16

Identification of Functional Diversity in the Enolase Superfamily Proteins

Kaiser Jamil[1] and M. Sabeena[2]
[1]*Head, Genetics Department, Bhagwan Mahavir Medical Research Centre, Mahavir Marg, Hyderabad, & Dean, School of Life Sciences (SLS), Jawaharlal Nehru Institute of Advanced Studies, (JNIAS)*
[2]*Research Associate, Centre for Biotechnology and Bioinformatics- (CBB), Jawaharlal Nehru Institute of Advanced Studies, (JNIAS), Secunderabad India*

1. Introduction

The Escherichia coli K12 genome is a widely studied model system. The members of the Enolase superfamily encoded by *E.coli* catalyze mechanistically diverse reactions that are initiated by base-assisted abstraction of the α-proton of a carboxylate anion substrate to form an enodiolate intermediate (Patricia C ,1996). Six of the eight members of the Enolase superfamily encoded by the Escherichia coli K-12 genome have known functions (John F, 2008). The members share a conserved tertiary structure with a two-domain architecture, in which three carboxylate ligands for the Mg^{2+} ion as well as the acid/base catalysts are located at the C-terminal ends of the β-strands in a $(\beta/\alpha)_7\beta$-barrel [modified $(\beta/\alpha)_8$- or TIM-barrel] domain and the specificity-determining residues are located in an N-terminal α+β capping domain.

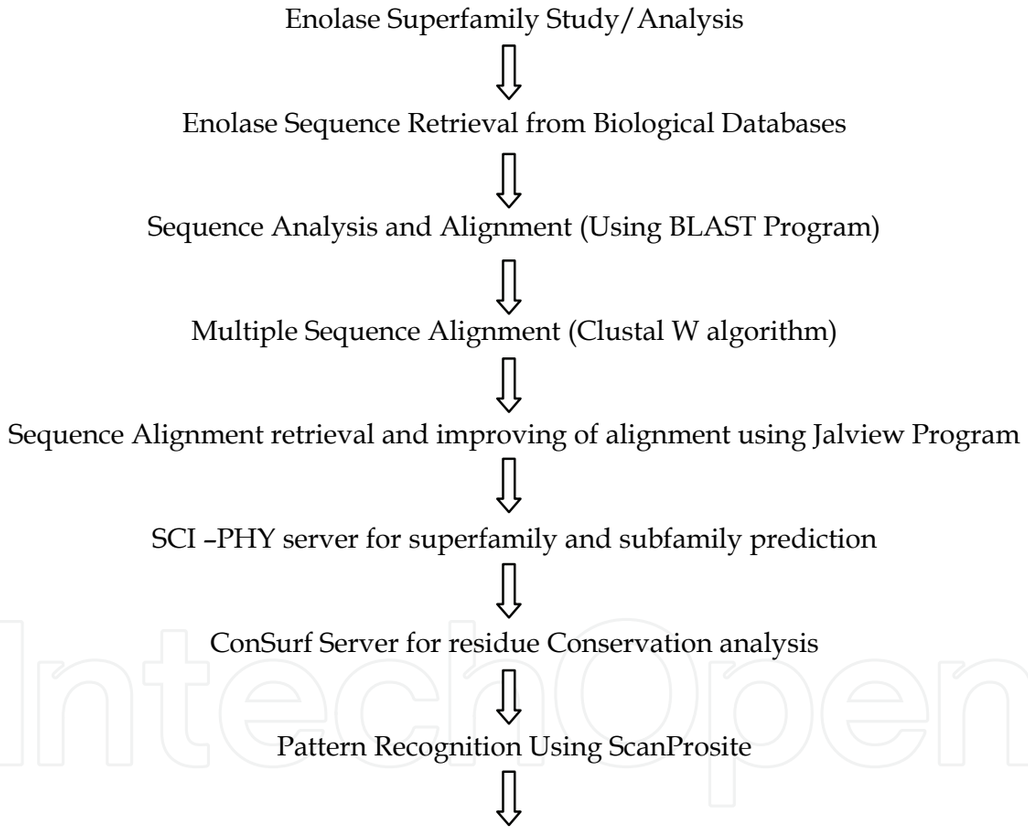
The rapid accumulation of data has led to an extraordinary problem of redundancy, which must be confronted in almost any type of statistical analysis. An important goal of bioinformatics is to use the vast and heterogeneous biological data to extract patterns and make discoveries that bring to light the ''unifying'' principles in biology. (Kaiser Jamil, 2008)Because these patterns can be obscured by bias in the data, we approach the problem of redundancy by appealing to a well known unifying principle in biology, evolution. Bioinformatics has developed as a data-driven science with a primary focus on storing and accessing the vast and exponentially growing amount of sequence and structure data (Gerlt JA, 2005)

Protein sequences and their three-dimensional structures are successful descendants of evolutionary process. Proteins might have considerable structural similarities even when no evolutionary relationship of their sequences can be detected (Anurag Sethi, 2005). This property is often referred to as the proteins sharing only a ''fold". Of course, there are also sequences of common origin in each fold, called a ''superfamily", and in them groups of sequences with clear similarities, are designated as ''family".

The concept of protein superfamily was introduced by Margaret Dayholff in the 1970 and was used to partition the protein sequence databases based on evolutionary consideration

(Lindahl E, 2000). The objective of this study was to analyse the functional diversity of the enolase gene superfamily. The gene superfamily consisting of twelve genes possess enzymatic functions such as L-Ala-D/L-Glu epimerase, Glucarate dehydratase, D-galactarate dehydratase, 2-hydroxy-3-oxopropionate reductase,l-o-succinylbenzoate synthase, D-galactonate dehydratase,[12]. 5-keto-4-deoxy-D-glucarate aldolase, L-rhamnonate dehydratase, 2-keto-3-deoxy-L-rhamnonate aldolase, Probable galactarate transporter, and Probable glucarate transporter (Steve EB ,1998)
This study was carried out to determine the Probable glucarate transporter (D-glucarate permease) features relating enolase superfamily sequences to structural hinges, which is important for identifying domain boundaries, and designing flexibility into proteins functions also helps in understanding structure-function relationships.

2. Methodology



Flowchart represents the materials and methods

2.1 UniProt KB for genomic sequence analysis

Enolase sequence from *E.coli* formed the basis for this study. The protein sequences were derived from UniProt KB, we found twelve sequences (Table 1). Most of the sequences in UniProt KB were derived from the conceptual translation of nucleotide sequences. The advantage of using UniProt KB was that it provides a stable, comprehensive, freely

accessible central resource on protein sequences and functional annotation. UniProt comprises of four major components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. We used this knowledge based computational analysis which helps for the functional annotation for the gene sequences shown below:

