International Journal of Bioinformatics Research, ISSN: 0975–3087, Volume 2, Issue 1, 2010, pp-53-55

Insilico Identification of novel Coding Regions from Archeal Genome - *Aeropyrum pernix*

Sivasubramaniam Arunmeena and Piramanayagam Shanmughavel*
*Computational Biology and Bioinformatics Lab, Department of Bioinformatics, Bharathiar University
Coimbatore-641046, Tamilnadu, India, shanvel_99@yahoo.com

Abstract- Researches on archaeal microorganisms continue to excite the scientific community. Their unique adaptations that cater to hypersaline, hyperthermic, and hypothermic circumstances have incited research to manipulate those attributes for use in virtually every aspect of life. Adaptations in membrane, enzymes, and protein structures and components have potential applications in areas including electronics, agriculture, aquaculture, medicine, pharmaceuticals, food science, and nutrition. Although the time and effort required to new find archaeal homologues may be great, many believe that the economic and environmental benefits of such a breakthrough would be considerable enough to outweigh the challenges. An analysis of the archeal genome *Aeropyrum pernix*, showed that certain regions earlier thought to be 'non-coding' have significant sequence similarity to other protein sequences from archaea and other species. The available sequence analysis tools were used to identify a number of potential protein coding regions in these putative 'non coding' regions. We could identify 907 such regions and 282 of them apparently code for proteins present in archeal or other species. The remaining 625 regions are mostly start /stop conflicts. Of the 282 protein coding regions, only 64 code for proteins with homologues of known function. A good number of proteins show homology to proteins that are important for the survival of the organism. Hence these novel regions may be referred as homologues to coding regions. In addition Genome sequence collections should be regularly checked to improve gene prediction by sequence similarity and greater effort is required to make gene definitions consistent across related species.
Key words – *Aeropyrum pernix*, Extremophiles, non-coding regions

Introduction

The discovery and recognition of Archaea as the third domain of life on earth have led to exciting developments and characterization of a wide array of previously unknown microorganisms and associated components in the last few decades. Differences in composition and properties of major components such as cytoplasmic membranes, enzymes, and proteins of these extreme Archaea were found to play major roles in maintaining archaeal stability in seemingly inhospitable environments. Unique archaeal adaptations to drastically varying biosystems have aroused special interests in their respective potential in biotechnological applications [17]. Under such circumstances identification of novel proteins in these organisms becomes imperative to understand the secret behind their successful adaptation to extreme conditions. *Aeropyrum pernix* is the first crenarchaeote and first aerobic member of archaea for which the complete genome has been determined. The genus aeropyrum does not belong to any of the taxa known so far and represents one of the deepest phylogenetic lineages within the archea domain [7].Its vigorous motility at both room temperature or at 90◦c, its strictly aerobic character, heterotrophic and hyperthermoneutrophilic characters paves new interest in investigating this genome [16]. The sequence analysis studies of *Aeropyrum pernix* point to the distinct nature and well-defined adaptations of these organisms and to attain a deeper understanding into the relationships between the three domains: Bacteria, archea and eukaryotes. The first and foremost objective of the analysis of a newly sequenced genome is identification of protein coding genes. Despite the availability of completed whole genome sequencing projects, which provide useful comparison with close relatives among other organisms during annotation, accurate gene prediction remains quite difficult [5]. It is important to develop fast yet reliable computational methods that predict genes or potential coding regions [6]. Besides, genome knowledge base has to be updated regularly to include and disseminate the new information. A method to detect possible coding regions from the non-coding regions of the archeal genome *Aeropyrum pernix* is presented in this paper. Similar report is available on the discovery of novel coding regions in four archeal genomes [14]. However our aim is to investigate the regions that exhibit significant similarity to proteins that are essential for the survival of the organism.

Methods

All the genome annotation databases were searched thoroughly to identify all the gaps between open reading frames (ORFs) (1699 regions) [1], out of which gaps longer than 50 nucleotides were extracted from the genomic sequence and obtained 907 inter ORF regions from the complete genome of *Aeropyrum pernix*. To detect significant sequence similarities that could indicate the occurrence of protein coding segments, all such extracted non-coding regions were matched against a non-redundant protein

sequence database. These 907 nt sequences were used as queries for the BLAST X 2.0 (Gish and States, 1993) runs against the non-redundant protein sequence database. Hits with P-value <10-6 were extracted and manually examined. All computations were performed on a sunblade2000 workstation running exclusively on UNIX platform. Certain regions may be missed because of a stringent threshold p-value 10-6. However, with ample availability of complete genome sequences of related species, this cut-off appears to be permissible.

