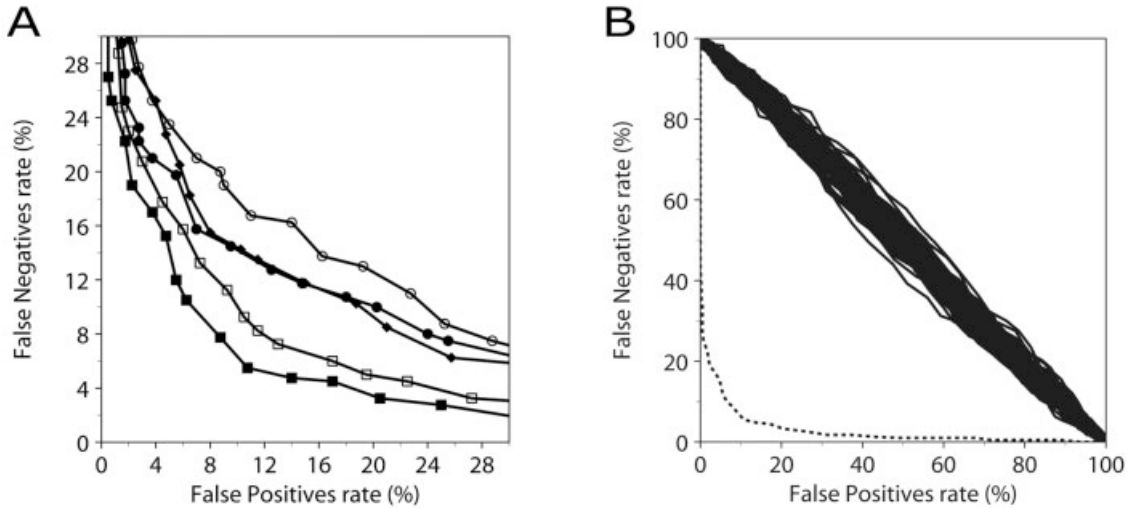

Fig. 2. Statistical potentials accuracies. ROC curves are used to assess the discriminative power of binary classifiers based on statistical potentials derived by using different amino acid types. (**A**) The statistical potential using 20 residue types (solid squares) is compared against statistical potentials based on optimally reduced alphabet definitions. The symbol correspondences are as follows: MM5 (open squares), SR5 (solid circles), WW5 (solid diamonds), and MU4 (open circles). (**B**) The statistical potential using 20 residue types from panel A (dashed line) is compared against 100 statistical potentials based on randomly reduced alphabet definitions (solid lines).

combined rates of false positives and false negatives. In third place, SR5 and WW5 exhibit a similar performance. The MU4 potential is the worst classifier for fold assessment, among the potentials based on reduced alphabets that were tested in this work. Finally, the reduced alphabets based on a random clustering of residues show a poor performance, comparable to that of a random classifier.

DISCUSSION

Several reduced amino acid alphabets have been described for many applications. The simplest is the two-letter amino acid alphabet, also known as the HP model.[30] This reduced alphabet clusters hydrophobic and polar residues into different groups and has been widely used to study protein folding in simplified 2D and 3D lattice models.[13,31] Despite its extreme simplicity, this alphabet probably describes the most important force that stabilizes a native protein structure: the partition of polar and hydrophobic residues that takes place in presence of water. Moreover, it allows complete enumeration of the sequence and structure spaces, a desirable property of a model to study and simulate the kinetics and thermodynamics of protein folding. However, in real-world applications, the HP alphabet is not accurate enough to properly describe the complex native protein structures. We have already demonstrated that the performance of statistical potentials is highly compromised when using the HP alphabet instead of the standard alphabet of 20 amino acids.[24] Although some arguments have been given that support the idea that statistical potentials performance is just a consequence of the partition propensity of residues in water,[32] this single argument does not account for the large differences in accuracy that are observed between the standard and the HP statistical potentials for fold assessment of real protein structures.[24] Therefore, substantial reductions on the amino acid alphabet, which are not as extreme as the HP model, could still be useful approximations. In this work, we have assessed the performance

for sequence alignment and fold assessment of some reduced amino acid alphabets and carried out a comparison against the standard alphabet. These two benchmarks were chosen because they constitute essential core elements of any protein structure prediction method and are also involved in other important applications such as functional genome annotation.

