Protein Engineering vol.13 no.3 pp.149–152, 2000

Simplified amino acid alphabets for protein fold recognition and implications for folding

Lynne Reed Murphy, Anders Wallqvist and Ronald M.Levy[1]

Department of Chemistry, Rutgers University, Wright-Rieman Laboratories,
610 Taylor Road, Piscataway, NJ 08854-8087, USA

[1]To whom correspondence should be addressed
E-mail: ronlevy@lutece.rutgers.edu

Protein design experiments have shown that the use of specific subsets of amino acids can produce foldable proteins. This prompts the question of whether there is a minimal amino acid alphabet which could be used to fold all proteins. In this work we make an analogy between sequence patterns which produce foldable sequences and those which make it possible to detect structural homologs by aligning sequences, and use it to suggest the possible size of such a reduced alphabet. We estimate that reduced alphabets containing 10–12 letters can be used to design foldable sequences for a large number of protein families. This estimate is based on the observation that there is little loss of the information necessary to pick out structural homologs in a clustered protein sequence database when a suitable reduction of the amino acid alphabet from 20 to 10 letters is made, but that this information is rapidly degraded when further reductions in the alphabet are made.

Keywords: minimal alphabet/protein fold recognition/sequence alignment

Introduction

A cell requires a large number of different proteins to execute and regulate cellular processes. Even though the structures of these individual proteins are highly complex and diverse on the atomic level, it is believed that there exists a finite number of protein folds. The constituent building blocks of these proteins are the 20 naturally occurring amino acids. It is from this set of amino acids that polypeptide chains are formed in the cell, which in turn rapidly fold into well-defined three-dimensional structures. From a combinatorial standpoint there is an almost endless variety of sequences that can be made from a 20-letter code, e.g. for a polypeptide chain of length 100 there are 20^{100} possible combinations. Of course, only a fraction of these chains can find unique and stable folds (Levinthal *et al*., 1975), a prerequisite for biological functionality. From the work of several groups investigating protein folding, it is strongly suggested that protein folding can be achieved with far fewer components than the 20 naturally occurring amino acids (Sander and Schulz, 1979; Regan and Delgrado, 1988; Heinz *et al*., 1992; Betz *et al*., 1993; Kamtekar *et al*., 1993; Davidson *et al*., 1995; Riddle *et al*., 1997; Plaxco *et al*., 1998). The simplest code is a division into non-polar and polar residues employed for folding four-helix bundles as illustrated by Kamtekar *et al*. (1993). These studies indicate that the underlying rule of a polar outside and a non-polar inside is the governing design criterion for this class of proteins (Kamtekar *et al*., 1993). Mutations within either the polar or non-polar groups are tolerated but mutations between them generally are not. In fact, the mutational tolerance of proteins is often high in many regions of the sequence (Matthews, 1993), although it is known that for many sequences there are a few key residues that must be strictly conserved for the protein to fold and function. In general, the allowed mutations follow from intuitive physical principles, e.g. hydrophobic groups and hydrophilic groups each tend to be conserved as a class, small residues are not replaced by large ones in the interior of a protein, etc. In addition to α-helical bundles, the recent work of Riddle *et al*. (1997) showed that for a small 57-residue β-barrel-like protein, 38 out of 40 targeted amino acids could be reduced to just a handful of residues. A combinatorial chemistry approach allowed the experimentalists to sample a wide range of possible mutations that could code for both foldability and function. Two fully functional constructs were detected in which 38 out of 40 selected sites mutated into five residues, I, K, E, A and G. These sequences yielded rapidly folding proteins that were viable *in vitro*.

Taken together, these experimental results suggest that there may exist reduced amino acid alphabets which could be used to fold many proteins by making appropriate substitutions in the original sequences. It is difficult to test this hypothesis directly on a sufficiently large number of proteins representative of the known folding families, but we can obtain insights into this problem by studying sequence patterns which characterize these protein folding families, and the effects of alphabet reduction on these sequence patterns.

