In-silico approaches to sequence and structure based scrutiny of nonsynonymous SNPs and synteny of *ACAA2* for its implicated role in metabolomics

Sidrah Anjum[1], Attya Bhatti[1] and Asma Gul[2]

[1] National University of Sciences and Technology, Kashmir Highway H-12, Islamabad
[2] Department of Bioinformatics and Biotechnology, International Islamic University, Islamabad H-10 Islamabad, Pakistan

E-mail: sidrah.bsbi269@iiu.edu.pk , gulasma@iiu.edu.pk*

Abstract. The science of metabolomics is exponentially growing with the boost in knowledge base of metabolites and development of bioinformatics data analysis tools. Human ACAA2 gene codes for a 397 amino acid protein, 3-Ketoacyl-CoA thiolase, which is a mitochondrial enzyme responsible for catalyzing the thiolytic cleavage of 3-ketoacyl-CoA by thiolase, the last reaction of fatty-acid β-oxidation. Bioinformatics approaches were used to scrutinize the human ACAA2 protein. The tertiary structure of the protein was predicted using a multitude of structure prediction servers including PHYRE2, (PS)-v2, I-TASSER and SAM-T08. Generated structures were then evaluated using ERRAT2, VERIFY3D, PROVE, PROCHECK and MoLProbity servers. PolyPhen-2 and SIFT servers were used to sort the damaging missense SNPs from the tolerable ones. The structural model of mutated sequence was also generated and superimposed on the normal structure using UCSF Chimera in order to find the regions of deviations. PoPMuSiC was used to analyze thermodynamic stability changes in mutated structure. Finally, syntenic analysis was performed using Cinteny in order to study the physical co-localization of the genetic loci with the ACAA2 gene. The model generated by PHYRE2 (normal mode) was found to be the best model so far which was selected for further studies. As a result of mutational analysis, substitution of D31Y caused by SNP rs11549282 was found to be damaging. Genomic context of Human and mouse ACAA2 was found to be the same up to 36 genes as predicted by Cinteny. This work will be helpful in detailed comprehension of the role of ACAA2 in fatty acid metabolism related disorders.

1 Background

ACAA2 gene encodes for human mitochondrial 3-oxoacyl-CoA thiolase which catalyzes the last step of fatty acid β-oxidation (Maher et al., 2009). The enzyme constitutes an amino acid sequence of 397 residues (Abe et al., 1993). Acaa2 gene expression profiles in mice with Diet-Induced Obesity were studied and 2-fold decrease in the levels of the gene in obese mice was observed (Soh et al., 2011). *ACAA2* gene can therefore be implicated to play a role in metabolomics because of its role in fatty acid metabolism. However, further research in this realm is required. The endeavor of the present study was to scrutinize the human *ACAA2* gene using the amino acid sequence of its protein product. Study entailed three modules which include tertiary structure prediction, mutational/SNP analysis and syntenic analysis.

2 Methodology

A multitude of web-servers based on modified approaches like iterative threading, comparative modeling, ab-initio contact prediction using neural networks, and profile-profile matching algorithms for the structure prediction were used for the prediction of tertiary structure of our protein in question. Complete amino acid sequence of *ACAA2* consisting of 397 residues was retrieved from NCBI in Fasta format (accession: NP_006102). A total of four structural models for ACAA2 protein were generated using SAM-T08, PHYRE2, (PS)2-v2 and I-TASSER. The models generated through various 3D structure prediction tools were evaluated to check their accuracy so as to select the best model. Different criteria determined the quality of these models. The criteria used for model assessment were Ramachandran plots, quality factor, Z-score, percentage of residues having average 3D-1D score greater than 2 etc. All the protein structural models were analyzed for these criteria. PROVE evaluated the models in terms of z-score. Zscore is a standard score that indicates how many standard deviations an observation or data is above or below the mean. Z-score of a model should be less than 1 (Laskowski et al, 1993). PROCHECK evaluated the models by giving the percentage of residues falling in most favored regions, allowed regions and generously allowed regions. ERRAT indicated overall quality factor of protein structure (Colovos, 1993) in the form of a plot. VERIFY3D provided percentage of residues having an average 3D-1D score > 2 (Bowie et al,1991). MoLProbity server displays the H-bond and van der waals contacts in the interfaces between components and detailed allatom contact analysis of any steric problems within the molecules (Davis et al., 2007). A total of 26 non-synonymous SNPs of *ACAA2* were retrieved from dbSNP. The retrieved SNPs were further screened for sorting of intolerant from tolerant SNPs using SIFT server. *ACAA2* protein with D31Y substitution was modeled and compared to the native *ACAA2* protein for investigating their structural variations. PoPMuSiC analyzed how the mutation affected the structure of the protein in

