# Environment specific substitution tables for thermophilic proteins

Mizugushi K (1,2), Sele M (3), Cubellis MV (3)

(1) Department of Biochemistry, University of Cambridge, UK
(2) Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
(3) Dipartimento di biologia strutturale e funzionale, Universita' di Napoli "Federico II", IT

## Motivation

Most organisms grow at temperatures from 20 to 50 °C, but some prokaryotes, including Archaea and Bacteria, are capable of withstanding higher temperatures, from 60 to >100 °C. Subtle differences between thermophilic and mesophilic molecules can be found when sequences or structures from homologous proteins are compared, but often they are family-specific and it is very difficult to derive general rules. The availability of complete genome sequences makes it feasible a large scale comparison between thermophilic and mesophilic proteins. Although most sequenced genomes of thermophilic organisms belong to archaea, a few are also available for eubacteria. We made independent comparisons of mesophilic proteins with their thermophilic counterparts of archaeal or eubacterial origins, since different mechanisms for the adaptation of proteins at high temperatures might have been exploited in the two kingdoms. Moreover we derived amino acid substitution tables that give the likely substitutions of amino acids in particular local environments because the conservation of amino acid residues has been shown to be strongly dependent on the environment in which they occur in the folded protein.

## Methods

A database of 19168 protein sequences derived from the genomes of 10 archaea living at or above 60 °C was compiled. First, 3763 protein structures belonging to 1057 different families were taken from HOMSTRAD, a database of protein structure alignments for homologous families (1). The sequence corresponding to each structure was used as a query to search with BLAST the database of archaeal thermophilic proteins. In this way, we built alignments, where the first sequence is for a protein of known structure and the other ones are for its homologues from archaeal thermophiles. The residues of the first protein, whose structure is known, were assigned to eight different structural environments, i.e. alpha helix (buried or exposed), beta strand (buried or exposed), positive mainchain phi angle (buried or exposed) and coil (buried or exposed). Environmentspecific amino acid substitution tables were calculated using the modified version of SUBST (K. Mizuguchi, unpublished). Substitution frequencies represent the likelihood of acceptance of a mutational event by a residue in the first sequence and in a particular structural environment, leading to any other residue in the archaeal thermophilic sequences. These tables, specific for thermophilic archaeal sequences, were compared to the standard substitution tables used in FUGUE (2). Environment specific substitution tables for thermophilic eubacteria were derived with the same method. In order to determine the effect of different substitution tables on sequence alignments, we used pairs of structures (one thermophilic and the other mesophilic) from HOMSTRAD. MELODY in the FUGUE suite of programs (2) was used to derive two profiles from the mesophilic protein structure; the first was obtained by using standard substitution tables and the second exploiting thermophile specific substitution tables. Each profile was aligned with FUGUE (2) to the sequence of the thermophilic protein, resulting in two alignments. The original structural alignment between the mesophilic and thermophilic proteins stored in HOMSTRAD, was used as a reference.

## Results

A few general rules for the adaptation of proteins at high temperatures have been put forward so far. Our substitution tables derived from an extremely high number of raw substitution counts allowed us to confirm or disprove some of them. For instance it has been claimed that an increase in thermostability is correlated with the location of branch points in amino acids and beta and gamma branched amino acids increase protein thermostability. Our substitution tables derived for thermophilic archaea confirm that in most environments Ile is a preferred amino acid. We aligned the sequences of thermophilic proteins to those of mesophilic proteins of known structure. These are the kind of alignments that must be used to model a thermophilic protein based on a mesophilic template. Unfortunately only a few pairs of homologous protein structures were available, where one member is mesophilic and the other thermophilic. On this limited set, we compared the alignments that exploited thermophile specific substitution tables and those with the standard substitution tables. The two kinds were comparably accurate, although for some families, thermophile specific substitution tables produced more accurate alignments.

**Contact email:** cubellis@unina.it

**References**

1. Mizuguchi K, Deane CM, Blundell TL, Overington JP. Protein Sci 1998;7:2469-2471
2. Shi J, Blundell TL, Mizuguchi K. J Mol Biol 2001;310:243-257