Sequence Codes for Extended Conformation: A Neighbor-Dependent Sequence Analysis of Loops in Proteins

Jin-an Feng feng@guanyin.fccc.edu
Chiquito J. Crasto crasto@guanyin.fccc.edu

Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, PA 19111, USA

Keywords: loops, protein structures, sequence analysis, PDB, loop propensity

1 Introduction

Far less is known about the sequence patterns of loops than that of the regular secondary structures in proteins. In large part this is because loop regions in proteins is the most sequentially diversified region. Due to lack of knowledge, our ability to identify loop regions of the proteins is limited. The loop regions of polypeptide chains are flexible and usually adopt extended conformations without any regular geometrical characteristics (Richardson, 1981). Loop structures are essential in allowing polypeptide chain to change directions and facilitating globular fold. No longer recognized as simple pieces of strings that hold secondary structures together, loops are shown to play an important role in stabilizing protein folds and regulating functions of the proteins (Dill *et al.*, 1993; Fetrow, 1995). Loops present on the surface of the proteins are often involved in mediating specific protein-protein interactions, receptor binding, and protein-DNA interactions. Residues on the surface loops are often targets of protein modifications, such as phosphorylation, that lead to subsequent functional activation or deactivation.

Here we present the work of an extensive analysis of the loop sequences in proteins. A loop databank is derived from the PDB. The databank is subdivided into different groups based on loop sizes, solvent accessibilities and connecting secondary structural elements. We calculated amino acid loop propensities in different loop groups in order to explore potential residue preferences in these loop groups. We ask the question whether there is an inherent amino acid sequence preference for adopting loop conformations. To address this issue, we carried out a neighbor-dependent sequence analysis of loops in proteins, and systematically analyzed the effect of neighboring amino acid type on the loop propensities of residues in loops.

2 Methods

All analyses were performed using a loop database derived from the October 1999 release of the Brookhaven Protein Data Bank which includes 10280 entries. Loops in the database are further subdivided into different groups according to loop lengths, solvent accessibility, and connecting secondary structures. The neighbor-dependent analysis of loops in proteins is carried out as follows. The frequency of occurrence of the residue type x at neighboring positions of a loop residue (a) is calculated according to equation (1):

$$\varepsilon_{x(a\pm i)} = \varepsilon_a\left(\frac{x_{(a\pm i)L}/n_{(a\pm i)L}}{x_{(a\pm i)P}/n_{(a\pm i)P}}\right) \tag{1}$$

where $x_{(a\pm i)L}$ is the number of residue type x at the $\pm i$th positions of the residue a in loops; $n_{(a\pm i)L}$ is the total number of residue at the $\pm i$th positions of the residue a in loops; $x_{(a\pm i)P}$ is the number of residue type x at the $\pm i$th positions of the residue a in our protein databank; $n_{(a\pm i)P}$ is the total number of residue at the $\pm i$th positions of the residue a in our protein databank, and ε_a, the loop propensity of residue type a, is applied here as a weighting factor that removes the bias of uneven distribution of amino acids between loops and proteins. In this report, we present our analysis of the values of $\varepsilon_{x(a\pm 1)}$.

3 Results and Discussion

Our neighbor-dependent analysis of loop residues shows that the neighboring residues could have profound effect on the preference of certain amino acids adopting loop conformations. The most "influential" amino acids (the affecters) in loops are Pro and Gly. The loop propensities of all amino acids are significantly higher when they are neighbored by a Pro. Residues with low loops propensities (LLP) such as Cys, Ile, Leu, Trp, and Val ($\varepsilon_{\text{average}} = 0.68$) have an averaged loop propensity of 1.35 when they are neighbored with a Pro. When these residues are positioned next to Gly in sequence, their loop propensity is raised to an average value of 1.22. The neighbor-dependent sequence analysis also revealed positional preference of certain amino acids in loops. For example, the propensities of Pro positioned at +1 position of Val ($\varepsilon_{p(v+1)} = 1.10$) is significantly different from that of the Pro positioned at −1 position of Val ($\varepsilon_{p(v-1)} = 1.66$), suggesting the dyad Pro-Val prefers loop conformation while dyad Val-Pro has no preference for loop conformation. This type of sequence preference patterns is also found for other amino acid types in loops.

One of the most noticeable attributes of the neighbor-dependent analysis method is the enhancement of statistical significance of residue propensity. When the loop propensity ε_a is calculated based on the frequency of occurrence of each individual amino acid in loops, its value is in the range between 0.6 and 1.8. On the other hand, the propensity values, $\varepsilon_{x(a\pm 1)}$, derived from this neighbor-dependent analysis, have a range of 0.3-2.9. In light of this expanded statistical scale, we are able to explore the "hidden code" in loop sequences. Specifically, we can identify dyad signatures (a-b) that are highly favorable for loop conformation, while their dyad pairs (i.e., b-a) have little or no preference for loop conformation. Table 1 lists some of the asymmetric dyads that have high propensity for loop conformation.

Table 1

Total loops	Short loops	Medium Loops	Long Loops	HH loops	HS loops	SH loops	SS loops
Pro-Val	Asn-Gln	Asn-Glu	Ala-His	Arg-His	Arg-Cys	Asn-Trp	Cys-Glu
	Asp-Ala	His-Asp	Arg-Asn	Cys-Gly	Asp-Asn	Asp-Met	Gln-His
	Ile-Gly	Lys-Ser	Asn-Gln	Cys-Lys	Cys-Gln	Glu-Asn	Gly-Trp
	Ile-Pro	Pro-Met	Asp-Ser	Gly-Ile	Cys-Gly	His-Met	His-Glu
	Pro-Val	Pro-Val	Asp-Thr	His-Met	His-Asn	His-Phe	Pro-Cys
	Trp-Pro	Trp-Asp	Asp-Trp	Leu-Asn	His-Cys	Met-Cys	Pro-Val
			Gln-His	Lys-Asp	Ile-Gly	Phe-Trp	
			Gly-Asn	Lys-Ser	Lys-Ser	Pro-Met	
			His-Cys	Phe-Trp	Lys-Tyr	Pro-Val	
			His-Leu	Pro-Met	Pro-Met	Ser-Cys	
			His-Met	Pro-Val	Pro-Trp	Thr-Arg	
			Leu-Asp	Trp-Gly	Val-Gly	Trp-Gly	
			Lys-Asn	Val-Gly		Tyr-Pro	
			Phe-His				
			Thr-Trp				
			Trp-Gly				
			Tyr-Arg				

The power of the neighbor-dependent sequence analysis of protein sequence is that it allows us to explore the positional preference of amino acid in protein sequences. The dyad signatures established in this study can serve as basic parameters for further investigation of sequence/structure relationship of proteins. A natural extension of this study would be to explore the neighboring effect beyond $+1$ and -1 position of the loop residues. Work is in progress to identify potential triads, or tetrads of polypeptide sequences that have preference for loop conformation in proteins.

References

[1] Dill, K.A., Fiebig, K.M. and Chan, H.S., Cooperativity in protein-folding kinetics, *Proc Natl Acad Sci U S A.*, 90(5):1942–1946, 1993.

[2] Fetrow, J. S., Omega loops: nonregular secondary structures significant in protein function and stability, *FASEB J.* 9:708–717, 1995.

[3] Richardson, J.S., The anatomy and taxonomy of protein structure, *Adv. Protein Chem.*, 34:167–339, 1981.