terms of its energy state i.e. thermodynamic stability. Syntenic blocks refer to chromosomal segments in two species where syntenic anchors are in consecutive order (Mural et al., 2002). To study the physical co-localization of the genetic loci with the *ACAA2* gene, syntenic analysis was performed using Cinteny Server.

3 Results and Discussion

3.1 Structural modeling and Model Evaluation

The evaluation results obtained from PROCHECK, ERRAT2, VERIFY3D and PROVE are summarized in table 1. Tables 2 and 3 include the evaluation results from MoLProbity server. The evaluation results indicate that the best structural model for ACAA2 was generated by PHYRE2 as the MoLProbity server provided the best results of ramachandran plots analysis for this model (See Table 2 and 3). It showed that 100% residues of PHYRE2-generated model lie in allowed regions of the ramachandran plot with highest percentage of residues lying in the favored regions (98.4%) Moreover, it gave the lowest percentage (0.33%) of poor rotamers with 0% Ramachandran outliers, and zero percent residues with bad angles and bad bonds. It has 0.00 C β deviations >0.25Å. Cβ deviation measures a particularly significant kind of bond angle distortion in proteins. A large Cβ deviation (>0.25 Å) often signals an incompatibility between the side chain and main chain conformation; for instance, a side chain fit 180° backward (Davis et al., 2007). PROCHECK also gave highest percentage of residues lying in the core regions of the ramachandran plot for PHYRE2-generated model. ERRAT2 gave a 67.905 overall quality score. VERIFY3D results indicated that 85.46 % of residues had an average 3D-ID score greater than 2. Z-score provided by PROVE for this model was 0.296. PHYRE2-generated model constitutes 19 helices and 10 β-sheets. 49% of the ACAA2 amino acid residues make up alpha helices, 20% make up the β-sheets and the rest of 31% residues make up the loops. PHYRE2 uses protein threading approach which has been recently reported to be one of the most effective and accurate protein structural modeling techniques (Shao et al., 2011).

3.2 Mutational analysis

Two SNPs (rs34783635 and rs11549282), were categorized as intolerant on the basis of their tolerance index <0.05. This implies that these two SNPs have the potential to change the structure of the ACAA2 protein thereby affecting its function. Furthermore, nine SNPs were predicted to be 'damaging' by Polyphen-2 server on the basis of their PSIC values. Analysis of the results provided by Polyphen-2 and SIFT servers, indicate that out of the screened SNPs, rs11549282 was categorized as intolerant by SIFT (tolerance index 0) as well as damaging by PolyPhen-2 server (PSIC value

0.997). SNP rs11549282 causes substitution of D→Y at position 31. Modeling of mutated structure: Figure 2 shows the mutated structure of ACAA2 predicted by PHYRE2. Analyzing the effect of mutations on structure: The thermodynamic stability value for the mutated protein was predicted to be 0.79 kcal/mol. The high thermodynamic stability value indicates that the mutant protein structure is not in its lowest energy state and hence, it is not stable (Tokuriki et al., 2008). This implies that D31Y mutation in the native ACAA2 leads to its destabilization. Results of PoPMuSiC are shown in figure 3.

