Amino Acid Preferences of Small Proteins
Implications for Protein Stability and Evolution

Stephen H. White

Department of Physiology and Biophysics
University of California
Irvine, CA 92717, U.S.A.

(Received 28 April 1992; accepted 19 June 1992)

The dependence of amino acid frequency on sequence length has been examined for the 20 natural amino acids using a set of 2275 protein sequences with little sequence identity. As expected, the frequency of cysteine increases dramatically for sequences shorter than 100 amino acids with a length-dependence that corresponds to an average of two Cys per sequence independent of length. Surprisingly, dramatic changes were also observed for the frequencies of arginine, lysine, aspartic acid, and glutamic acid: Arg and Lys frequencies increase for short sequences whereas Asp and Glu frequencies decrease. These changes do not appear to be due to an over-abundance of DNA- and membrane-binding proteins in the database and may, therefore, be related to protein stability. Possible stabilizing mechanisms include increased hydrogen bonding by Arg and increased hydrophobic stabilization due to the amphiphilic character of Arg and Lys. These observations suggest that amino acid composition played an important role in the evolution of small proteins.

Keywords: protein stability; protein evolution; amino acid frequencies; lengths of protein sequences

Efforts to design proteins de novo are at present focused on short amino acid sequences of 30 to 50 residues (Richardson & Richardson, 1989a; Hill et al., 1990; DeGrado et al., 1991). However, it has been shown on theoretical grounds that monomeric proteins with fewer than about 70 amino acids are unlikely to be stable if their folding is determined solely by hydrophobic forces (Dill, 1985). Disulfide bonds can improve the stability of proteins (Schultz & Schirmer, 1979; Wetzel, 1987; Creighton, 1988) and anecdotal evidence suggests that they are found in most small proteins (Wetlauffer, 1981; Richardson & Richardson, 1989a). If this is a general phenomenon, then the frequency of cysteine in small proteins should be generally elevated relative to large proteins. But, this raises a more general question: are there amino acids besides Cys that might be preferred by small proteins for reasons of enhanced stability? To examine this question, the length-dependence of the average frequency of each of the 20 natural amino acids was determined for a carefully selected set of 2275 protein sequences with little sequence identity. The results show that the frequency of Cys in short sequences is greatly elevated relative to the frequency in long sequences. Furthermore, for short sequences, the observed length-dependence of the average frequency corresponds to that expected for two Cys per sequence. The frequencies of other amino acids also depend significantly on length, but the behavior of Arg, Lys, Asp and Glu is most striking: Arg and Lys are significantly elevated relative to Asp and Glu, with Arg being elevated relative to Lys for the shortest sequences. These unanticipated changes have implications for the thermal stability and evolution of small proteins.

The set of protein sequences examined (the Superfamily Set) was drawn from release 27 (31 December, 1990) of the National Biomedical Research Foundation's Protein Identification Resource (PIR) database (George et al., 1986; Oreutt et al., 1983). Only the first qualifying member encountered in each superfamilly was included to avoid redundancy. Qualifying sequences were those that were complete (i.e. not fragments), unambiguous and longer than 21 amino acids. The sequences were sorted by length into bins 20 amino acids wide and the ensemble amino acid frequencies for the \( N_j \) sequences (nominal length \( L_j \) in the jth bin were determined (see legend to Fig. 2). A plot of \( N_j \) versus \( L_j \) (Fig. 1) reveals a distribution that declines exponentially after peaking at about 115 amino acids. The mean length of the population is about 300 and the great majority of the sequences
have lengths of 600 amino acids or fewer. None of the amino acid frequencies shows a statistically significant dependence on length for $L_j > 300$. The mean amino acid frequencies for sequences with $300 \leq L_j \leq 600$ were used as the "long-sequence" reference standards for comparison with the amino acid frequencies of short sequences ($100 \leq L_j \leq 200$ and $20 \leq L_j < 100$). Two short-sequence regimes were chosen because the rate of change of frequency with length varies among the amino acids. The rate for a particular amino acid can be estimated by comparing the two short-sequence frequencies.