Results

The list of 907 regions contains both complete ORFS (282) and sections of previously characterized ORFs with conflicting start/end sites (625). The former can be identified by similarity to homologues protein sequences in the database, while the latter detected with reference to similarity, direction and annotation of downstream ORFs.A major portion of the 282 newly discovered coding regions appear to code for homologues of previously reported genome sequences. Most of these protein coding regions match other archeal proteins within the same or in a related species, therefore making the predictions more reliable. All of the newly identified protein coding genes except a hypothetical protein in chloroflexus aurantiacus have archeal proteins as closest homologues. Our analysis showed that *Aeropyrum pernix* has a good amount of newly identified proteins coding regions (282). There are 64 cases which have similarity to proteins of known functions (Table.1). Since a.pernix is strictly aerobic the regions that show homology with the proteins involved in TCA cycle are crucial. ORF upstream of APE_1056.1 codes for 2 oxoacid:ferredoxin oxidoreductase.The presence of 2 oxoacid: ferredoxin oxidoreductase is exclusively reported in *Aeropyrum pernix* (10 Kawarabayasi et al., 1999) and sulfolobus sp strain 7 [18] whereas prokaryotes use alpha –ketoglutarate in the TCA cycle. The others which show closest homologues to proteins or enzymes of TCA cycle are upstream of APE_0006.1, APE_0367.1, APE_1035.1 that codes for ferrodoxin, isocitrate dehydrogenase (NADP), glucokinase respectively. The ORF upstream of APE_0010.1 code for dihydroxy-acid dehydratase, the ORF upstream of APE_0107 codes for lysyl-tRNA synthatase, the ORF upstream of APE_0436B.1 codes for threonyl -tRNA synthatase , the ORF upstream of APE_0471.1 codes for alanyl –tRNA synthatase, the ORF upstream of APE_0472B.1 codes for mu-crystallin, the ORF upstream of APE_0560 codes for aspartate kinase , the ORF upstream of APE_0621.1codes for cystathionine beta-synthase, the ORF upstream of APE_0702.1 codes for glutamate dehydrogenase, the ORF upstream of APE_0734 codes for threonine synthase, the ORF upstream of APE_0880b codes for valyl-tRNA synthetase, the ORF upstream of APE_0966.1 codes for seryl-tRNA synthetase, the ORF upstream of APE_1078 codes for thiazole biosynthetic enzyme, the ORF upstream of APE_1081.1 codes for ABC-type cobalt transport system, ATPase component ,enzymes involved in amino acid biosynthesis. The ORF upstream of APE_0067.1 codes for DNA-directed DNA polymerase pfu polymerase, the ORF upstream of APE_0492.1 codes for methylated-DNA--protein-cysteine methyltransferase, enzyme involved in DNA replication. The ORF upstream of APE_0756.1 codes for translation initiation factor 2 beta subunit, the ORF upstream of APE_1029 Codes for translation initiation factor 5A, the ORF upstream of APE_1232 codes for 30S ribosomal protein S28 e which plays role in translation. The ORF upstream of APE_1013.1 codes for thermosome, subunit (alpha) which plays vital role in protein folding. The ORF upstream of APE_0050 codes for molybdopterin biosynthesis mog protein,the ORF upstream of APE_0313.1 codes for putative transketolase,the ORF upstream of APE_0419a codes for serine/threonine protein phosphatase, the ORF upstream of APE_0537a codes for succinyl-CoA synthetase beta chain,the ORF upstream of APE_0825.1 codes for acyl-CoA dehydrogenase, the ORF upstream of APE_0826a codes for inorganic pyrophosphatase, the ORF upstream of APE_0957 codes for alcohol dehydrogenase, the ORF upstream of APE_1184 codes for long-chain-fatty-acid--CoA ligase, enzymes involved in other metabolic processes. Some of the detected regions which match hypothetical proteins cannot always be corroborated by consistent predictions in other genomes, and they may be isolated cases of false positive ORF assignments. However, in cases where all related species contain at least one such protein, it is compulsory to include the regions predicted herein. Additional evidence can also be provided by the presence of duplicated genes within the same genome.

Discussions

Transcription, replication, translation, energy giving processes such as glycolysis, TCA cycle, and biosynthesis processes like amino acid biosynthesis and vitamin biosynthesis are considered to be important for the survival of the organism. So, the organism will have multiple copies of those genes that are required for its survival. This is further proved in our analysis, that the coding regions of enzymes or proteins involved in the above said metabolic processes are found to be homologues to the the non-coding regions of *Aeropyrum pernix*. Aldehyde ferredoxin oxido reductase is noted to play primary role in the catabolism of sugars or amino acids in *Pyrococcus furiosus* [12] and

