Characterization of a family B DNA polymerase and sequence analysis of replicative proteins from the hyperthermophilic marine archaeon
Thermococcus sp. OGL-20P

Damien Marsic, Owen K. Garriott, Jim Hudson and Joseph D. Ng

Laboratory for Structural Biology and Department of Biological Sciences, University of Alabama in Huntsville, MSB 203C, Huntsville, AL 35899, USA, marsicd@uah.edu

Thermococcus sp. OGL-20P is a sulfur-reducing hyperthermophilic archaeon that has been isolated from mud collected at the Rainbow hydrothermal vent field on the North-Atlantic Ridge, at a depth of 2300 m. The purpose of this study was to investigate this organism’s replicative protein network. A first approach using conventional cloning of unknown sequences resulted in the cloning, expression and characterization of a family B DNA polymerase. A second approach consisted in starting a genome sequencing project and searching for target protein genes in the genomic sequence data.

A fragment of the family B DNA polymerase gene was PCR amplified using primers based on conserved sequences among related species. Sequences of gene termini were then determined using a custom version of the adaptor-PCR technique, allowing amplification of the full-length coding region. The DNA polymerase gene contained an intein-coding sequence which was subsequently removed by the overlap extension method. The gene was cloned in a pCR-T7-TOPO expression vector and expressed in Escherichia coli BL21-Rosetta. The mature form of the DNA polymerase has been purified to homogeneity as determined by SDS-PAGE analysis through heat treatment followed by three chromatographic steps. Its biochemical and enzymological properties have been investigated, as well as its performance in PCR. OGL-20P DNA polymerase appears to be more robust and versatile in PCR applications than other archaeal family B polymerases, making it an excellent replacement for the bacterial family A Taq polymerase as a multi-purpose PCR enzyme. Compared to Taq, specific activity and thermal stability are 3 times higher, both Km (for DNA and dNTP substrates) are lower and it has a proof-reading 3’-5’ exonuclease activity conferring a 1.7 x higher fidelity.

DNA sequences encoding other proteins involved in DNA replication were obtained from a preliminary draft of the OGL-20P genomic sequence. The sequencing was performed by the company 454 Life Sciences using an experimental massively parallel technology allowing the generation of several hundreds of thousands of short reads (100-120 nt) in just a few hours. A high coverage level (19.5-fold) was necessary for an efficient assembly, in order to compensate for the short read length. The 402516 obtained reads were assembled into 128 major contigs which were ordered based on complementarity between partial protein sequences encoded by contig ends. A search of replicative proteins based on hidden Markov models from the Pfam database revealed the presence of only 2 DNA polymerases (1 family B and 1 family D) and a single occurrence of CDC6, MCM-helicase, PCNA, RFC. All replicative protein sequences were highly similar to their homologues in the Thermococcales order, except for a different distribution of inteins. OGL-20P is the second representative of the Thermococcus genus (after T. kodakaraensis) and the fifth of the Thermococcales order (after Pyrococcus horikoshii, P. furiosus and P. abyssi) whose genome has been completely sequenced.