3.3 Structural Comparison of Normal and Mutated Protein

Secondary structure analysis

In order to analyze the effect of D31Y mutation on ACAA2 secondary structure elements, the secondary structures of wild and mutated protein ACAA2 were predicted through PHYRE2. Results indicate that the substitution D31Y (rs11549282) caused an increase in number of α- helices from 19 to 20 while number of β-strands remained the same i.e. 10. 14 out of 19 α- helices and 7 out of 10 β-strands observed in the native protein remained conserved after substitution. Changes were observed in the rest of the helices and strands. Summary of the affected residue positions constituting secondary structure elements (helices and strands) of wild and D31Y (rs11549282) mutated ACAA2 structure is given in table 5.

Tertiary structure analysis

3D structure is more conserved than sequence (Capriotti and Marti-Renom, 2010), thus native and mutant (D31Y) structures were superimposed in three dimensions using UCSF Chimera which actually produced a measure to assess the level of similarity of the aligned structures. The resultant superimposed structures are shown in fig. 6 where regions of deviations are encircled. SDM value for superimposition was 16.524. SDM is 0 for identical structures and its value increases as the similarity between the structures decreases (Krissinel et al., 2004). This implies that the mutant and wild structure were not exactly alike. Q-score values range from 0 to 1. For completely dissimilar structures its value is 0 and1 for identical structures (Johnson et al., 1990). Q-score value for our superimposed structures is 0.915 respectively. RMSD value of 0.831Å was observed for D31Y (rs11549282). RMSD value lying between 0 and 1.5Å indicates significantly similar structures while a greater RMSD value implies increase in structural differences (Waheed et al., 2012). The observed RMSD values mean significant structural variations, reflecting significant functional variations have similar sequences. These results show that the wild and mutated structures are significantly similar but even the subtle changes in wild structure have led to the destabilization of ACAA2 protein.

3.4 Syntenic Analysis.

Synteny is a valid deduction that two or more genomic regions are derived from a single ancestral genomic region as all the genes within a syntenic block are likely to be orthologous with a preserved gene order (Lane et al., 2001). Results of Cinteny server indicate that the bidirectional genomic context of human ACAA2 gene is the same as that of the mouse gene up to 36 genes. This implies that these 36 genes belong to the same 'syntenic block'. It is an indication that mice and humans are sharing a highly conserved gene sequences and that this region in mice and humans has derived from a common ancestor as a causatum of a genome duplication event (Ansari-Lari et al., 1998).

4 Conclusion

The functional characterization of a protein sequence is one of the most frequent problems in molecular biology. Knowledge of protein structure provides an insight into its interactions (Aydin et al., 2011), which define the protein's biological role and functions (Cheng et al., 2005). The predicted 3D structure of the protein will provide valuable insights into its interactions. The alternations of key residues in a protein cause loss of its normal biological function SNPs cause changes in the secondary structure elements and physiochemical properties thus it was inferred that these changes may be translated into the tertiary structure (Waheed et al., 2012). Damaging SNP causing D31Y mutation in ACAA2 leads to the addition or the deletion of one or more alpha helices or beta sheets thereby destabilizing the protein tertiary structure and its associated function . Hence it can be implicated to play a role in metabolic disorders such as obesity in humans as well. Moreover, our study shows that bidirectional genomic context of human ACAA2 gene is the same as that of the mouse gene up to 36 gene. This syntenic region in mouse will be helpful in gaining insights into the evolution of coding and noncoding segments of this region in humans and will serve as a first step in elucidation of the regulatory elements of this conserved genomic region. The syntenic region in mouse chromosome 18 can be isolated and sequenced for further studies which would save researchers from extensive experimentation on humans which would otherwise take years and years. Moreover, the syntenic block can be helpful for additional studies such as genome evolution, ancestral genome reconstruction, gene family evolution and origin of gene family duplicates. The analysis of synteny in the gene order sense can also be performed which has numerous other applications in genomics.