S.No	Accession Id	Sequence Name	Sequence
1.	P0A6P9	ENO_ECOLI Enolase OS=*Escherichia coli* (strain K12) GN=eno PE=1 SV=2	MSKIVKIIGREIIDSRGNPTVEAEVHLEGGFVGMA AAPSGASTGSREALELRDGDKSRFLGKGVTKAVA AVNGPIAQALIGKDAKDQAGIDKIMIDLDGTENK SKFGANAILAVSLANAKAAAAAKGMPLYEHIAE LNGTPGKYSMPVPMMNIINGGEHADNNVDIQEF MIQPVGAKTVKEAIRMGSEVFHHLAKVLKAKGM NTAVGDEGGYAPNLGSNAEALAVIAEAVKAAG YELGKDITLAMDCAASEFYKDGKYVLAGEGNKA FTSEEFTHFLEELTKQYPIVSIEDGLDESDWDGFAY QTKVLGDKIQLVGDDLFVTNTKILKEGIEKGIANS ILIKFNQIGSLTETLAAIKMAKDAGYTAVISHRSGE TEDATIADLAVGTAAGQIKTGSMSRSDRVAKYN QLIRIEEALGEKAPYNGRKEIKGQA
2.	P51981	AEEP_ECOLI L-Ala-D/L-Glu epimerase OS=*Escherichia coli* (strain K12) GN=ycjG PE=1 SV=2	MRTVKVFEEAWPLHTPFVIARGSRSEARVVVVEL EEEGIKGTGECTPYPRYGESDASVMAQIMSVVPQL EKGLTREELQKILPAGAARNALDCALWDLAARR QQQSLADLIGITLPETVITAQTVVIGTPDQMANSA STLWQAGAKLLKVKLDNHLISERMVAIRTAVPD ATLIVDANESWRAEGLAARCQLLADLGVAMLEQ PLPAQDDAALENFIHPLPICADESCHTRSNLKAL KGRYEMVNIKLDKTGGLTEALALATEARAQGFSL MLGCMLCTSRAISAALPLVPQVSFADLDGPTWLA VDVEPALQFTTGELHL
3.	P0AES2	GUDH_ECOLI Glucarate dehydratase OS=*Escherichia coli* (strainK12) GN=gudD PE=1 SV=2	MSSQFTTPVVTEMQVIPVAGHDSMLMNLSGAHA PFFTRNIVIIKDNSGHTGVGEIPGGEKIRKTLEDAIP LVVGKTLGEYKNVLTLVRNTFADRDAGGRGLQT FDLRTTIHVVTGIEAMLDLLGQHLGVNVASLLGD GQQRSEVEMLGYLFFVGNRKATPLPYQSQPDDSC DWYRRHEEAMTPDAVVRLAEAAYEKYGFNDFK LKGGVLAGEEEAESIVALAQRFPQARITLDPNGA WSLNEAIKIGKYLKGSLAYAEDPCGAEQGFSGRE VMAEFRRATGLPTATNMIATDWRQMGHTLSLQS VDIPLADPHFWTMQGSVRVAQMCHEFGLTWGS HSNNHFDISLAMFTHVAAAAPGKITAIDTHWIW QEGNQRLTKEPFEIKGGLVQVPEKPGLGVEIDMD QVMKAHELYQKHGLGARDDAMGMQYLIPGWT FDNKRPCMVR

S.No	Accession Id	Sequence Name	Sequence
4.	P39829	GARD_ECOLI D-galactarate dehydratase OS=*Escherichia coli* (strain K12) GN=garD PE=1 SV=2	MANIEIRQETPTAFYIKVHDTDNVAIIVNDNGLK AGTRFPDGLELIEHIPQGHKVALLDIPANGEIIRYG EVIGYAVRAIPRGSWIDESMVVLPEAPPLHTLPLA TKVPEPLPPLEGYTFEGYRNADGSVGTKNLLGITT SVHCVAGVVDYVVKIIERDLLPKYPNVDGVVGLN HLYGCVAINAPAAVVPIRTIHNISLNPNFGGEVM VIGLGCEKLQPERLLTGTDDVQAIPVESASIVSLQD EKHVGFQSMVEDILQIAERHLQKLNQRQRETCPA SELVVGMQCGGSDAFSGVTANPAVGYASDLLVR CGATVMFSEVTEVRDAIHLLTPRAVNEEVGKRLL EEMEWYDNYLNMGKTDRSANPSPGNKKGGLAN VVEKALGSIAKSGKSAIVEVLSPGQRPTKRGLIYA ATPASDFVCGTQQVASGITVQVFTTGRGTPYGLM AVPVIKMATRTELANRWFDLMDINAGTIATGEET IEEVGWKLFHFILDVASGKKKTFSDQWGLHNQL AVFNPAPVT
5.	P29208	MENC_ECOLI o-succinylbenzoat e synthase OS=*Escherichia coli* (strain K12) GN=menC PE=1 SV=2	MRSAQVYRWQIPMDAGVVLRDRRLKTRDGLYV CLREGEREGWGEISPLPGFSQETWEEAQSVLLAW VNNWLAGDCELPQMPSVAFGVSCALAELTDTLP QAANYRAAPLCNGDPDDLILKLADMPGEKVAK VKVGLYEAVRDGMVVNLLLEAIPDLHLRLDANR AWTPLKGQQFAKYVNPDYRDRIAFLEEPCKTRD DSRAFARETGIAIAWDESLREPDFAFVAEEGVRAV VIKPTLTGSLEKVREQVQAAHALGLTAVISSSIESS LGLTQLARIAAWLTPDTIPGLDTLDLMQAQQVRR WPGSTLPVVEVDALERLL
6.	Q6BF17	DGOD_ECOLI D-galactonate dehydratase OS=*Escherichia coli* (strain K12) GN=dgoD PE=1 SV=1	MKITKITTYRLPPRWMFLKIETDEGVVGWGEPVIE GRARTVEAAVHELGDYLIGQDPSRINDLWQVMY RAGFYRGGPILMSAIAGIDQALWDIKGKVLNAPV WQLMGGLVRDKIKAYSWVGGDRPADVIDGIKTL REIGFDTFKLNGCEELGLIDNSRAVDAAVNTVAQ IREAFGNQIEFGLDFHGRVSAPMAKVLIKELEPYR PLFIEEPVLAEQAEYYPKLAAQTHIPLAAGERMFS RFDFKRVLEAGGISILQPDLSHAGGITECYKIAGM AEAYDVTLAPHCPLGPIALAACLHIDFVSYNAVL QEQSMGIHYNKGAELLDFVKNKEDFSMVGGFFK PLTKPGLGVEIDEAKVIEFSKNAPDWRNPLWRHE DNSVAEW
7.	P23522	GARL_ECOLI 5-keto-4-deoxy-D-glucarate aldolase OS=*Escherichia coli* (strain K12) GN=garL PE=1 SV=2	MNNDVFPNKFKAALAAKQVQIGCWSALSNPIST EVLGLAGFDWLVLDGEHAPNDISTFIPQLMALKG SASAPVVRVPTNEPVIIKRLLDIGFYNFLIPFVETKE EAELAVASTRYPPEGIRGVSVSHRANMFGTVADY FAQSNKNITILVQIESQQGVDNVDAIAATEGVDGI FVGPSDLAAALGHLGNASHPDVQKAIQHIFNRA SAHGKPSGILAPVEADARRYLEWGATFVAVGSDL GVFRSATQKLADTFKK

