STRUCTURAL CONSERVATION OF REMOTE HOMOLOGUES: BETTER AND FURTHER IN CONTACT FRAGMENTS

Clovis Galiez[1]*, François Coste[1]

[1] *Dyliss team, IRISA/INRIA Rennes - Bretagne Atlantique, France.* *clovis.galiez@inria.fr

We address here a basic question on sequence-structure relationships in proteins: does a protein sequence depict a structure with a uniform faithfulness all along the sequence ? We investigate this question by defining contact fragments *and show that their sequence homologs are significantly more faithful to structure than randomly chosen fragments.*

INTRODUCTION

While it is a key point to predict the structures and functions of the ever-increasing amount of available sequences, sequence-structure relationships in proteins is not well-understood [1]. We introduce here the *contact fragments*, whose definition conciliates both sequential and structural neighborhood, and show experimentally that their sequence to structure relation is better conserved than for random fragments and remains significant at lower sequence homology levels, making them appealing for the characterization of distant sequences sharing a common structure.

METHODS

We define a *contact fragment (CF)* to be a pair of backbone segments from a protein which are close in the tertiary structure, or more precisely which 1) share a contact (C_α atoms are closer than σ=7Å), 2) interact (every C_α atom is at most at a distance τ=13Å of a C_α of the other segment) and 3) do not overlap each other (see Figure 1.).

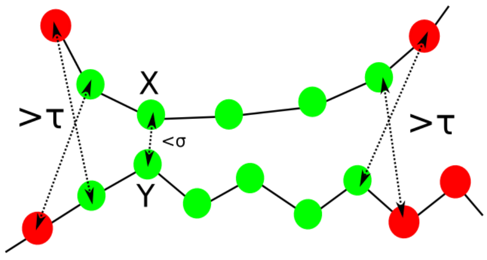

FIGURE 1. Definition of a contact fragment (in green) with respect to σ *and* τ *thresholds.*

To quantify structure conservation with respect to sequence homology in proteins, we make the following experiment: we mine the non-redundant PDB (90% of sequence identity) with protein sequence fragments. Then, for each sequence query, we compare (using TM-align [2]) the structures of the retrieved sequences with the structure of the original query. There are 3 types of sequences we use as queries: 1) the underlying sequence of CF (773 CF extracted from the dataset Astral64 described in [3], the length per segment of CF ranging from 2 to 70, with a mean of 15 residues), 2) the underlying sequence of a randomly chosen pair of fragments (*RP*, extracted from the *same* dataset with *same lengths than* CF) and 3) the underlying sequence of randomly chosen contiguous fragments (*RC*, extracted from the same dataset and whose length equals the sum of length of the two segments of CF).

RESULTS & CONCLUSIONS

The results are shown on Figure 2. It should be read like this : for a given *p*-value threshold (on the *x*-axis), then 80% of the retrieved sequences have a better TM-score with the original query than the corresponding value indicated by the thick line. Thin lines represent the number of hits below the given *p*-value threshold of the *x*-axis. As expected, for the three types of fragments, the lower is the *p*-value threshold for the sequence retrieval, the more the retrieved sequences are similar, and thus, the higher lie the TM-score values. Interestingly, we see that contact fragments show a far better structure conservation for a given sequence similarity. For instance, if one wants to know what is the p-value threshold such that 80% of the hits below this threshold have a strong structure conservation (i.e. let us say that has a TM-score above 0.8 with the original query : ① on Figure 1), then, from our dataset one has to fix this threshold around 7.10-1 in the case of contact fragments (allowing for a bigger recall) whereas one would fix it around 1.10-8 in case of random pair of sequences, and to 3.10-5 in case of contiguous fragments. These threshold would retrieve 3524 sequences for CF ②, 1653 for RP ③ and 388 for RC ④. From this experiment, we see that for the same sequence similarity, the structure is more conserved in contact fragments.

A corollary is that when mining protein sequences, one may allow more sequence divergence in contact fragments than elsewhere in the sequence.

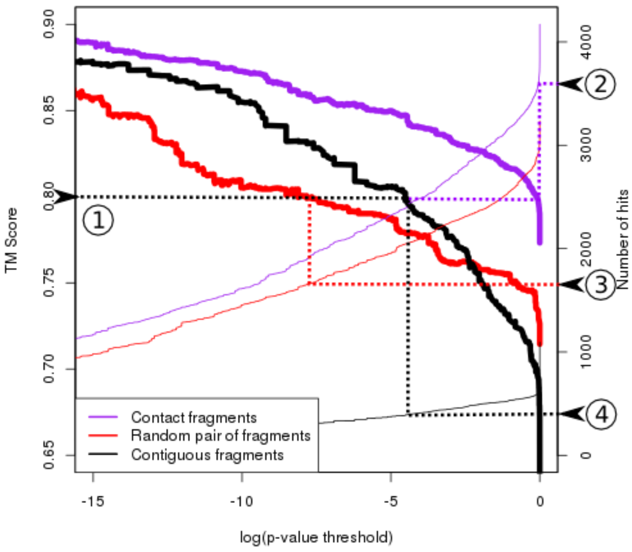

FIGURE 2. Thick lines are the 20% quantile of the TM-score values for hits below the given p-value threshold of x-axis. Thin lines show the corresponding number of hits.

This study suggests first that contact fragments carry a strong sequence-structure relationship, allowing them to be used as accurate building blocks for structure prediction. On the primary structure counterpart, weighting sequence similarity according to the position in the protein structure should improve the current tools for protein sequence retrieval.

REFERENCES

1. H. H. Gan et al., *Analysis of Protein Sequence/Structure Similarity Relationships*, Biophysical Journal, 83 (2002)
2. Y. Zhang, J. Skolnick, *TM-align: A protein structure alignment algorithm based on TM-score* , Nucleic Acids Research, 33: 2302-2309 (2005)
3. C. Galiez, F. Coste, *Amplitude Spectrum Distance: measuring the global shape divergence of protein fragments*, Submitted (2015)