Fig. 1. Wild type ACAA2 structure

Fig. 2. Mutated ACAA2 structure

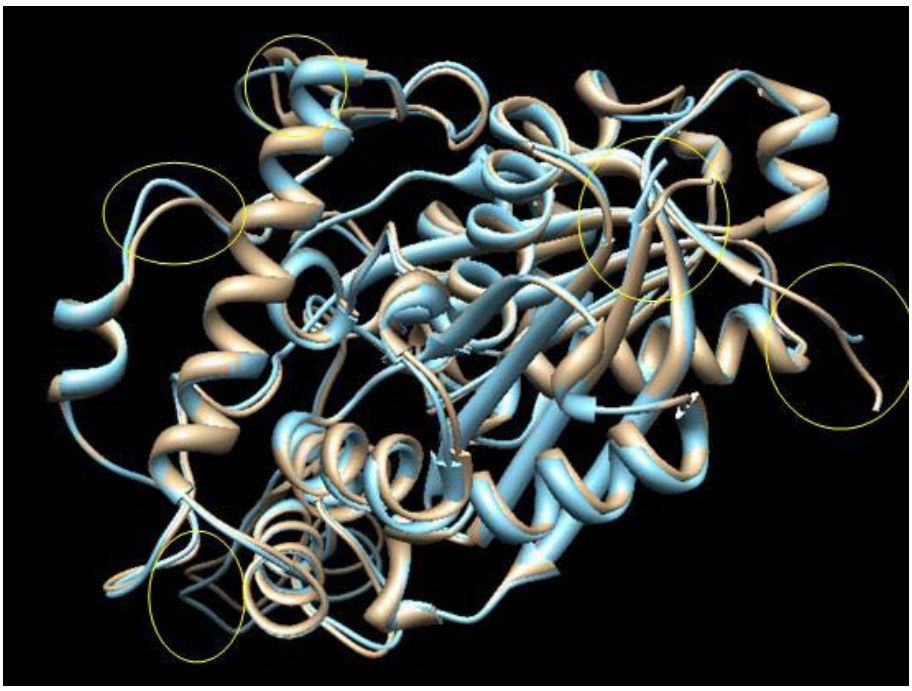

Fig. 3. Wild and mutated ACAA2 superimposed structures.

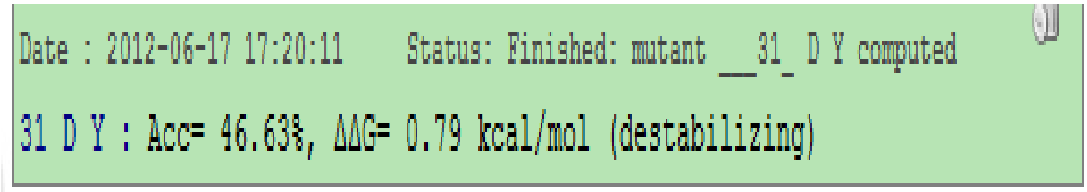

Fig. 4. PoPMuSiC results indicating the stability, Solvent accessibility and the thermodynamic stability of the mutated ACAA2 protein