S.No	Accession Id	Sequence Name	Sequence
8.	P77215	RHAMD_ECOLI L-rhamnonate dehydratase OS=*Escherichia coli* (strain K12) GN=yfaW PE=1 SV=2	MTLPKIKQVRAWFTGGATAEKGAGGGDYHDQG ANHWIDDHIATPMSKYRDYEQSRQSFGINVLGTL VVEVEAENGQTGFAVSTAGEMGCFIVEKHLNRFI EGKCVSDIKLIHDQMLSATLYYSGSGGLVMNTISC VDLALWDLFGKVVGLPVYKLLGGAVRDEIQFYA TGARPDLAKEMGFIGGKMPTHWGPHDGDAGIR KDAAMVADMREKCGEDFWLMLDCWMSQDVN YATKLAHACAPYNLKWIEECLPPQQYESYRELKR NAPVGMMVTSGEHHGTLQSFRTLSETGIDIMQPD VGWCGGLTTLVEIAAIAKSRGQLVVPHGSSVYSH HAVITFTNTPFSEFLMTSPDCSTMRPQFDPILLNEP VPVNGRIHKSVLDKPGFGVELNRDCNLKRPYSH
9.	P76469	KDRA_ECOLI 2-keto-3-deoxy-L-rhamnonate aldolase OS=*Escherichia coli* (strain K12) GN=yfaU PE=1 SV=1	MNALLSNPFKERLRKGEVQIGLWLSSTTAYMAEI AATSGYDWLLIDGEHAPNTIQDLYHQLQAVAPY ASQPVIRPVEGSKPLIKQVLDIGAQTLLIPMVDTAE QARQVVSATRYPPYGERGVGASVARAARWGRIE NYMAQVNDSLCLLVQVESKTALDNLDEILDVEGI DGVFIGPADLSASLGYPDNAGHPEVQRIIETSIRRI RAAGKAAGFLAVAPDMAQQCLAWGANFVAVG VDTMLYSDALDQRLAMFKSGKNGPRIKGSY
10.	P0AA80	GARP_ECOLI Probable galactarate transporter OS=*Escherichia coli* (strain K12) GN=garP PE=1 SV=1	MILDTVDEKKKGVHTRYLILLIIFIVTAVNYADRA TLSIAGTEVAKELQLSAVSMGYIFSAFGWAYLLM QIPGGWLLDKFGSKKVYTYSLFFWSLFTFLQGFVD MFPLAWAGISMFFMRFMLGFSEAPSFPANARIVA AWFPTKERGTASAIFNSAQYFSLALFSPLLGWLTF AWGWEHVFTVMGVIGFVLTALWIKLIHNPTDHP RMSAEELKFISENGAVVDMDHKKPGSAAASGPK LHYIKQLLSNRMMLGVFFGQYFINTITWFFLTWFP IYLVQEKGMSILKVGLVASIPALCGFAGGVLGGVF SDYLIKRGLSLTLARKLPIVLGMLLASTIILCNYTN NTTLVVMLMALAFFGKGFGALGWPVISDTAPKEI VGLCGGVFNVFGNVASIVTPLVIGYLVSELHSFNA ALVFVGCSALMAMVCYLFVVGDIKRMELQK
11.	P0ABQ2	GARR_ECOLI 2-hydroxy-3-oxopropionate reductase OS=*Escherichia coli* (strain K12) GN=garR PE=1 SV=1	MKVGFIGLGIMGKPMSKNLLKAGYSLVVADRNP EAIADVIAAGAETASTAKAIAEQCDVIITMLPNSP HVKEVALGENGIIEGAKPGTVLIDMSSIAPLASREI SEALKAKGIDMLDAPVSGGEPKAIDGTLSVMVGG DKAIFDKYYDLMKAMAGSVVHTGEIGAGNVTKL ANQVIVALNIAAMSEALTLATKAGVNPDLVYQA IRGGLAGSTVLDAKAPMVMDRNFKPGFRIDLHIK DLANALDTSHGVGAQLPLTAAVMEMMQALRA DGLGTADHSALACYYEKLAKVEVTR

S.No	Accession Id	Sequence Name	Sequence
12.	Q46916	GUDP_ECOLI Probable glucarate transporter OS=*Escherichia coli* (strain K12) GN=gudP PE=1 SV=1	MSSLSQAASSVEKRTNARYWIVVMLFIVTSFNYG DRATLSIAGSEMAKDIGLDPVGMGYVFSAFSWAY VIGQIPGGWLLDRFGSKRVYFWSIFIWSMFTLLQG FVDIFSGFGIIVALFTLRFLVGLAEAPSFPGNSRIVA AWFPAQERGTAVSIFNSAQYFATVIFAPIMGWLT HEVWSHVFFFMGGLGIVISFIWLKVIHEPNQHPG VNKKELEYIAAGGALINMDQQNTKVKVPFSVKW GQIKQLLGSRMMIGVYIGQYCINALTYFFITWFPV YLVQARGMSILKAGFVASVPAVCGFIGGVLGGIIS DWLMRRTGSLNIARKTPIVMGMLLSMVMVFCNY VNVEWMIIGFMALAFFGKGIGALGWAVMADTA PKEISGLSGGLFNMFGNISGIVTPIAIGYIVGTTGSF NGALIYVGVHALIAVLSYLVLVGDIKRIELKPVAG Q

Table 1. Enolase Sequences from *E.coli* –K12 Strain (from UNIPROT-KB in Fasta format)

2.2 BLAST program for sequence analysis and alignment

Basic Local Alignment Search Tool (BLAST) is one of the most heavily used sequence analysis tools we have used to perform Sequence Analysis and Alignment. BLAST is a heuristic that finds short matches between two sequences and attempts to start alignments. In addition to performing alignments, BLAST provides statistical information to help decipher the biological significance of the alignment as 'expect' value. (Scott McGinnis, 2004). Using this BLAST program the twelve gene sequences were aligned against archaea and bacteria. The sequences were sorted out according to the existing gene names with similarity and the fused genes were removed.

2.3 Clustal W program for multiple sequence alignment

Multiple sequence alignments are widely acknowledged to be powerful tools in the analysis of sequence data.(Sabitha Kotra et al 2008) Crucial residues for activity and for maintaining protein secondary and tertiary structures are often conserved in sequence alignments. Hence, multiple sequence alignment was done for all the enolase gene sequences based on the ClustalW algorithm using the tool BioEdit software program. We determined the alignments which is the starting points for evolutionary studies. Similarity is a percentage sequence match between nucleotide or protein sequences. The basic hypothesis involved here was that similarity relates to functionality, if two sequences are similar, they will have related functionalities.

Realigned the obtained Multiple Sequence Alignments (MSA) using ClustalW (Muhummad Khan and Kaiser Jamil, 2010). Using MSA we could obtain high score for the conserved regions, compared to the reported query sequences. So we viewed the multiple alignment result using a program 'Jalview' which improved the multiple alignment. With this program we could extract and get the complete alignment of all sequences for realigning to the query sequence to get better results (Fig. 1). Jalview is a multiple alignment editor written in Java. It is used widely in a variety of web pages which is available as a general purpose alignment editor. The image below shows the result when Jalview has taken the

full length sequences and realigned them (using Clustalw) to the query sequence. The alignment has far fewer gaps and more similarities to the entire portion of the query sequences.