Thermococcus litoralis [13]. It is interesting to note that two different non coding regions upstream of APE_0167 and APE_0168.1 that are seen very close to each other show homology with coding region of the same enzyme aldehyde ferrodoxin oxido reductase which reveals that the activity of this enzyme is crucial in *Aeropyrum pernix*. Similar methodology is also reported in the discovery of bacterial genes [15] in the databases, in particular for E-Coli [6]. One can also use this approach to detect sequencing errors. Without access to primary information, it is not possible for us to reconstruct the actual coding sequences, therefore only the boundaries of the potential coding regions are reported. Some of the sequences that are predicted in our analysis, however, may indeed be correct, and those genes should then be considered as cryptic genes [9] or vestigial sequences. The genome of Rickettsia prowazekii is known to contain non-coding regions that appear to be genes deactivated by several instances of mutations [2]. Indeed, the two categories are not mutually exclusive, as genes that have been deactivated by evolutionary mechanisms may in principle revert to a functional state. It is possible to have some degree of uncertainty in the results brought out by the widely used genome sequencing annotation procedures and in particular ORF detection [11] or automatic ORF translation by TrEMBL [14]. In addition to that as each run may correspond to different coding regions, it is important to consider every uniquely matching region. During annotation of the four genomes, it is found that the results differ drastically with respect to the reported false negative ORF calls. This indicates an inadequacy of standardization and an inconsistency in genome sequence annotation [3]. It can be concluded that it is important to perform frequent searches to match non coding regions of nucleotide sequences against protein databases after each genome sequencing projects to get rid of inconsistencies and false positives (or negative) ORF assignments .In addition since good number of non-coding regions show homology to proteins (64)) of known function, instead of referring these regions as non coding regions it may be referred as homologues of coding regions.

Acknowledgements

This work was supported by DBT Bioinformatics Facility, Delhi, India

References

[1] Altschul S.F., Boguski M.S., Gish W. and Wootton J.C. (1994) *Nat Genet*,6(2), 119-29.

[2] Andersson S.G., Zomorodipour A., Andersson J.O., Sicheritz-Pontén T., Alsmark U.C., Podowski R.M., Näslund A.K., Eriksson A.S., Winkler H.H. and Kurland C.G. (1998) *Nature*, 396(6707),133-40.

[3] Andrade M.A. and Sander C. (1997) *Curr Opin Biotechnol*,8(6),675-83.

[4] Apweiler R., Gateau A.,Contrino S., Martin M.J., Junker V., O'Donovan C., Lang F.,Mitaritonna N., Kappus S. and Bairoch A.(1997) *Proc Int Conf Intell Syst Mol Biol*, 5,33-43.

[5] Bocs S., Danchin A. and Médigue C. (2002) *BMC Bioinformatics*, 3,5.

[6] Borodovsky M., Rudd K.E. and Koonin E.V. (1994) *Nucleic Acids Res*. 22(22),4756-67.

[7] Faguy D.M. and Doolittle W.F. (1999) *Curr Biol*, 9,R883-6.

[8] Gish W.and States D.J. (1993) *Nat Genet*, 3,266-72.

[9] Hall B.G., Yokoyama S. and Calhoun D.H.(1983) *Mol Biol Evol*, 1(1),109-24.

[10] Kawarabayasi Y., Hino Y., Horikawa H., Yamazaki S., Haikawa Y., Jin-no K., Takahashi M., Sekine M., Baba S., Ankai A., Kosugi H., Hosoyama A., Fukui S., Nagai Y., Nishijima K., Nakazawa H., Takamiya M., Masuda S., Funahashi T., Tanaka T., Kudoh Y., Yamazaki J., Kushida N., Oguchi A., Kikuchi H et al. (1999) *DNA Res*, 6(2),83-101,145-52

[11] McIninch J.D., Hayes W.S. and Borodovsky M. (1996) *Proc Int Conf Intell Syst Mol Biol,* 4,165-75.

[12] Mukund S. and M. W. W. Adams. (1991) *J. Biol. Chem.*, 266(22),14208-14216.

[13] Mukund S. and M. W. W. Adams. (1993) *J. Biol. Chem.*, 268(18),13592-13600.

[14] Ragavan S. and Ouzounis A. (1999) *Nucleic Acids Res.*, 27(22), 4405-4408.

[15] Robison K., Gilbert W., Church G.M. (1994) *Nat Genet.*, 7(2),205-14.

[16] Sako Y., Nomura N., Uchida A., Ishida Y., Morii H., Koga Y., Hoaki T. and Maruyama T. (1996) *Int J Syst Bacteriol*, 46(4),1070-7.

[17] Yihwa Yang., Daniel T. Levick., Caryn K. Just. **(**2008) *Journal of Young Investigators,* 17(4).

[18] Zhang Q., Iwasaki T., Wakagi T., Oshima T. (1996) *J Biochem*. ,120(3),587-99.