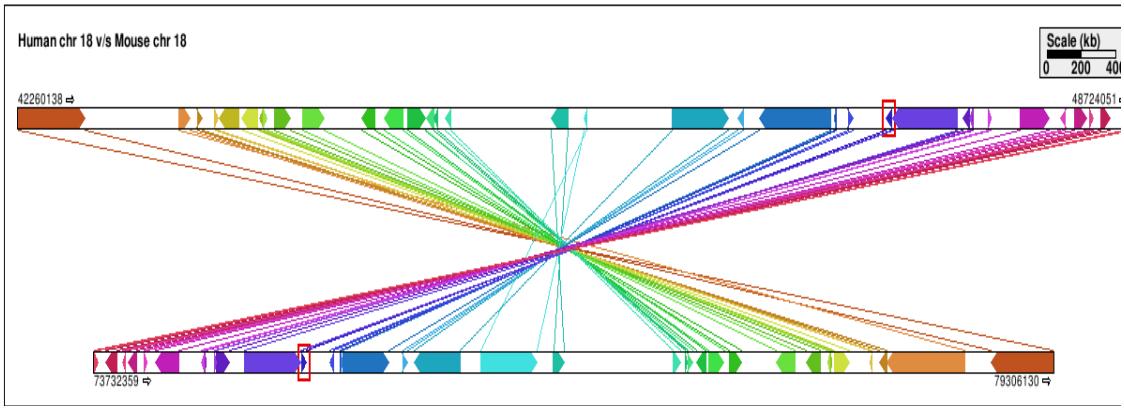

Fig. 5. The synteny blocks in human chromosome 18 (top) and mouse chromosome 18 (bottom) which contain the ACAA2 gene with adjust parameters of 70kb of minimum length of synteny block and 700kb of maximum gap between adjacent marker.

Table 1. Summary of Evaluation results of structural models obtained from PROCHECK, ERRAT2, VERIFY3D and PROVE.

Models	PROCHECK				ERRAT2	VERIFY3D	PROVE
	Core (%)	Allowed (%)	Generously (%)	Disallowed (%)			
I-TASSER	88.8	8.09	0.69	2.7	48.586	86.93	-0.096
$(PS)^2$-v2	90.2	8.3	0.9	0.6	69.657	83.67	0.031
SAM-T08	92.2	6	1.2	0.6	77.604	89.70	0.384
PHYRE	92.4	6.9	0.3	0.3	67.905	85.46	0.296

Table 2. Summary of evaluation results of structural models obtained from MoLProbity server.

Models	Poor rotamers %	Ramachandran outliers %		Cβ deviations >0.25Å	Residues with bad bonds%	Residues with bad angles%
		Outliers	Favored			
I-TASSER	4.14	1.01	92.41	16	0.00	4.79
$(PS)^2$-v2	2.89	0.51	95.95	1	0.00	1.26
SAM-T08	1.29	1.28	97.44	17	0.50	1.01
PHYRE2	0.33	0.00	98.45	0.00	0	0

Table 3. Summary of evaluation results obtained from MoLProbity Server.

Models	MoLProbity Ramachandran plot analysis	
	Favored regions (%)	Allowed regions (%)
I-TASSER	85.5	94.9
$(PS)^2$-v2	89.4	96.1
SAM-T08	97.4	98.7
PHYRE2	97.2	98.7

Table 4. Summary of affected residue positions constituting secondary structure elements of wild and mutated ACAA2.

Secondary Structure element	Residues positions in normal structure	Residue positions in structure with D31Y substitution
Beta strand	53-57	53-58
Alpha Helix	145-148	144-148
Beta strand	203-208	205-208
Alpha Helix	219-221	219-222
Alpha Helix	-------	247-249
Beta strand	275-285	275-286
Alpha Helix	295-306	296-306
Alpha Helix	355-372	354-372
Beta strand	376-379	376-378
Alpha Helix	380-381	379-382

Table 5. Genomic context of human ACAA2 and its syntenic region in mouse.

Gene Name	Chromosomal Location			
	Human 18		Mouse 18	
SETBP1	42260138	42648475	78947117	79306130
SLC14A2	43194766	43263072	78342883	78793689
SLC14A1	43304092	43332485	78296830	78338858