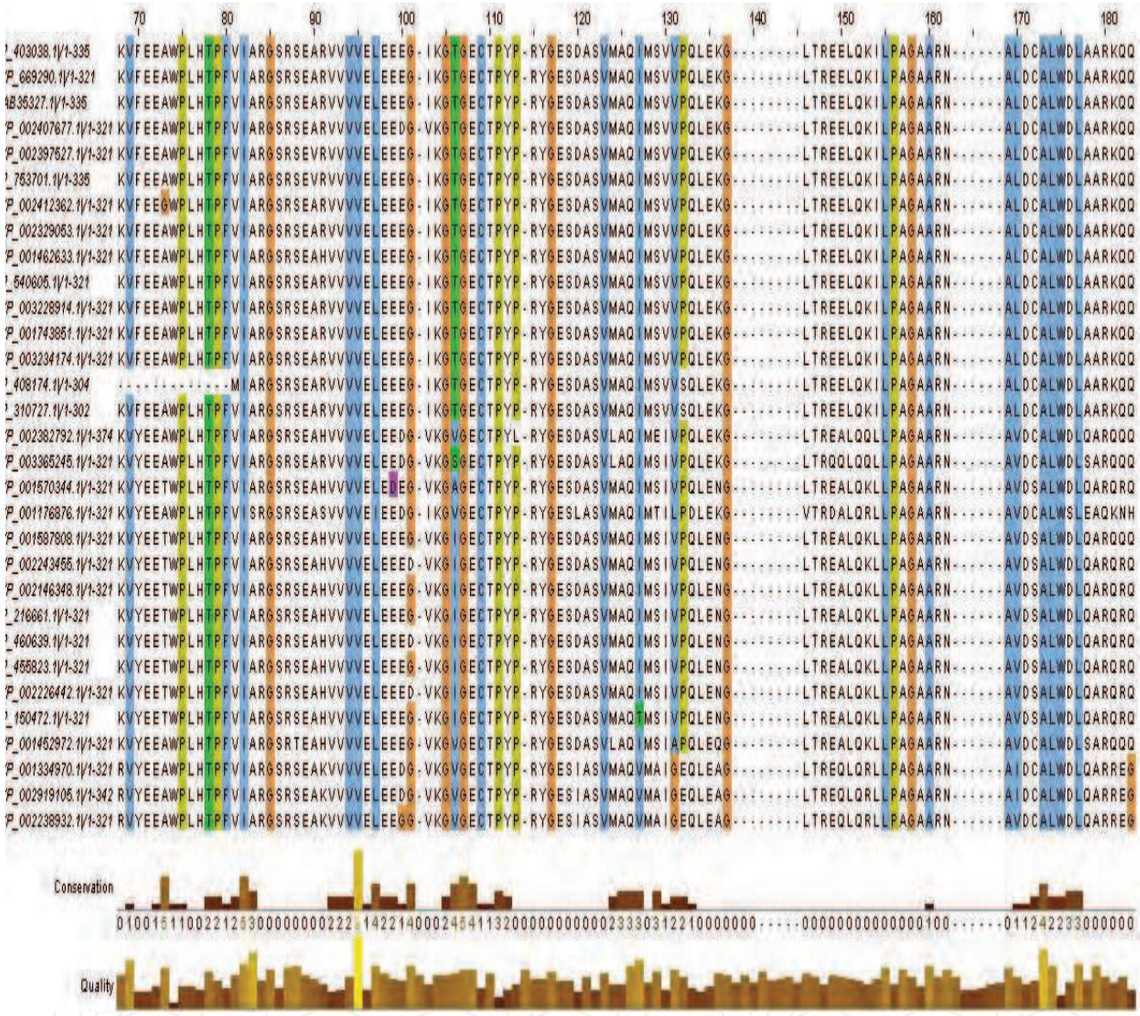


Fig. 1. Multiple Sequence Alignment as shown in Jalview

2.4 SCI –PHY server for superfamily and subfamily prediction

Using SCI-PHY server we found subfamilies/subclasses present in the aligned sequences, which merged into five groups. The corresponding pattern for each group of subfamily sequences was found by using ScanProsite and PRATT. A low-level simple pattern-matching application can prove to be a useful tool in many research settings (Doron Betel, 2000). Many of these applications are geared toward heuristic searches where the program finds sequences that may be closely related to the query nucleotide/protein sequences.

2.5 ConSurf server for conservation analysis

For each subfamily sequences the corresponding PDB ID using ConSurf Server was determined. ConSurf-DB is a repository of ConSurf Server which used for evolutionary

conservation analysis of the proteins of known structures in the PDB. Sequence homologues of each of the PDB entries were collected and aligned using standard methods. The algorithm behind the server takes into account the phylogenetic relations between the aligned proteins and the stochastic nature of the evolutionary process explicitly. The server assigned the conservation level for each position in the multiple sequence alignment (Ofir Goldenberg, 2002). Identified specific pattern for each of the FASTA format sequence from PDB files using ScanProsite and some of the key residues that comprise the functionally important regions of the protein (Ofir Goldenberg, 2002). We determined the residues present in each of PDB files denoting subfamilies using Swiss PDB Viewer. Mapped out all the residues in color with the help of Rasmol by finding the specific pattern.

3. Results and discussion

This study is an attempt to determine the functional diversity in enolase superfamily protein. The approach we used is a all pairwise alignment of the sequences followed by a clustering of statistically significant pairs into groups or subfamilies by making sure that there is a common motif holding all the members together. Multiple sequence alignment and pattern recognition methods were included in this. The study analyzed the possible subfamilies in Enolase protein superfamily which shares in organisms such as archaea, bacteria with respect to *E.coli* and finally predicted five superfamilies which may play a role in functional diversity in Enolase superfamily protein.

Generally 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. Protein function divergence is according to local structural variation around the active sites (Changwon K, 2006). Even when proteins have similar overall structure, the function could be different from each other. Accurate prediction of residue depth would provide valuable information for fold recognition, prediction of functional sites, and protein design. Proteins might have considerable structural similarities even when no evolutionary relationship of their sequences can be detected. This property is often referred to as the proteins sharing ie; a ''fold". Of course, there are also sequences of common origin in each fold, called a ''superfamily", and in them there are groups of sequences with clear similarities designated as ''family". These sequence-level superfamilies can be categorized with many Bioinformatics approaches (LevelErik L , 2002)

3.1 Functional/ structural validation

The functions of the five identified protein family include:

3.1.1 Group 1

Mandelate racemase / muconate lactonizing enzyme family signature-1: which is an independent inducible enzyme cofactor. Mandelate racemase (MR) and muconate lactonizing enzyme (MLE) catalyses separate and mechanistically distinct reactions necessary for the catabolism of aromatic acids Immobilization of this enzyme leads to an enhanced activity and facilitates its recovery

MR_MLE_1 Mandelate racemase / muconate lactonizing enzyme family signature 1: (Fig.2)

Polymer: 1

Type: polypeptide(L)

Length: 405 Chains:A, B, C, D, E, F, G, H

Functional Protein: PDB ID: 3D46 chain A in E-val 0.0.

