References

Go,M. (1981) *Nature*, **291**, 90.
Go,M. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 917.
Saitô,N. (1989) *Adv. Biophys.*, **25**, in press.
Saitô,N., Shigaki,T. and Yamamoto,M. (1988) *Prot. Struct. Funct. Genet.*, **3**, 199.
Watanabe,K., Nakamura,A., Fukuda,Y. and Saitô,N. (1989) Submitted.
Yoshimura,T., Noguchi,H., Inoue,T. and Saitô,N. (1989) Submitted.

Systematic search of structural elements in proteins

S.Saitoh, K.Nishikawa and T.Yao
Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

Secondary structures of proteins are classified into helices, sheets and various types of turns. However, rather large regions are left, named loops. In order to classify protein structures systematically, especially for these loops, a method named the conformational probe method was developed. Preliminary results of this work were presented previously (Saitoh *et al.*, 1987). α helices and β sheets which were assigned by Kabsch and Sanders' (1983) algorithm were first extracted. Then all fragments (main chain atoms) in the remaining region (i.e. turns and loops) were superimposed mutually and if the root mean square deviation value (RMSD) was less than some cut-off value, the fragments were classified into the same group. This method was applied to 59 proteins of known X-ray structure (better than 2 Å resolution), and the length of four residues were considered. When the cut-off value was set to 0.6 Å, 342 groups were needed for the grouping of all fragments. If doubly grouped fragments were excluded, the number of groups which contained more than 30 members was 30. In each group the Ramachandran plots of second and third residues of its fragments were almost in the same region respectively. The 30 groups were analyzed individually, among them a group named FXN8, the fragments of which have a similar shape (RMSD $<$ 0.6 Å) with the fragment from residues 8 to 11 in Flavodoxin (4FXN in PDB code), were especially interesting. The fragments have almost always glycines at the third position, in accord with the fact that the phi psi value is in the left-handed helix. Some fragments of the group were situated at the junction between the carboxyl end of the α helix and the beginning of the β sheet, and some others were part of β hairpins. The fragments collected with this method have rather common features, so this method would be useful for the classification of loop regions. FXN8 may be the newly found structural element which often appears in loop regions of various proteins.

References

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Saitoh,S., Sato,S., Yao,T. and Go,N. (1987) Abstracts of the 27th National Meeting of the Biophysical Society of Japan, Tokushima, October 1987, S510 (in Japanese).

Estimation of the average energy increment by amino acid exchange in proteins and its use in evaluating a homology scoring matrix

S.Miyazawa and R.J.Jernigan[1]
National Institute of Genetics, Mishima, Shizuoka 411, Japan and [1]National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

The average energy increments of protein native structures caused by amino acid exchange are estimated, and used to evaluate a transition probability matrix of codon substitutions and then a log relatedness odds matrix which is used as a scoring matrix to measure the similarity between protein sequences. The average fitness of an amino acid exchange is approximated by the average degree of instability of the protein native structure which is then approximated to be equal to the Boltzman factor of the average energy increment of the protein native structure caused by the amino acid exchange. In a previous study (Miyazawa and Jernigan, 1985) we estimated the effective inter-residue contact energy of each type of amino acid pair for proteins in solution from 18 192 residue–residue contacts observed in 42 globular proteins. By using the contact energy and the number of contacts for each type of amino acid pair, we evaluate the average energy increment of protein native structures caused by an amino acid exchange. The estimates of average energy increments of exchanging each of the 20 kinds of amino acids for any other show reasonable characteristics of physico-chemical similarities of amino acids. A transition probability matrix for codon substitutions is evaluated from those estimates of average energy increments of amino acid exchanges with a simple assumption for the rule of base mutation. Codons are assumed to mutate to another codon with the probability that is proportional to the equilibrium frequency of the codon, only if it can occur with a single base mutation. The equilibrium frequencies of amino acids are taken to be equal to those compiled by Dayhoff *et al.* (1978) in order to compare the transition matrix with the mutation probability matrix evaluated by them. Degenerated codons are assumed to be equally used, even though it may not be true. This transition matrix reflects base mutation rates, and includes the effects of the genetic code and the conservative selection for amino acid exchanges. The diagonal elements of the transition matrix of 1 PAM except for Ser correlate well with those of the mutation probability matrix of 1 PAM evaluated by Dayhoff *et al.* (1978). The log relatedness odds matrix, which is calculated from the transition matrix of 250 PAM and can be used as a scoring matrix to detect distant relationships between protein sequences, is compared with the Dayhoff's matrix and other scoring matrices. The diagonal and off-diagonal elements of the log odds matrix correlate well with those of the Dayhoff's matrix, when the most infrequent and the least mutable amino acids, Trp, Met, Cys and Tyr, are excluded; the correlation coefficient is 0.82. The poor correlation of off-diagonal elements including those amino acids may come from statistical errors due to small numbers of substitutions. Alignment scores obtained from this scoring matrix show that this matrix tends to yield significantly higher scores than the unitary matrix and genetic code matrix, and also may yield higher scores in the region of low alignment score than the Dayhoff's scoring matrix. Thus, this scoring matrix will be useful for homology search as well as the Dayhoff's one.

References

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) *Atlas of Protein Sequence and Structure*. Vol. 5, Supplement 3, pp. 345–352.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.