

Approaches to determine biological function- dissecting the diverse crotonase super-family

Ashwin Sivakumar^{1,2}, Swapan Mallick¹, Liisa Holm^{1,2}

Keywords: Crotonases, Enoyl CoA hydratase, bacteria, archaea, functional annotation, Escherichia coli, Sub-family classification, patterns, motifs, function, proteins, PCA (Principal Component Analysis), Density points clustering, fused genes, key residues.

1 Introduction.

A protein's function is encoded within putatively functional signatures or motifs that represent residues involved in both functional conservation and functional divergence within a set of homologous proteins at various levels of hierarchy that is, super-families, families and sub-families. The crotonases represent a diverse set of proteins sharing a common structural scaffold at a super-family level and a common catalytic strategy by catalyzing reactions that usually involve a thioester enolate anion intermediate stabilized by hydrogen bonding with two peptidic NH groups in an oxyanion hole. The crotonase sub-classes diversify to catalyze different overall reactions. Here we describe a pipeline approach that uses only sequence information to find 'putative' functional motifs and residues of functional significance in a sequence level sub-family of proteins, which together make up diverse super-family of proteins. This approach addresses key issues like defining sequence level super-families, defining meaningful sub-groups within super-families, characterizing putative functional fingerprints and residues. Tree based approaches solely based on sequence similarity do not work in such super-families due to the diversity amongst its sequences. Here we describe the application of this approach in case of crotonases. Homologues of seven known crotonases in Escherichia coli k12 genome have diversified to catalyze ten different catalytic reactions in other archaeal and bacterial genomes in addition to the seven reactions known to be catalysed by the crotonases belonging to the Escherichia coli k12 genome.

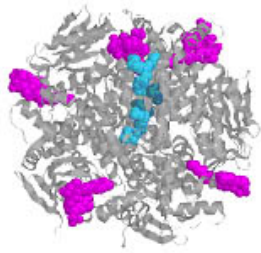
Super-family	Homologous group of proteins having a common ancestor.
Supra-family	Homologous group of proteins whose members catalyze different overall reactions and their reactions share no common mechanistic strategy but active site residues might be conserved.
Family	Each super-family is made of one or more domain families. Here definition of a domain corresponds to that of structural domains.
Sub-family/Sub-Group	Smaller group of sequences within a super-family, which share a function specific to this group.
Sequence pattern/Motifs	Are regular expressions that can be represented as prosite like patterns.
Specificity motifs/Phylo-motifs	Motifs or sequence patterns which are specific to a sub-family that usually contain residues involved in the functional specificity of the sub-family. These patterns are found in a multiple alignment of a sub-family and specificity is inferred by searching this pattern against an NRDB [6]
Paralogous groups/gene families	Groups of sequences within a genome thought to have a common evolutionary origin or ancestor. These genes have arisen by duplication of an ancestral gene.
Multi-Modular proteins/Composite Proteins	Are proteins, which exists as fused gene in one or more genomes while occurring as two or more homologous genes remaining separate in other genomes. Modules/component proteins can be identified by automation of the Riley approach [1].
Module/Component Proteins	These are complete functional proteins, which exist as two or more components of a fused gene in another genome.

¹ Bioinformatics group, Institute of Biotechnology, University of Helsinki, PO Box 56

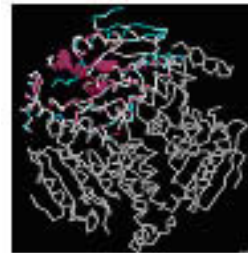
² Corresponding authors: ashwin.sivakumar@helsinki.fi, liisa.holm@helsinki.fi

2 Methods

Paralogous groups were constructed for 52 bacterial and archaeal genomes using the paralogous groups classified for the *Escherichia coli* k12 genome [3]. These groups were generated after the proteins were broken down into modules of component proteins [1,4], which represent the function of a gene product. One of the paralogous groups were the crotonases, which catalyze diverse reactions while sharing a common catalytic scaffold. The crotonases were then classified into 17 distinct sub-families using an iterated Principal component approach [Casari *G et al*] and density of points clustering approach [5]. The clusters/sub-families correspond with each other. We then mine specificity motifs for each sub-family both manually and Tierias [Rigoutsos *I et al*]. Mutual information scores for columns of a multiple sequence alignment of a group of proteins divided into specificity groups can be used to identify individual residues, which are specific to each of the groups [Mirny *et al*]. Every column in the multiple alignment was assigned mutual information scores. Residues scoring highly on mutual information within the specificity motifs are likely to be the residues involved in functional specificity of each-sub-groups. Sub-groups found within the crotonases belong to the same class and catalyse the same reaction. We validate our sub-families using experimental data and structures available for some of the predicted sub-families. Here we also propose a new approach for an hierarchical functional classification system which uses only sequence information which would move from a super-family level (where all proteins are homologous) to a sub-family level which are functional. This would be a complementary system to the existing domain classification systems like ADDA [2].



Generic Enoyl CoA hydratase
[ILV]-x(3)-E-x(7)-V-[GA]-x-[IVL]-x-L-N-R-P



Methylmalonyl CoA decarboxylase
R-x(2)-[IV]-[IVL]-x(8)-F-x(2)-G-x-D-[ILV]

References

- [1] Serres M, Riley, M. 2004 Structural domains, protein modules and sequence similarities enrich our understanding of the *Shewanella oneidensis* MR-I: *Proteome. OMICS Winter*; 8(4): 306-21
- [2] Heger A, Wilton CA, Sivakumar A, Holm L. 2005. ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res. Jan 1*; 33 Database Issue:D188-91
- [3] Sivakumar A, Wilton, C, Holm L. A database of homologous groups in bacterial and archeal genomes. *unpublished*
- [4] Sivakumar A, Holm, L. From sequences to a functional unit. *Submitted*
- [5] Wicker N, Dembele D, Raffelsberger W, Poch O. *Nucleic Acids research* 2002 Sep 15, 30 (18): 3992-4000.
- [6] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein Knowledgebase *Nucleic Acids Res.* 32: D115-D119.