Possible amino acid pattern found in chain A

I-x(1,3)-Q-P-D-[ALV]-[ST]-H-[AV]-G-G-I-[ST]-E-x(2)-K-[IV]-A-[AGST]-[LM]-A-E-[AS]-[FY]-D-V-[AGT]-[FLV]-[AV]-[LP]-H-C-P-L-G-P-[IV]-A-[FL]-A-[AS]-[CS]-L-x-[ILV]-[DG]
Key Residues

THR 136, SER 138, CYS 139,VAL 140, Asp 141, ALA 143, LEU 144, ASP 146, LEU 147, GLY 149, LYS 150, PRO 155, VAL 156, LEU 159, LEU 160, GLY 161

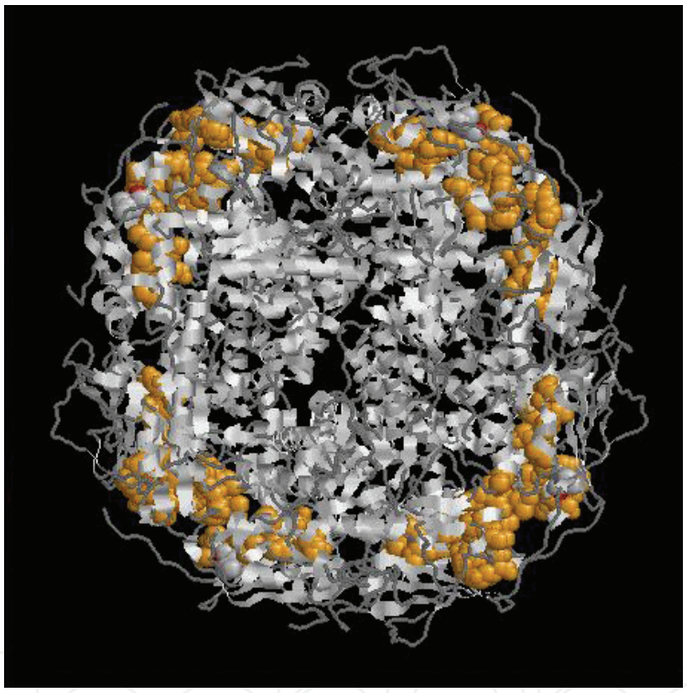


Fig. 2. Functional Protein Information (PDB Id: 3D46) The residues in yellow colour represents the identified functional residues in Group 1

3.1.2 Group 2

TonB-dependent receptor proteins signature-1 : TonB-dependent receptors is a family of beta-barrel proteins from the outer membrane of Gram-negative bacteria. The TonB complex senses signals from outside the bacterial cell and transmits them via two membranes into the cytoplasm, leading to transcriptional activation of target genes
TONB_DEPENDENT_REC_1 TonB-dependent receptor proteins signature 1 : (Fig.3)

Polymer:1

Type:polypeptide(L)

Length:99

Chains:A, B

Functional Protein: PDB ID: 3LAZ

Possible amino acid pattern found in 3LAZ

T-K-R-G-L-I-Y-A-A-T-P-A-S-D-F-V-C-G-T-Q-Q-V-A-S-G-I-T-V-Q-V-F-T-T-G-R-G-T-P-Y-G-L-M-A-V-P-V-I-K-M-A

Key Residues

GLU 88, SER89, VAL91, VAL92, PRO94, GLU95

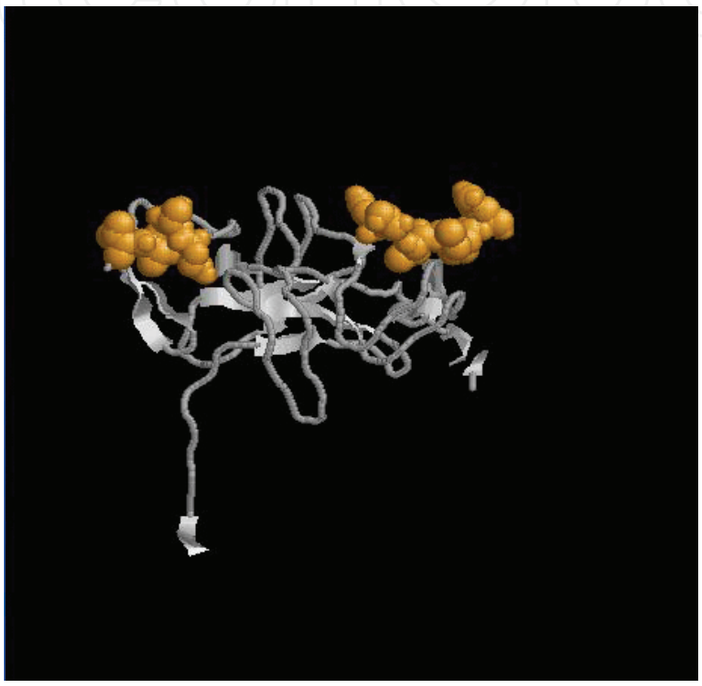


Fig. 3. Functional Protein Information (PDB Id: 3LAZ). The residues in yellow colour represents the identified functional residues in Group 2

3.1.3 Group 3

3-hydroxyisobutyrate dehydrogenase signature : This enzyme is also called beta-hydroxyisobutyrate dehydrogenase. This enzyme participates in valine, leucine and isoleucine degradation.

3_HYDROXYISOBUT_DH 3-hydroxyisobutyrate dehydrogenase signature : (Fig.4. a and Fig.4. b)

Polymer:1

Type:polypeptide(L)

Length:295

Chains:A, B

Functional Protein: PDB ID: 1YB4

Possible amino acid pattern found in 1YB4

G-[IMV]-[EK]-F-L-D-A-P-V-T-G-G-[DQ]-K-[AG]-A-x-E-G-[AT]-L-T-[IV]-M-V-G-G-x(2)-[ADEN]-[ILV]-F-x(2)-[LV]-x-P-[IV]-F-x-A-[FM]-G-[KR]-x-[IV]-[IV]-[HY]-x-G

Key Residues

PHE5, ILE6, GLY7, LEU8, GLY 9, GLY 12, ALA 16, ASN 18

Polymer:1

Type:polypeptide(L)

Length:299

Chains:A

Alternate: 3_HYDROXYISOBUT_DH 3-hydroxyisobutyrate dehydrogenase signature :