SIGLEC15	43405545	43422521	78240353	78254007
KIAA1632	43427574	43547305	----------	
PSTPIP2	43563502	43652250	78033289	78121618
ATP5A1	43664110	43684199	78012507	78021608
HAUS1	43684298	43708299	77996306	78006519
C18orf25/8030462N17Rik	43753988	43846955	77872020	77952749
RNF165	43914187	44040783	77694849	77803875
ST8SIA5	44259081	44337039	77424586	77494189
PIAS2	44392060	44497466	77303947	77394447
KATNAL2	44526787	44628614	77231939	77286035
HDHD2	44633781	44676871	77182854	77210910
IER3IP1	44681413	44702745	77168766	77180353
FUSSEL18/652991	44746293	44775554	77095144	77139081
SMAD2	45359466	45457515	76401579	76465385
ZBTB7C	45553733	45567494	75979832	76308218
KIAA0427/9811	46065427	46389588	75590859	75857350
SMAD7	46446223	46477081	75527019	75555588
DYM	46570172	46987079	75178426	75446620
C18orf32/497661	47008030	47013601	75165554	75169587
RPL17	47014854	47018906	75160131	75163035
LIPG	47088427	47119278	75098976	75120917
ACAA2	47309874	47340251	74938866	74965861
MYO5B	47349156	47721451	74602273	74931131
CCDC11	47753563	47792865	74442754	74519638

References

1. Abe, H. (1993). Cloning and sequence analysis of a full length cDNA encoding human mitochondrial 3-oxoacyl-CoA thiolase. Biochim. Biophys. Acta., 1216:304-306.

2. Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, Lu J, Gorrell JH,

3. Chinault AC, Belmont JW, Miller W, Gibbs RA. Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. Genome Res. 1998;8:29–40

4. Aydin, Z., Singh, A., & Bilmes, J. (2011). Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure. BMC Bioinformatics, 12:154 .
5. Bowie, J.U., Luthy, R. and Eisenberg, D. (1991), A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253: 164-170.

6. Capriotti, E., & Marti-Renom, M. A. (2011). Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. BMC Bioinformatics, 11:322 .

7. Cheng, J., Randall, A.Z., Sweredoski, M.J., Baldi, P. (2005), SCRATCH: a protein structure and structural feature prediction server. Nucl. Acid. Res., 33: 72-76.

8. Colovos, C. and Yeates, S.O. (1993), Verification of protein structures: patterns of nonbonded atomic interactions. Protein Science, 2(9): 1511-1519.

9. Davis, I.W. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucl. Acids Res., 35: 375-383.

10. Johnson, M. S., Sutcliffe, M. J., & Blundell , T. L. (1990). Molecular anatomy: Phyletic relationships derived from three-dimensional structures of proteins. Journal of Molecular Evolution , 30(1), 43-59.

11. Lane, R.P., Cutforth, T., Young, J., Athanasiou, M., Friedman, C., Rowen, L., Evans, G., Axel, R., Hood, L. and Trask, B.J. (2001), Genomic analysis of ortholgous mouse and human olfactory receptor loci. Proc. Natl. Acad. Sci. USA., 98: 7390-7395.

12. Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993), PROCHECK: a program to check the stereochemical quality of protein structures. J.Appl.Cryst., 26: 283-291.

13. Maher, A.C., Fu, M.H., Isfort, R.J., Varbanov, A.R., Qu, X.A.. (2009). Sex Differences in Global mRNA Content of Human Skeletal Muscle. PLoS ONE, 4(7).

14. Mural RA, MD Adams, EW Myers et al, (2002). A comparison of whole-genome shotgun-derived mouse chromosome 15 and the human genome. Science, 296 (1661-1971)

15. Shao, M., Wang, S., & Wang, C. (2011). Incorporating Ab Initio energy into threading approaches for protein structure prediction. BMC Bioinformatics, 12(Suppl 1):S54.

16. Soh, J., Kwon, D. Y. and C, YS. (2011), Hepatic Gene Expression Profiles Are Altered by Dietary Unsalted Korean Fermented Soybean (Chongkukjang) Consumption in Mice with Diet-Induced Obesity. Journal of Nutrition and Metabolism.

17. Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How Protein Stability and New Functions Trade Off. PLoS Computional Biology 4(2)

18. Waheed, R., Khan, M. H., Bano, R., & Rashid, H. (2012). Sequence and structure based assessment of nonsynonymous SNPs in hypertrichosis universalis. Bioinformation. 8(7): 316–318.