Reduced amino acid alphabets can aid protein structure prediction methods. On the sequence space, a 20-letter code can generate a vast number of possible sequences, even for very short proteins. On the structural space, adding several conformations to each residue, the possible protein structure conformations are even larger. In the simplest case (i.e., a system described in a pairwise fashion), the accurate understanding of the energy forces governing the occurring interactions begins with the description of 400 (20×20) or 210 ($20 \times 21/2$) energy functions for asymmetric and symmetric interactions, respectively. Each energy function, in turn, will require additional variables or parameters to be properly described. To increase the accuracy of the energy functions, a further dimensional growth of the matrix that represents the system is required. Therefore, any statistical approach based on known experimental data will need to fill in most of the matrix bins to properly describe the system. Unfortunately, because the experimental data is finite, in practical terms there is a need to simplify the matrix used to describe the system. Thus, reducing the amino acid alphabet would allow a more detailed exploration of other properties in protein structures that could become relevant but have not been yet explored due to the outlined limitations. In this scenario, an important question arises: what reduction of the amino acid alphabet will allow us to describe with high accuracy the energy forces that govern protein structure stability? The answer to this question is ambitious, and it has not been yet fully addressed in the literature. Here we tried to address this question by studying the accuracy of sequence alignment and fold assessment methods using optimally reduced amino acid alphabets.

We first focused on the use of amino acid substitution matrices derived from reduced alphabets to align two sequences. Our intention was to assess how the reduction of the amino acid alphabet affected the accuracy of the alignments produced. The results of this work showed that a small drop in accuracy is observed for sequence alignments generated with substitution matrices derived from reduced amino acid alphabets. The reduced matrices still encode enough information to outperform randomly generated matrices. The reduced similarity matrices could be used as a starting point for incorporating additional information into the alignment process. For example, information derived from the 3D structure of one of the aligned sequences, such as the environment-dependent substitution matrices[33] is likely to further improve the utility of sequence-structure alignment in comparative modeling applications. This additional information could now be included and be computationally treatable over a reduced number of amino acid types from 20 to just 5, which translates into an overall reduction of matrix bins from 210 to 15, respectively.

Second, we focused on the use of statistical potentials derived from reduced amino acid alphabets to assess the accuracy of protein structure models. One of the major problems in deriving statistical potentials is the limited size of the database of known protein structures and the large number of parameters required to properly define an interaction of two or more bodies in three-dimensional space. This limitation could be overcome by reducing the size of the amino acid alphabet. Thus, a more detailed description of the interactions describing native protein structures could be achieved. The fact that the statistical potential based on the JO20 alphabet only clearly outperforms other potentials based on reduced alphabets for the very small models, may suggests that a large number of interactions is required to properly assess a protein model when a potential based on a reduced amino acid alphabet is used. This is in agreement with our previous observations, where the increase of the distance range of a pairwise potential substantially improved its performance when assessing very small models (i.e., the total number of interactions is increased and more confidence is gained in the total observed energy of the model).[24]

As it should have been expected before hand, the JO20 potential exhibited the best performance in model assessment, irrespective of the sensitivity/specificity balance at any given classification threshold. The potential based on the novel MM5 alphabet was the second best classifier for fold assessment. Then, the SR5 and WW5 potentials exhibited an overall similar performance, although that for high specificities the SR5 potential was more accurate. Finally, the MU4 potential showed the worst performance in fold assessment among all the reduced potentials tested. However, it has been previously demonstrated that the information content of amino acid alphabets decreases when the total number of clusters or residue types are reduced from five to four types.[9,34] This may explain the differences in accuracy observed for the potential based on the MU4 alphabet compared to the other optimally reduced alphabets. Therefore, it is nearly impossible to fairly compare reduced alphabets of different sizes. The potentials derived from the JO20, MM5, and SR5 alphabets show similar behavior and capabilities required for high specificity (i.e., in automated and large-scale protein structure fold assessment that requires a good detection of false positives). This finding suggests that potentials based on optimally reduced alphabets with additional three-dimensional features may exhibit an even higher specificity for automated protein fold assessment, compared to the assessed potential based on the standard alphabet.

The differences for clustering amino acids into different groups may explain the varying accuracy of potentials based on different optimally reduced alphabets. For example, the SR5 and MM5 alphabets cluster into different groups two structurally important amino acids such as glycine and proline. The SR5 alphabet creates two separate groups containing these residues as their unique

members. In contrast, for the MM5 alphabet, glycine clusters into a group with alanine, the second smallest amino acid, and proline clusters within a large group constituted of polar amino acids. It is somehow surprising that WW5 and MU4 alphabets have glycine and proline defined within the same cluster. In the case of WW5, these two residues are clustered as unique members of their group. The cysteine amino acid, also a structurally important residue in proteins, is differentially clustered by the optimally reduced alphabets. The MM5 alphabet clusters this amino as the unique member of its group. The other reduced alphabets cluster the cysteine amino acid in groups containing hydrophobic residues. Most of the polar and hydrophobic amino acids are clustered into different groups, with the exception of the MU4 alphabet, which clusters the hydrophobic residues into two different groups. Therefore, most of the differences between the optimally reduced alphabets arise from the partitioning of glycine, proline, cysteine, and polar residues. Those differences may partially explain the varying accuracy of reduced alphabets for fold assessment. Some random alphabets successfully clustered glycine, proline, and cysteine into different groups, but none of them properly separated hydrophobic and polar residues (supplemental material).

A significant reduction of the standard amino acid alphabet to five or four different types did not result in a substantial reduction of performance for sequence alignment and model assessment. Although varying performances were observed, this is true for most of the optimally reduced amino acid alphabets tested in this work. This finding opens the possibility of expanding the description of the system to derive new amino acid substitution matrices and statistical potentials. To mention some: structural features such as secondary structure preferences could be added to the substitution matrices; statistical potentials would not be blind to the orientation between residues in three-dimensional space; distance-dependence and secondary structure or local backbone conformation could be considered simultaneously; three body interactions could be better described by having several observations of each particular state; etc. We will certainly continue the exploration of some of these new variables to seek for better methods that may prove useful for protein structure prediction.

CONCLUSIONS

We have assessed the performance of reduced amino acid alphabets for the tasks of sequence alignment of remote homologs and fold assessment of protein structure models. Residue–residue substitution matrices and statistical potentials based on several optimally reduced alphabets were calculated and tested using a large benchmark of protein sequences and structures. Based on the results obtained, we draw the following conclusions:

1. residue–residue substitution matrices and statistical potentials calculated on the basis of optimally reduced alphabets exhibit a similar performance as those based on the standard alphabet in the tasks of sequence alignment of remote homologs and fold assessment of protein structure models.
2. The little loss of performance or accuracy of substitution matrices and statistical potentials based on optimally reduced alphabets is clearly compensated by the significant reduction of matrix bins that is achieved (i.e., a reduced alphabet of five residue types leads to a 14 or 16 times reduction of the total number of matrix bins, for symmetric and asymmetric matrices respectively).

ACKNOWLEDGMENTS

We are grateful to Prof. Andrej Sali for all his support, encouragement, and access to his CPU Linux cluster. We would like to thank Dr. M.S. Madhusudan for his suggestions. We acknowledge the helpful comments and suggestions made by the two anonymous reviewers of this manuscript.

REFERENCES

1. Creighton TE. Protein folding. 2nd ed. New York: W.H. Freeman Company; 1993.
2. Schafmeister CE, LaPorte SL, Miercke LJ, Stroud RM. A designed four helix bundle protein with native-like structure. Nat Struct Biol 1997;4:1039–1046.
3. Trinquier G, Sanejouand Y. Which effective property of amino acids is best preserved by the genetic code? Protein Eng 1998;11: 153–169.
4. 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 Biol 1997;4:805–809.
5. Kuznetsov IB, Rackovsky S. Discriminative ability with repect to amino acid types: assessing the performance of knowledge-based potentials without threading. Proteins 2002;49:266–284.
6. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
7. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
8. Andersen CAF, Brunak S. Representation of protein-sequence information by amino acid subalphabets. AI Magazine 2004;25:97–104.
9. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins 2000;38:149–164.
10. Esteve JG, Falceto F. A general clustering approach with application to the Miyazawa-Jernigan potentials for amino acids. Proteins 2004;55:999–1004.
11. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Protein Eng 2003;16:323–330.
12. Fan K, Wang W. What is the minimum number of letters required to fold a protein? J Mol Biol 2003;328:921–926.
13. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
14. Wolynes PG. As simple as can be? Nat Struct Biol 1997;4:871–874.
15. Johnson MS, Overington JP. A structural basis for sequence comparisons. An evaluation of scoring methodologies. J Mol Biol 1993;233:716–738.
16. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
17. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
18. Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley; 1989.
19. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
20. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:226–229.

21. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
22. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
23. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D 2002;58:899–907.
24. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
25. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213: 859–883.
26. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure 2002;10:435–440.
27. Marti-Renom MA, Ilyin VA, Sali A. DBAli: A database of protein structure alignments. Bioinformatics 2001;17:746–747.
28. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
29. Sanchez R, Sali A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 1998; 95:13597–13602.
30. Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. A test of lattice protein folding algorithms. Proc Natl Acad Sci USA 1995;92:325–329.
31. Chan HS, Dill KA. Comparing folding codes for proteins and polymers. Proteins 1996;24:335–344.
32. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
33. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1992;1:216–226.
34. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Proteins 2002;48:463–486.