Functional Protein: PDB ID: 1VPD

Possible amino acid pattern found in1VPD

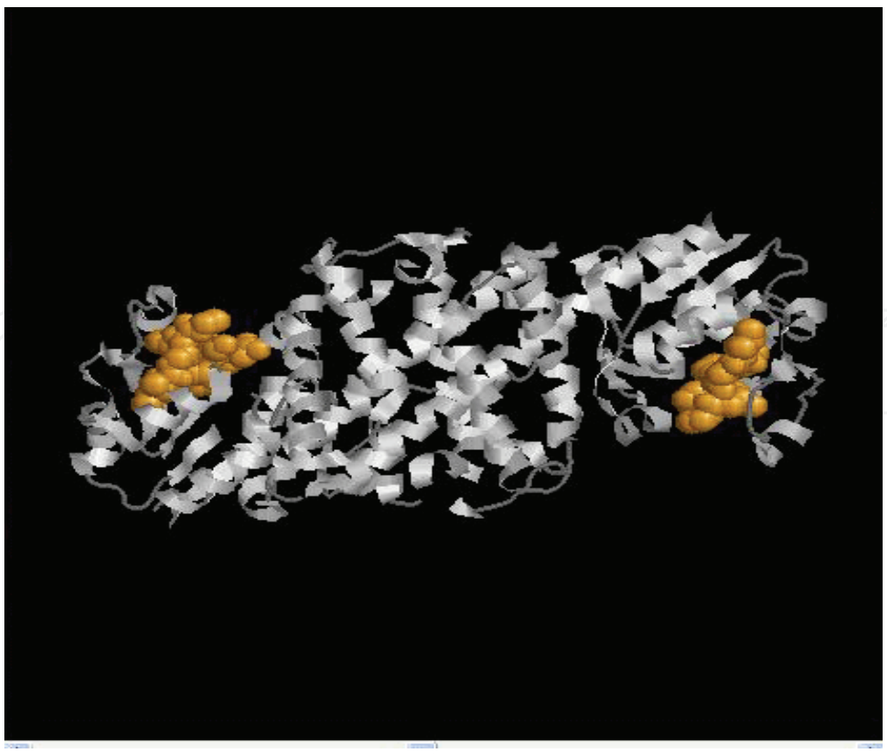
G-[ADET]-x-G-[AS]-G-x(1,2)-T-x(0,1)-K-L-[AT]-N-Q-[IV]-[IMV]-V-[AN]-x-[NT]-I-A-A-[MV]-[GS]-E-A-[FLM]-x-L-A-[AT]-[KR]-[AS]-[GV]-x-[ADNS]-[IP]

OR

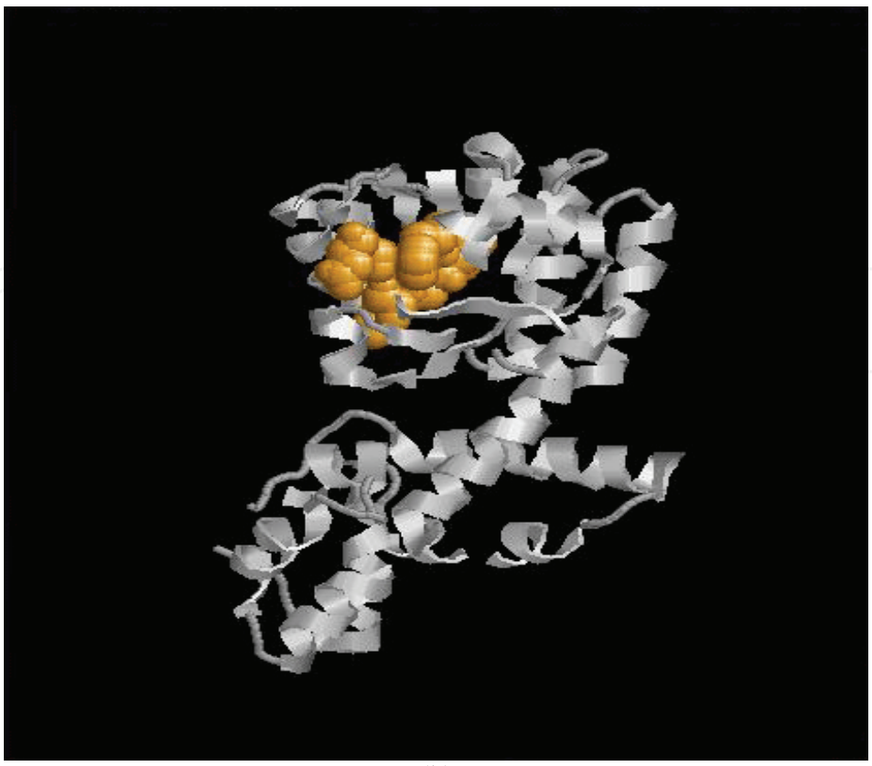
K-L-A-N-Q-x(0,1)-I-x(0,1)-V-[AN]-x-N-I-[AQ]-A-[MV]-S-E-[AS]-[FL]-x-L-A-x-K-A-G-[AIV]-[DENS]-[PV]-[DE]-x-[MV]-[FY]-x-A-I-[KR]-G-G-L-A-G-S-[AT]-V-[LM]-[DN]-A-K

Key Residues

PHE7, ILE8, GLY9, LEU10, GLY11, GLY14, SER18, ASN20



(a)



(b)

Fig. 4. a. Functional Protein Information (PDB Id: 1YB4) The residues in yellow colour represents the identified functional residues in Group 3. Also. b Functional Protein Information (PDB Id: 1VPD). The residues in yellow colour represents the identified functional residues in Group 3

3.1.4 Group 4

Enolase signature : Enolase, also known as phosphopyruvate dehydratase, is a metalloenzyme responsible for the catalysis of the conversion of 2-phosphoglycerate (2-PG) to phosphoenolpyruvate (PEP), the nineth and penultimate step of glycolysis. Enolase can also catalyze the reverse reaction, depending on environmental concentrations of substrates.

Polymer:1

Type:polypeptide(L)

Length:431

Chains:A, B, C, D

Functional Protein: PDB Id: 1E9I

ENOLASE Enolase signature: (Fig.5. a and Fig.5. b)

Possible amino acid pattern found in 1E9I

G-x(0,1)-D-D-[IL]-F-V-T-[NQ]-[PTV]-[DEKR]-x-[IL]-x(2)-G-[IL]-x(4)-[AGV]-N-[ACS]-[ILV]-L-[IL]-K-x-N-Q-[IV]-G-[ST]-[LV]-x-[DE]-[AST]-[FILM]-[ADES]-A-[AIV]-x(2)-[AS]-x(3)-[GN]

Key Residues

ILE 338, LEU339, ILE340, LYS341, ASN343, GLN344, ILE 345, GLY346, SER347, LEU348, THR349, GLU350, THR351

Alternate : ENOLASE Enolase signature

Polymer:1

Type:polypeptide(L)