A central problem of structural genomics is to predict the folding family given a newly sequenced gene. In many instances this can be accomplished by aligning the ‘query’ sequence against a database of sequences which have been clustered into folding families according to structural criteria. We have analyzed how this sequence-based fold recognition procedure depends on the size of the amino acid alphabet from which the sequences were constructed. As discussed below, we observe the fold recognition is minimally degraded when the amino acid alphabet is reduced from 20 to 10 letters by appropriately grouping chemically similar amino acids, but the sequence-encoded information needed to differentiate folding families is rapidly degraded when further reductions in the alphabet are made.

The analogy between protein folding and protein fold recognition is, of course, incomplete. A sequence which has no detectable homology within a clustered database of proteins may well fold to a structure whose family is represented in the database – divergent evolution has produced many such examples. In this sense an estimate of the minimal alphabet size based on homology detection by sequence alignments represents an upper bound. Conversely, there is no guarantee that a synthetic sequence will actually fold even if its alignment scores against sequences in a particular family are very much larger than can be expected by chance – there could be kinetic barriers to the folding of such sequences, for example. Yet it is intriguing to suggest that if it is possible to construct representative sequences from a large number of different folding families which will fold using a reduced set of amino acids (a reduced alphabet), then the alignment scores of the corresponding sequence pairs will be very much larger than expected by chance. That is to say, if the different sequence patterns that encode for a diversity of folding families can be preserved when the sequences are synthesized using reduced alphabets, then it should be possible to probe this relationship by carrying out computer alignment experiments on sequences constructed with reduced alphabets and comparing these results with those for the parent native sequences.

In this work we evaluate the extent to which sequence patterns derived from reduced alphabets preserve the information needed to detect homologs in a clustered database. The amino acid reduction scheme is based on the analysis of correlations among similarity matrix elements used for sequence alignments. We find that as the alphabet size is reduced, the information encoded in the amino acid sequences responsible for protein fold recognition is degraded. We estimate that for proteins of many different families a minimal alphabet requires 10–12 letters.

Materials and methods

The alignments between protein sequences were performed using the global alignment algorithm of Myers and Miller (1988) as coded in the FASTA program package suite by Pearson (1990). The BLOSUM50 matrix derived by Henikoff and Henikoff (1992) was used for alignments based on the 20-letter alphabet and as the starting point for constructing the reduced similarity matrices. The gap insertion and elongation parameters used for alignments were set to –12/–2.

The 20-letter amino acid alphabet is reduced to smaller alphabets based on correlations indicated by the BLOSUM50 similarity matrix, i.e. amino acid pairs with high similarity scores are grouped together. The procedure for grouping like amino acids together is as follows: first, the correlation coefficients between similarity matrix elements are calculated for all pairs of amino acids, i.e. for alanine (A) and valine (V) the coefficient would be evaluated as

$$^{C}\mathrm{A,V} = \frac{\Sigma_{i=1}^{20} M_{\mathrm{A},i} \cdot M_{\mathrm{V}i}}{(\Sigma_{i=1}^{20} M_{\mathrm{A},i} \cdot M_{\mathrm{A},i})\,(\Sigma_{i=1}^{20} M_{\mathrm{V},i} \cdot M_{\mathrm{V},i})} \tag{1}$$

the summation of i being taken over the 20 amino acids; second, the two amino acids with the highest correlation coefficient are grouped together, then the pair with the next highest correlation is either added to the first group if one member is already in the group or separated into a new group if not, and the process is repeated until all the amino acids are divided into the desired number of groups.

Reduction of the similarity matrix based on the groupings is performed by calculating new matrix elements as the average of the appropriate old similarity matrix elements. For example, the score between a group consisting of (A) and one consisting of (ST) is computed as the average of the A–S and A–T terms. Thus whereas in the original similarity matrix an alignment of A with S contributes M_{AS} to the overall alignment score, in the new similarity matrix the contribution is $M_{12} = (M_{\mathrm{AS}} + M_{\mathrm{AT}} + M_{\mathrm{SA}} + M_{\mathrm{TA}})/4$, and alignment of A with T is equivalent to that of A with S.