The amino acid frequencies for the three length ranges are plotted in Figure 2(a) and the ratios of the frequencies of the short sequences to the long ones are plotted Figure 2(b). As $L_j$ decreases, statistically significant increases in frequency are observed for Cys, Lys and Arg, and decreases in frequency for Glu, Gln, Pro, Asn, His, Gly, Ala and Asp. The frequency of Met also increases but this is because its absolute frequency is low and virtually all of the sequences in the database begin with Met. The plot of frequency versus length for His, shown in Figure 3(a), is typical of the behavior of Gln, Pro, Asn, Gly and Ala, which is generally characterized by a slow steady decline in frequency with decreasing $L_j$. Interestingly, Figure 2(a) suggests that among amino acids with frequencies greater than 5%, the increases in Lys and Arg just about balance the decreases in Ala, Gly and Asp. The most dramatic and interesting frequency changes with length are those for Cys, Arg, Lys, Glu and Asp shown in Figures 3 and 4.
The expectation of a higher frequency of Cys in small proteins is confirmed by the results presented in Figure 3(b). In contrast, the frequency of His (Fig. 3(a)), which has about the same long-sequence frequency as Cys, shows a much less impressive dependence on length. If the general role of Cys is to stabilize small proteins by disulfide bonds, then the increased frequency should correspond to at least two Cys residues per sequence. This condition is fulfilled as shown by the broken curves in Figure 3(b), which represent the frequency distributions expected if every sequence had precisely two or four Cys residues regardless of length. It thus appears that small proteins frequently require covalent or ligand-mediated stabilization. This implies a decreased hydrophobic stability of many small proteins and indicates the importance of amino acid composition and sequence length in protein stability (Dill, 1985).

Figure 2(a) shows that the long-sequence frequency of Glu is larger than that of Arg, Lys or Asp, which are approximately equal to each other. However, as $L_j$ decreases, the frequencies of Arg and Lys increase, while those for Glu and Asp decrease, with the decrease in Asp being greater than that for Glu (see arrows, Fig. 2(a)). These changes produce a disparity between the frequencies of positively and negatively charged residues in the population of small proteins, which is perhaps not surprising in that many of them may be involved in interactions with DNA or membranes. However, the difference in behavior of the length-dependence of the frequencies of Arg and Lys shown in Figure 4(a) and (b) suggests that charge requirements alone are not the sole determinant of frequency. After Cys, the most dramatic frequency change with length occurs for Arg (Fig. 4(a)).

These observations suggest an important role for Arg in the stabilization of small proteins. But, several questions must be addressed: (1) is the population of short protein sequences biased by an overabundance of proteins that interact with DNA and membranes? (2) Do individual proteins with high frequencies of Arg and Lys also have low frequencies of Asp and Glu? (3) Do proteins with fewer than two Cys residues have higher frequencies of Arg? While the large size of the Superfamily Set makes it impractical to assign individual proteins to specific classes, the question of population bias can be approached by searching the description lines of the proteins in the database for relevant keywords. A search on the keywords DNA-binding, regulatory, histone, protamine, zinc finger and repressor yielded 75 proteins from the Superfamily Set. Of these, only 18 were members of the population of 390 short-sequence proteins ($L \leq 100$ amino acids). For membrane-associating proteins, a search on the keywords toxin and hormone yielded 28 sequences of which 11 were in the short-sequence population. Combined, the proteins that may have specific interactions with DNA and membranes account for about 7 to 8% of the short-sequence population and 4 to 5% of the total population. One can reasonably conclude that the frequency variations are not
seriously biased by DNA- and membrane-binding proteins.

The frequencies reported earlier are ensemble frequencies calculated from the total numbers of amino acids in the populations. The second and third questions, however, require knowledge of the frequencies for the individual sequences based upon Cys usage. The short-sequence population was therefore divided into low-Cys (<2 per sequence) and high-Cys (≥2 per sequence) populations and the frequencies of the individual sequences were determined. Interestingly, 229 of the 390 proteins fall into the low-Cys class with mean $f_{Cys} = 0.5\%$ compared to 6.4% for the high-Cys class, which tend to have three or more Cys/sequence. This suggests two broad stability classes for small proteins distinguished by Cys usage. The average frequencies of Arg (7.1%) and Lys (7.2%) in the low-Cys group are elevated relative to the high-Cys group (6.1% and 5.7%, respectively) even though the fraction $F_{arg} = (f_{arg} + f_{lys})/(f_{arg} + f_{lys} + f_{glu} + f_{asp})$ for each group is about the same (0.57 to 0.58). The behavior of the individual sequences can be judged from the fraction of sequences with $F_{arg} > 0.5$. The observed values are 0.64 for low-Cys and 0.63 for high-Cys, with significance levels of 0.003 and 0.02, respectively. That is, the probability that these ratios occur by chance is 0.02 or less.