Length:427

Chains:A, B

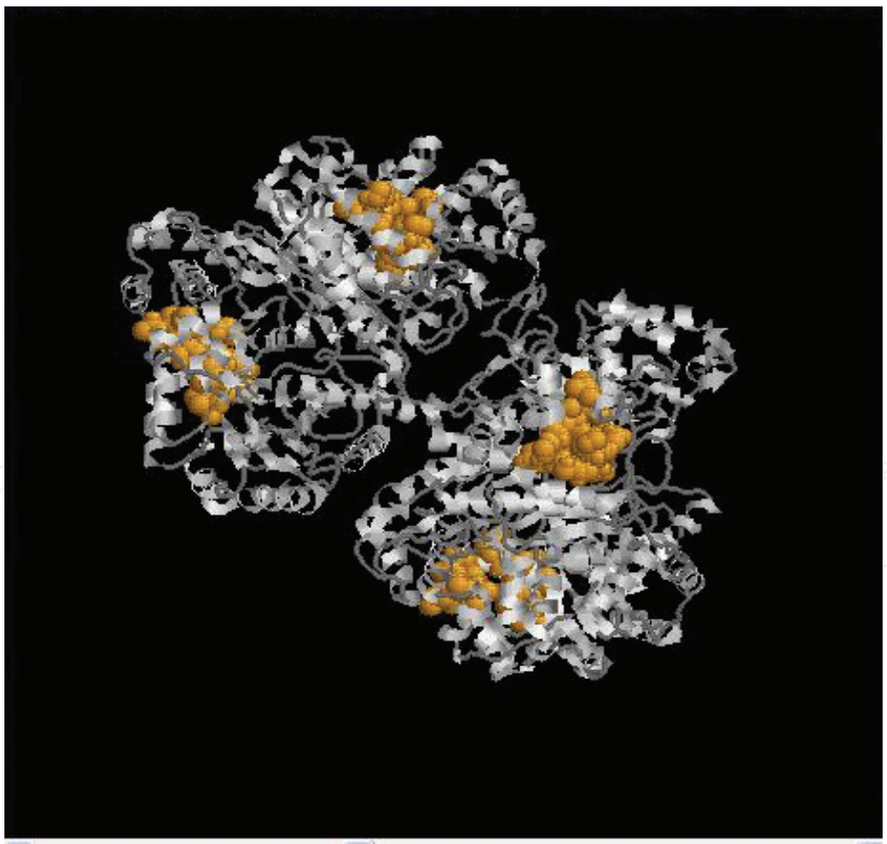
Functional Protein*:* PDB ID: 2PA6

Possible amino acid pattern found in 2PA6

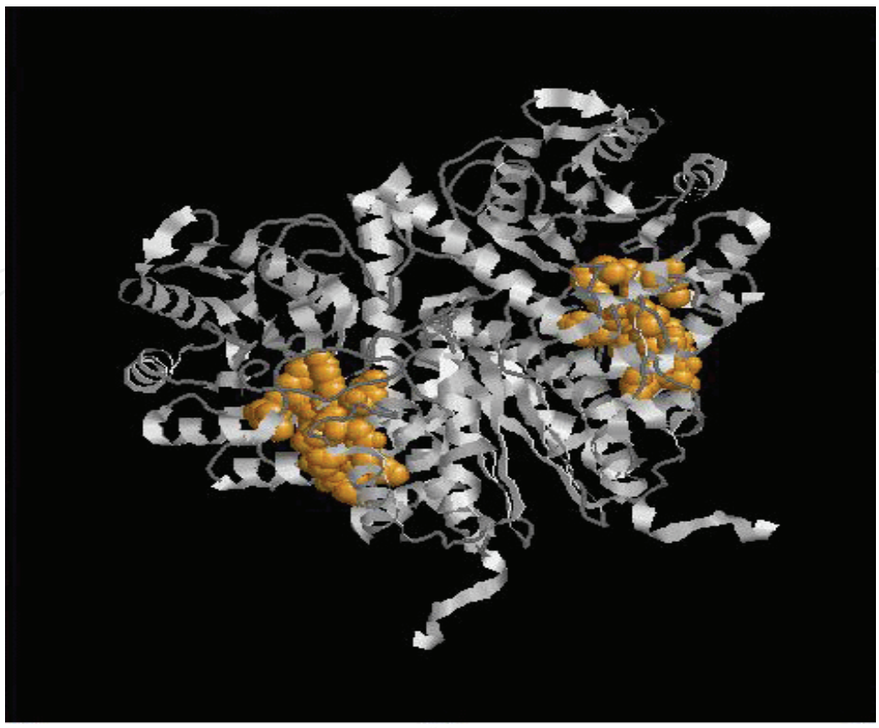
S-x(1,2)-S-G-[DE]-[ST]-E-[DG]-[APST]-x-I-A-D-[IL]-[AS]-V-[AG]-x-[AGNS]-[ACS]-G-x-I-K-T-G-[AS]-x-[AS]-R-[GS]-[DES]-R-[NTV]-A-K-Y-N-[QR]-L-[ILM]-[ER]-I-E-[EQ]-[ADE]-L-[AEGQ]

Key Residues

LEU 336, LEU337, LEU338, LYS339, ASN341, GLN342, ILE343, GLY344,THR345, LEU 346, SER347, GLU348, ALA 349



(a)



(b)

Fig. 5. a Functional Protein Information (PDB Id: 1E91) The residues in yellow colour represents the identified functional residues in Group 4. Also. b Functional Protein Information (PDB Id: 2PA6). The residues in yellow colour represents the identified functional residues in Group 4

3.1.5 Group 5

Glycerol-3-phosphate transporter (glpT) family of transporters signature :(Fig.6)
The major facilitator superfamily represents the largest group of secondary membrane transporters in the cell.

Molecule:Glycerol-3-phosphate transporter

Polymer:1

Type:polypeptide(L)

Length:451

Chains:A

Functional Protein: PDB ID: 1PW4

Possible amino acid pattern found in 1PW4

P-x(2,3)-R-x(0,1)-G-x-A-x-[AGS]-[FILV]-x(3)-[AGS]-x(3)-[AGS]-x(2)-[AILV]-x-[APST]-[IPV]-x(2)-[AG]-x-[ILV]-[ASTV]-x(3)-G-x(3)-[ILMV]-[FY]-x(3)-[AGV]-[AGILPV]-x-[GS]-[FILMV]

Key Residues

GLU153, ARG154, GLY155, SER159, VAL160, TRP161, ASN162, ALA164, ASN166, VAL167, GLY168, GLY169

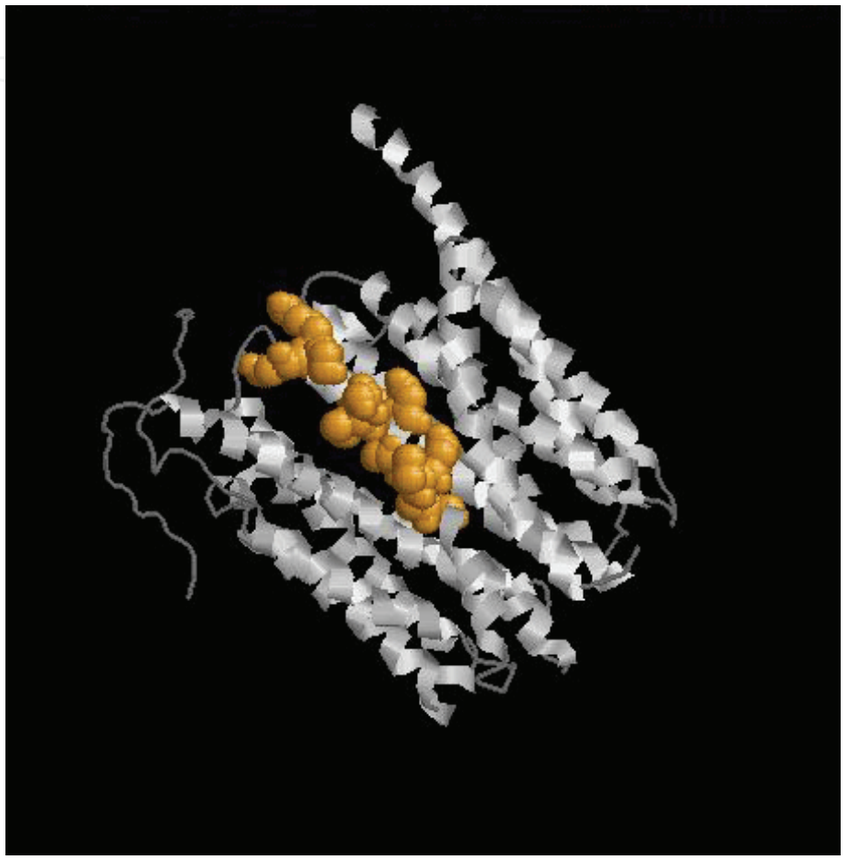


Fig. 6. Functional Protein Information (PDB Id: 1PW4). The residues in yellow colour represents the identified functional residues in Group 5