Alphabet reductions derived from the similarity matrix are shown in Figure 1. The complete group of reduced alphabets studied in addition to those delineated in the figure are as follows: 3 letters, [(LASGVTIPMC), (EKRDNQH), (FYW)]; 5 letters, [(LVIMC), (ASGTP), (FYW), (EDNQ), (KRH)]; 6 letters, [(LVIM), (ASGT), (PHC), (FYW), (EDNQ), (KR)]; 12 letters, [(LVIM), (C), (A), (G), (ST), (P), (FY), (W), (EQ), (DN), (KR), (H)]; and 18 letters, [(LM), (VI), (C), (A), (G), (S), (T), (P), (F), (Y), (W), (E), (D), (N), (Q), (K), (R), (H)]. These groupings are similar to those previously proposed from examining amino acid side-chain properties (Miyata *et al.*, 1979; Santibanez and Rohde, 1987) and other similarity matrices (Collins and Coulson, 1987; Risler *et al.*, 1988; Landes and Risler, 1994).

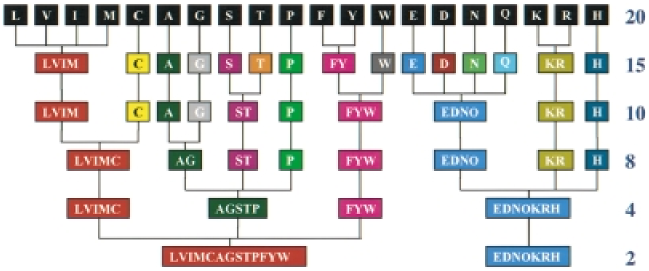

Fig. 1. Schemes for reducing amino acid alphabet based on the BLOSUM50 matrix by Henikoff and Henikoff (1992) derived by grouping and averaging the similarity matrix elements as described in the text. The most correlated amino acids naturally form groups which have similar physiochemical properties. Hydrophobic residues, especially (LVIM) and (FYW), are conserved in many reduced alphabets, as are the polar (ST), (EDNQ) and (KR) groups. The most basic alphabet reduces to two groups that can be categorized broadly as hydrophobic/small (LVIMCAGSTPFYW) and hydrophilic (EDNQKRH).

Homology detection with reduced alphabets was tested using ‘all-against-all’ alignments of sequences within the SCOP40 database, extracted by Brenner *et al.* (1998) from SCOP (Murzin *et al.*, 1995) (version 1.36) and representing all distantly related proteins in the Protein Data Bank with an amino acid identity of 40% or less. The total number of sequences is 1323, which are divided into 639 homologous superfamilies. Detection of homology, i.e. identification of the superfamily for each sequence in the database, is illustrated by coverage as a function of errors per query, for a set of expectation value thresholds in Figure 2 (inset). The coverage is defined as the number of homologous pairs detected divided by the total number of homologous pairs present in the database. For the SCOP40 database there are a total of 9044 homologous pairs to detect among 1 750 329 aligned sequence pairs. The error per query is defined as the total number of non-homologous protein sequences detected with expectation values equal to or greater than the threshold divided by the total number of aligned sequence pairs. Plots of error per query (EQP) versus coverage were constructed for each reduced alphabet. These results were constructed by systematically varying the *e*-value cutoffs used to identify homologous sequences. The effects of alphabet size on fold-recognition among the SCOP40 sequences shown in Figure 2 correspond to searching SCOP40 at an EPQ of 0.001. The low error rate (0.001) effectively eliminates error in the analysis shown in Figure 2 arising from the misassignment of fold sequences to folding families. Using less stringent *e*-value cutoffs, corresponding to an EPQ of 0.01, does not have a significant effect on the results shown in the figure.

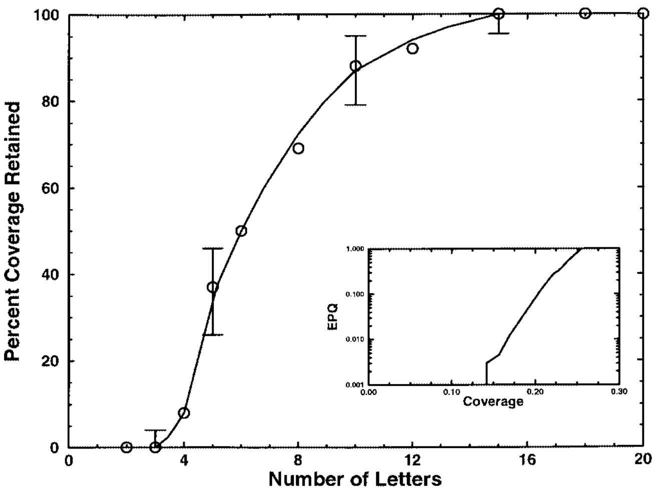

Fig. 2. Retention of coverage relative to the 20-letter alphabet as a function of the number of amino acids in the alphabet. The solid black line is the average retention of homology detection for the reduced alphabets studied at an errors per query (EPQ) value of 0.001. While two groups, hydrophilic and hydrophobic, are not sufficient to retain any homology detection at this error level, the retention increases steeply with the number of letters in the alphabet. The increase is much faster than a linear, additive model. The homology detection with 10–12 letters is >90% of that observed with 20 letters. The error bars in the figure were derived from looking at the variation of coverage retained with five separate groups of proteins: all α-helical protein domains, all β-sheet domains, an α/β class containing mainly parallel β-sheets, an α + β class of mainly antiparallel β-sheets and a small protein class. Inset: homology detection with the SCOP40 database. The graph shows the coverage versus EPQ within the SCOP40 database for the naturally occurring amino acid alphabet.

Results

Many schemes for constructing reduced alphabets based on chemical similarity have been proposed. The specifics of one reduced alphabet for a 57-residue src SH3 domain were obtained by Riddle *et al.* (1997) using a phage display selection strategy. This protein has a β-barrel-like structure and provides a binding site for proline-rich peptides (Feng *et al.*, 1994). Residues involved in the binding of substrate were not targeted for mutations to allow for the use of a binding assay to ascertain functionality of the mutated proteins. The mutation in the allowed regions resulted in two folded and functional proteins, FP1 and FP2. Five residues (IKEAG) represent the minimal alphabet for the sequence regions targeted for mutations. The mutations from WT to the two folded and functional sequences can be categorized in terms of mutations between hydrophilic residues (EDNQKRH), large hydrophobes (LVIF) and small residues (AGSTP). Mutations within these groups are more frequent than between groups, although some mutations cross groups.

In this work we formulate an amino acid reduction scheme based on the analysis of correlations among similarity matrix elements used for sequence alignments. Our procedure averages the matrix elements of the most closely related residues, and constructs reduced similarity matrices using these average values. Alphabet reductions derived from the underlying similarity matrix as explained in the Materials and methods section are shown in Figure 1.

This reduction of the amino acid alphabet followed several clearly recognizable paths, i.e. initially residues with similar physical/chemical properties are grouped together; large hydrophobes (LVIM), amino acids with large and mainly hydrophobic aromatic side chains (FY[W]) and long-chain positively charged residues (KR). The 10-letter alphabet of Figure 1 contains five amino acid groups, including the three groups mentioned plus two additional hydrophilic groups, alcohols (ST) and charged/polar residues (EDNQ). Further reductions of the 20-letter code coalesce smaller residues, and ultimately the code reduces to two basic groups, hydrophobic and hydrophilic. The initial divisions of the reduced alphabets are similar to the five-letter code of Riddle *et al.* (1997), i.e. the I, K, E, A, G alphabet found for the β-barrel-like protein. Down to and including the level of reduction corresponding to the 10-letter alphabet in Figure 1, these amino acids are maintained in separate groups. Thus the specific example of the reduced alphabet that Riddle *et al.* determined for SH3 is consistent with the proposed simplification scheme through 10 letters, but not below 10 because of AG pairing at the eight-letter level. Figure 1 serves as a guide to the construction of reduced alphabets which may be useful for correlating sequence patterns and folding patterns in a *statistical* sense as observed across a large number of folding families rather than as a recipe for constructing a particular protein using a very small alphabet.

Having constructed similarity matrices corresponding to reduced alphabets, we proceed to evaluate the extent to which sequence patterns derived from the reduced alphabets preserve the patterns needed to detect homologs in a clustered database. For this analysis we use the SCOP40 clustered database (Murzin *et al.*, 1995; Brenner *et al.*, 1998), containing 1323 proteins assigned to 639 folding families; no two homologous sequences share more than 40% sequence identity in this database. Coverage versus errors per query (EPQ) plots (Brenner *et al.*, 1998) used to assess sequence-based methods for homology detection provide the context for our analysis of the effect of reduced alphabets on fold recognition. Figure 2 (inset) shows the coverage versus EPQ plot for the SCOP40 database constructed using the complete 20-letter amino acid code. The coverage is defined as the fraction of homologous sequence pairs that have alignment scores above a threshold, while the EPQ is defined for the same threshold as the total number of non-homologous proteins with alignment scores above the threshold divided by the total number of queries made. As shown in Figure 2 (inset), there is, for example, a 20% coverage of SCOP40 at an EPQ of 0.1 when making use of the full information content of the 20-letter code. This means that we can detect approximately 20% of the true homologs in the SCOP40 database by sequence comparison with a 10% error rate (i.e. the alignment score threshold is set to a value such that 90% of the aligned pairs with scores greater than this threshold are homologous).

The effects of alphabet reduction on protein fold recognition were tested in the following way. The similarity matrices for 10 sets of increasingly reduced alphabets obtained by grouping the amino acids as shown in Figure 1 were assembled. For each of these reduced alphabets, all-against-all sequence alignments were performed using the SCOP40 database. Coverage versus EPQ plots were evaluated for each of the datasets. As the alphabet size is reduced, the information encoded in the amino acid sequences that is responsible for the protein fold recognition is lost. One way to characterize this is to compare the coverages of SCOP40 at a chosen error rate using the different reduced alphabets. The fractional coverage retained relative to the 20-letter alphabet at an EPQ value of 0.001 is shown in Figure 2. There is a strong non-linear dependence of the fold recognition (the coverage) on the number of amino acids in the alphabet from which the similarity matrices (and thus the sequences) were constructed. As the alphabet is reduced from 20 letters to 12 or 10, the percentage coverage retained is reduced by only ~10%; further reduction of the alphabet is accompanied by a steep loss of fold recognition. With a four-letter alphabet, the coverage of the SCOP40 database at an EPQ of 0.001 is reduced by 90% relative to the complete 20-letter code. When the alphabet is reduced to two types of residues, hydrophilic and hydrophobic, there is no detectable fold recognition.

The results shown in Figure 2 correspond to the analysis of fold recognition using the entire SCOP40 database (Murzin *et al*., 1995; Brenner *et al*., 1998). We also have examined the effects of alphabet reduction on fold recognition using subsets of this database corresponding to five major fold categories of the SCOP classification scheme (Murzin *et al*., 1995). We did not detect a strong dependence of the results on fold type; hence this analysis does not support the suggestion that β-sheet containing proteins are less tolerant to an amino acid reduction than α-helical proteins (Riddle *et al*., 1997). However, effects due to the differences in size and distribution of sequences within families among the different folding classes could obscure differences in the dependence of the coverage-EPQ plots on the alphabet.

Discussion

Is there a minimum number of letters required to fold a protein? Combinatorial protein synthesis indicates that some proteins can fold and function with far fewer than 20 amino acids (Sander and Schulz, 1979; Regan and Delgrado, 1988; Heinz *et al*., 1992; Betz *et al*., 1993; Kamtekar *et al*., 1993; Davidson *et al*., 1995; Riddle *et al*., 1997). From a physiological point of view, a viable protein must possess three characteristics: (1) it must fold into a stable and unique three-dimensional structure, (2) the folding process must be realizable on an appropriate time scale and (3) the protein must be able to perform its function. All of these criteria are met by naturally occurring proteins.

Theoretical considerations concerning the folding of heteropolymers indicate that a certain minimum complexity in the polymeric building blocks is required for folding on both kinetic and thermodynamic grounds (Bryngelson *et al*., 1995; Wolynes *et al*., 1995; Hinds and Levitt, 1996; Klimov and Thirumalai, 1996; Shaknovich, 1996; Dill and Chan, 1997; Wolynes, 1997). From the results of simplified lattice model simulations, it has been postulated that a minimum of three different amino acid types is required for protein folding, However, the true minimal alphabet may well require additional complexity for the creation of the large number of protein fold types of the kind actually observed in nature. Although the present studies do not address the physics of the problem in the way that lattice simulations are designed to do, this analysis of simplified amino acid alphabets required for protein fold recognition does have implications for the protein folding problem, particularly with regard to the relationship between reduced alphabets and the diversity of fold types. In the context of protein fold recognition by sequence alignment, we find that sequences constructed from 10-letter alphabets obtained by grouping amino acids appropriately contain nearly as much information as the natural sequences do.

Acknowledgement

This work was supported in part by a grant from the National Institutes of Health (GM-30580).

References

Betz,S.F., Raleigh,D.P. and DeGrado,W.F. (1993) *Curr. Opin. Struct. Biol*., **3**, 601–610.

Brenner,S.E., Chothia,C. and Hubbard,J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Bryngelson,J.D., Onuchic,J.N., Socci,N.D. and Wolynes,P.G. (1995) *Proteins: Struct. Funct. Genet*., **21**, 167–195.

Collins,J.F. and Coulson,A.F.W. (1987) In Bishop,M.J. and Rawlings,C.J. (eds), *Nucleic Acid and Protein Sequence Analysis; a Practical Approach*, Vol. 3. IRL Press, Washington, D.C, pp. 323–358.

Davidson,A.R., Lumb,K.J. and Sauer,R.T. (1995) *Nature Struct. Biol*., **2**, 856–864.

Dill,K.A. and Chan,H.S. (1997) *Nature Struct. Biol*., **4**, 10–19.

Feng,S., Chen,J.K., Yu,H., Simon,J.A. and Schreiber,S.L. (1994) *Science*, **266**, 1241–1247.

Heinz,D.W., Baase,W.A. and Matthews,B.W. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 3751–3755.

Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hinds,D.A. and Levitt,M. (1996) *J. Mol. Biol*., **258**, 201–209.

Kamtekar,S., Schiffer,J.M., Babik,J.M. and Hecht,M.H. (1993) *Science*, **262**, 1680–1685.

Klimov,D.K. and Thirumalai,D. (1996) *Proteins: Struct. Funct. Genet*., **26**, 411–441.

Landes,C. and Risler,J. (1994) *CABIOS*, **10**, 453–454.

Levinthal,C., Wodak,S.J., Kahn,P. and Dadivanian,A.K. (1975) *Proc. Natl Acad. Sci. USA*, **72**, 1330–1334.

Matthews,C.R. (1993) *Annu. Rev. Biochem*., **62**, 139–160.

Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) *J. Mol. Evol*., **12**, 219–236.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.

Myers,E.W. and Miller,W. (1988) *CABIOS*, **4**, 11–17.

Pearson,W.R. (1990) *Methods Enzymol*., **183**, 63–98.

Plaxco,K.W., Riddle,D.S., Grantcharova,V. and Baker,D. (1998) *Curr. Opin. Struct. Biol*., **8**, 80–85.

Regan,L. and Delgrado,W.F. (1988) *Science*, **241**, 976–978.

Riddle,D.S., Santiago,J.V., Bray-Hall,S.T., Doshi,N., Grantcharova,V.P., Yi,Q. and Baker,D. (1997) *Nature Struct. Biol*., **4**, 805–809.

Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) *J. Mol. Biol*., **204**, 1019–1029.

Sander,C. and Schulz,G.E. (1979) *J. Mol. Evol*., **13**, 245–252.

Santibanez,M. and Rohde,K. (1987) *CABIOS*, **3**, 111–114.

Shaknovich,E.I. (1996) *Folding Des*., **1**, R50–R54.

Wolynes,P.G. (1997) *Nature Struct. Biol*., **4**, 871–874.

Wolynes,P.G., Onuchic,J.N. and Thirumalai,D. (1995) *Science*, **267**, 1619–1620.

Received June 1, 1999; revised November 11, 1999; accepted November 29, 1999