

Interlock structural motif is typical for beta sandwich domains

Aksianov E.¹, Alexeevsky A.², Kister A.³, Gelfand I.⁴

Keywords: structural motif, interlock, beta-sandwich

The goal of this work is to define common structural and sequence features in a large group of sandwich-like proteins. This type of architecture unites 96 different protein superfamilies, which have no detectable sequence homology. We hypothesize here that proteins grouped together on the basis of common architecture have some commonality on the level of supersecondary and primary structure as well. Recently, our preliminary analysis has revealed a certain arrangement of four strands – interlock [1] (fig. 1) which can serve as the invariant characteristics of the sandwich proteins and distinguish this group of proteins. The delineation of an invariant supersecondary substructure made it possible to identify and align strands from very diverse proteins with a dissimilar arrangement of strands and variable numbers of the strands. The result of the multialignment of the sandwich proteins gives 8 hydrophobic positions with fixed structural roles. These positions lie at the center of the interface between the beta-sheets and form the common geometrical core of sandwich proteins. Earlier it was shown for certain protein groups that just interlocked strands play an important role in human fibronectin folding [2]; also interlocked strands are the most spatially conserved strands among immunoglobulines [3].

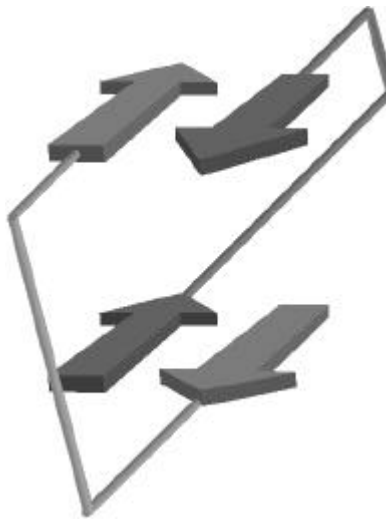


Fig. 1. The scheme of the interlock.

To check our hypothesis about the invariant core, in the current work we perform the systematic screening of the interlocks in all known protein structures. To this purpose we developed *FINE* (*F*inder of *I*nterlocked *E*lements) computer program. In a given protein structure, *FINE* either detects an interlock (or several interlocks) or reports an interlock absent. The algorithm includes three steps: (1) Detect all sets of four beta strands with an appropriate interlock topology; (2) Select two residues in each of four strands in such way that the mutual positions of eight C_α atoms satisfy the certain geometrical characteristics; (3) Calculate the score of similarity of spatial arrangement of eight C_α atoms by compare them with the “canonical” interlock in the training set of 74 Ig-like domain structures.

¹ Virology department, Biological faculty, Moscow State University, Moscow 199992, Russia, E-mail: evaksianov@belozersky.msu.ru

² Belozersky Institute, Moscow State University, Moscow 199992, Russia, E-mail: aba@belozersky.msu.ru

³ Department of Health Informatics, SHRP, University of Medicine and Dentistry of New Jersey, Newark, NJ, 07107, USA. E-mail: akister@math.rutgers.edu

⁴ Department of Mathematics, Rutgers University, Piscataway, NJ, 08855, USA. E-mail: igelfand@math.rutgers.edu

To create scoring system, the distances between all pairs of C_a atoms from 8 key amino acid residues were computed for each structure. The score for an analyzed set of 8 C_a atoms is defined as $S = -S \log p_{ij}$, where p_{ij} is the frequency in the training set of the observed distance between i -th and j -th C_a atoms. The less is score S , the more similar to C_a of interlock key residues is the analyzed set of 8 C_a atoms. The threshold value 20 were used for interlock detection. Parameters of *FINE* program were adjusted to avoid false hits of interlocks.

	Analysed	Interlocks found
PDB files	27954	2105
SCOP domains	54745	5431
SCOP families	2460	153
SCOP superfamilies	1408	80
SCOP folds	905	35

Table 1. Distribution of interlocks in protein structures.

All PDB databank [4] was screened with a *FINE* program. The results were compared with a SCOP classification. Interlocks were found in 5431 SCOP domains (approximately 10% of all 54745 domains in 1.65 SCOP release – see table 1). In SCOP 50 folds are annotated as ‘sandwich-like’. Interlocks were detected in 23 of them. Among other 12 folds, containing at least one domain with interlock detected by *FINE* program, 5 folds could be also considered also as ‘sandwich-like’. In the rest 7 folds interlocks were detected in only 28 domains. All these domains were analyzed and in a number of cases detected interlocks were considered as false-positive mistakes of the algorithm.

It should be noted that inside many SCOP folds interlocks were detected in some domains and were not detected in other domains. Partially, it is caused by the peculiarities of the algorithm of *FINE* program. On the other hand, we found a few SCOP families, in which some domains contain an interlock and the other do not contain.

The obtained results show that interlocks are present almost exclusively in sandwich-like domains, but not in all of them.

This work is partially supported by Ludwig Institute for Cancer Research (CRDF GAP grant RB0-12771) and RFBR grant 030790157.

References

- [1] Kister, A.E., Finkelstein A.V. and Gelfand I.M. 2002. Common features in structures and sequences of sandwich-like proteins. *Proceedings of the National Academy of Sciences USA*. 99(22): 14137-14141.
- [2] Cota E., Steward A., Fowler, SB. and Clarke J. 2001. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. *Journal of Molecular Biology* 305(5):1185-1194.
- [3] Bork P., Holm L. and Sander C. 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. *Journal of Molecular Biology*. 242(4):309-320.
- [4] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne P.E. 2000. The Protein Data Bank. *Nucleic Acids Research*. 28, pp. 235-242.