Similar trends are observed for the identified DNA- and membrane-associating proteins but there are some interesting differences between the two sub-populations. The DNA-associating proteins have Arg and Lys frequencies of 15% and 64%, respectively, whereas the membrane-associating proteins have frequencies of 42% and 90%. A possible explanation for this reversal is that hydrogen-bond interactions may be less important for membrane-associating proteins than for DNA-associating proteins. Another interesting difference is that the Cys frequency of the membrane-associating proteins is very high (11%) with an average of about seven per sequence (none with fewer than 2). This makes sense if one supposes that these small proteins must have a stable tertiary structure in solution to expedite highly stereo-specific interactions with receptors. While such interactions are also important for DNA-associating proteins, one can imagine that their structure in solution prior to binding is of lesser importance.

Overall, the results of this closer examination of the short-sequence proteins support the idea that Arg and the proportion of Arg + Lys relative to Glu + Asp are of general (but not, of course, exclusive) importance in protein stability. Although one must be cautious about inferring specific mechanisms from global statistical analyses such as this one, some interesting possibilities are suggested by these and other data. By comparing the natural amino acid substitutions in a number of thermophilic and mesophilic proteins, Argos and his colleagues obtained statistical evidence that Lys-to-Arg substitutions improve thermal stability (Argos et al., 1979; Menendez-Arias & Argos, 1989). A physical basis for such improvements was provided by Mrabet et al. (1992), who showed that the increased thermal stability of D-xylose isomerase occurs because Arg forms a larger number of hydrogen bonds than Lys. Increased hydrogen bonding may therefore be one of the major reasons that Arg is favored in small proteins. But, the data presented here suggest additional stabilizing mechanisms. The general increases in the frequencies of Arg and Lys may contribute to protein stability because they have significant hydrophobic as well as polar character that would permit them to fulfill simultaneously polar exterior and non-polar interior functions (Jacobs & White, 1980).

This might be the explanation for changes in the relative frequencies of Arg, Lys, Glu and Asp. Other than charge, the main differences between the positive and negative residues are smaller hydrophobic components, which may make Asp and Glu less able to fulfill both exterior and interior functions. Consistent with this general idea, the average frequency of Glu in the short-sequence protein set is 5.4% compared to 0.4% for Asp. Furthermore, Rose et al. (1985) found that the average accessible surface area buried upon folding increased in the order Asp < Glu < Lys < Arg. Finally, as noted earlier, the very abundant Gly and Ala residues are disfavored in small proteins, consistent with a greater role for charged amino acids in hydrophobic stabilization. However, the relative frequencies of Glu and Asn and their changes with length do not agree with this interpretation. But, this may be because their absolute frequencies are considerably smaller than those for Glu and Asp, causing them to have a smaller overall influence.

The analysis presented in this paper provides information that must be accounted for by theories of protein evolution. It is generally assumed, for example, that proteins originated by the assembly of short exon-encoded sequences (Gilbert, 1978; Blake, 1979, 1983). Dori et al. (1990) have assembled an exon database which shows that the average length of modern exons is about 40 to 50 amino acids. If the primordial exons were of this size and if they represented independently stable folding units, then many of them would have been expected to be enriched in Cys, Arg and Lys. An increased frequency of Arg with its enhanced hydrogen bonding potential could have been particularly advantageous for the evolution of proteins in a hot reducing primordial environment. This might explain why Arg is represented by six codons in the genetic code despite the fact that its frequency (≈6%) is presently unremarkable. Based upon codon frequency (King & Jukes, 1969), the frequency of Arg would be expected to be about 9% as observed for Leu, which is also represented by six codons. Provocatively, the frequency of Arg approaches this value in the shortest sequences.

The conclusions presented here are strengthened by comparisons of the amino acid frequencies for the Superfamily Set with those for proteins from
Escherichia coli (data not shown). An examination of 835 qualifying E. coli sequences from the 1990 PIR database shows virtually no dependence of Cys frequency on length. Further, the global average frequency of Cys in E. coli (1.2%) is smaller than for the Superfamily Set proteins (1.8%). Cysteine apparently conferred little evolutionary advantage in the reducing intracellular environment of E. coli (Wetzel, 1987). In contrast, the length variations of the frequencies of Arg, Lys, Asp and Glu in E. coli are similar to those of the Superfamily Set except that Lys increases with decreasing length in a manner similar to that of Arg. These differences indicate that small proteins can be stabilized without disulfide bonds, although it should be noted that the peak of the length distribution for E. coli sequences is about 150 amino acids rather than 115. This shift toward longer sequences suggests the importance of sequence length in protein stability.

I am grateful to Drs George Rose, Ken Dill, David Eisenberg, Tom Poulos, Bill Winley and Jay Edelman for their comments on the manuscript. This work was supported in part by the NSF and the NIH.

References


Edited by B. W. Matthews