4. Conclusion

Identification of the specificity-determining residues in the various protein family studies has an important role in bioinformatics because it provides insight into the mechanisms by which nature achieves its astonishing functional diversity, but also because it enables the assignment of specific functions to uncharacterized proteins and family prediction. Genomics has posed the challenge of determination of protein function from sequence or 3-

D structure. Functional assignment from sequence relationships can be misleading, and structural similarity does not necessarily imply functional similarity. Our studies on the analysis of the superfamily revealed, for the first time, that in these species (archaea and bacteria) using *E. coli*. as a genomic model, we can contribute important insights for understanding their structural as well as functional relationships. The computational prediction of these functional sites for protein structures provides valuable clues for functional classification.

5. Acknowledgement

The authors gratefully acknowledge the support from JNIAS for the successful completion of this project.

6. References

Muhummad Khan and Kaiser Jamil (2008) Genomic distribution, expression and pathways of cancer metasignature genes through knowledge based data mining. International Journal of Cancer Research 1 (1), PP1-9, ISSN 1811-9727

Muhummad Khan and Kaiser Jamil (2008), Study on the conserved and polymorphic sites of MTHFR using bioinformatic approaches. Trends in Bioinformatics 1 (1) 7-17.

Sabitha Kotra, Kishore Kumar Madala and Kaiser Jamil (2008), Homology Models of the Mutated EGFR and their Response towards Quinazolin Analogues; J. Molecular Graphics and modeling , Vol-27, pp244-254.

Muhummadh Khan and Kaiser Jamil (2010) Phylogeny reconstruction of ubiquitin conjugating (E2) enzymes. Biology and Medicine Vol 2 (2), 10-19.

Patricia C, Babbitt, Miriam S. Hasson, Joseph E. Wedekind, David R. J. Palmer, William C. Barrett, George H. Reed, Ivan Rayment, Dagmar Ringe, George L. Kenyon, and John A. Gerlt (1996); The Enolase Superfamily: A General Strategy for Enzyme-Catalyzed Abstraction of the α-Protons of Carboxylic Acid *Biochemistry* 35 (51), pp 16489–16501

Babbitt PC Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids, Biochemistry. 35(51):16489-50

John F. Rakus, Alexander A. Fedorov, Elena V. Fedorov, Margaret E. Glasner, Brian K. Hubbard, Joseph D. Delli, Patricia C. Babbitt, Steven C. Almo and John A. Gerlt, (2008) Evolution of Enzymatic Activities in the Enolase Superfamily: l-Rhamnonate Dehydratase, *Biochemistry 47* (38), pp 9944–9954

Gerlt JA, Babbitt PC, Rayment I.(2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Arch Biochem Biophys. 1;433(1):59-7

Anurag Sethi, Patrick O'Donoghue, and Zaida Luthey-Schulten (2005) Evolutionary profiles from the QR factorization of multiple sequence alignments, PNAS vol. 102 no. 11 4045-4050

Lindahl E, Elofsson A, (2000) Identification of related proteins on family, superfamily and fold level. Journal of Molecular Biology 295: 3, 613-625

Steven E. Brenner, Cyrus Chothia, and Tim J. P. Hubbard (*1998*) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships PNAS May 26: 95 6073-6078

Dayhoff, M.O. (1974) Computer analysis of protein sequences, Fed. Proc. 33, 2314-2316,.

Scott McGinnis (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Research, Vol. 32

Hubbard BK, Koch M, Palmer DR, Babbitt PC, Gerlt JA. (1998) Evolution of enzymatic activities in the enolase superfamily: characterization of the (D)-glucarate/galactarate catabolic pathway in Escherichia coli. Biochemistry. 13;37(41):14369-75.

David R. J. Palmer, James B. Garrett,V. Sharma, R. Meganathan, Patricia C. Babbitt, and John A. Gerlt, (1999) Unexpected Divergence of Enzyme Function and Sequence: ''N-Acylamino Acid Racemase" Is o-Succinylbenzoate Synthase, Biochemistry, 38 (14), pp 4252–4258

Satu Kuorelahti Paula Jouhten, Hannu Maaheimo, Merja Penttila and Peter Richard (2006) l-galactonate dehydratase is part of the fungal path for d-galacturonic acid catabolism Molecular Microbiology 61:4 1060 – 1068

Brian K. Hubbard, Marjan Koch, David R. J. Palmer, Patricia C. Babbitt, and John A. Gerlt (1998) Evolution of Enzymatic Activities in the Enolase Superfamily: Characterization of the (D)-Glucarate/Galactarate Catabolic Pathway in Escherichia coli Biochemistry, 37 (41) 14369–14375

John F. Rakus, Alexander A. Fedorov, Elena V. Fedorov, Margaret E. Glasner, Brian K. Hubbard, Joseph D. Delli, Patricia C. Babbitt, Steven C. Almo and John A. Gerlt, (2008) Evolution of Enzymatic Activities in the Enolase Superfamily: l-Rhamnonate Dehydratase, Biochemistry, 47 (38), 9944–9954

Robert Belshaw and Aris Katzourakis (2005) Blast to Align: a program that uses blast to align problematic nucleotide sequences, Bioinformatics 21(1):122-123

Dmitry Lupyan, Alejandra Leo-Macias and Angel R. Ortiz (2005) A new progressive-iterative algorithm for multiple structure alignment Bioinformatics Volume 21:15 3255-3263

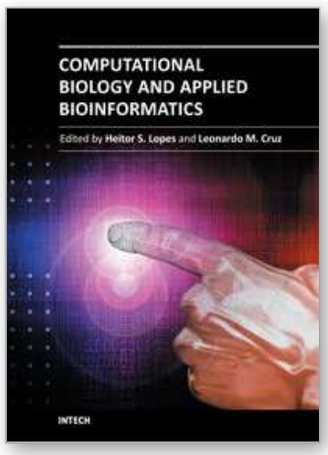
Doron Betel and Christopher WV Hogue, Kangaroo (2002) A pattern-matching program for biological sequences, BMC Bioinformatics, 1186/1471-2105-3-20

Ofir Goldenberg, Elana Erez, Guy Nimrod, and Nir Ben-Tal (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures Nucleic Acids Res. D323–D327.

Changwon Keum and Dongsup Kim (2006) Protein function prediction via ligand interface residue match, World Congress on Medical Physics and Biomedical Engineering 2006, August 27 – September 1, COEX Seoul, Korea ''Imaging the Future Medicine"

LevelErik Lindahl and Arne Elofsson, (2000) Identification of Related Proteins on Family, Superfamily and Fold Journal of MolecularBiology 295: 3, 613-625

Neidhart DJ, Kenyon GL, Gerlt JA, Petsko GA (1990) Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. Nature. 347(6294):692-4.



Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4
Hard cover, 442 pages
Publisher InTech
Published online 02, September, 2011
Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen