Academic Sciences International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 4, Issue 3, 2012

Research Article

PREDICTION OF DELETERIOUS NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS (nsSNPs) OF GALC GENE BY COMPUTATIONAL METHOD

M. MADHUMATHI AND V. SHANTHI*

School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India. Email: shanthi.v@vit.ac.in

Received: 28 Feb 2012, Revised and Accepted: 13 April 2012

ABSTRACT

Single Nucleotide Polymorphisms (SNPs) are the most abundant sequence variations encountered in a genome and play a major role in understanding of the genetic basis of many complex human diseases. The genetics of human phenotype variation could be understood by knowing the functions of these SNPs. It is still a major challenge to identify the functional SNPs in a disease-related gene. In this work, we have analyzed the genetic variation that can alter the expression and function of GALC gene using computational methods. Of the total 1615 SNPs, 12 were found to be nonsynonymous SNPs (nsSNPs). Among these 12 nsSNPs, 3 are deleterious analyzed by SIFT program. 4 nsSNPs are damaged found by PolyPhen. 3 nsSNPs that were observed to be deleterious and damaged by using both SIFT and PolyPhen program. From this a comparison of stabilizing residues are done by RMSD of native and mutant proteins, higher the RMSD value is the more deviation between two structure, so we propose that an nsSNP rs73312829 with a mutation of Arginine to Cystine at position 79 could be the main target mutation for the Krabbe disease.

Keywords: SIFT, PolyPhen, RMSD, GALC gene, Krabbe disease, Modeled structure.

INTRODUCTION

In human genome, single base change, called single nucleotide polymorphism (SNP) and is the most frequent type of genetic variation [1, 2]. Up to March 25, 2010, a total of 23,653,737 SNPs in human have been identified and deposited in the NCBI dbSNP. When SNPs occur in coding regions and cause amino acid change in corresponding proteins called nonsynonymous single nucleotide polymorphisms (nsSNPs) [3].These nsSNPs occurring in protein coding region and affecting protein functions and causing common diseases [4]. nsSNPs affect gene regulation by altering DNA and transcriptional binding factors and maintenance of the structural integrity of cells and tissues [5, 6] and also nsSNPs affect the proteins functional role in signal transduction of hormonal, visual and other stimulants [7, 8]. Half of all genetic changes related to human diseases are attributed to nsSNP variants [9, 10]. SNPs report for the more common form of human genetic deviation. About 500,000 SNPs fall in the coding regions of the human genome [11]. Recent studies about distinguishing disease-causing amino acid substitutions from neutral nsSNPs in human proteins generally focus on predicting numeric scores of the mutations (indicating the likelihood of causing diseases) [12, 13].

Krabbe disease is autosomal recessive, rare, often fatal disorder that affects the myelin sheath of the nervous system. Symptoms begin at the age of 3 and 6 months with fevers, irritability, vomiting, limb stiffness, seizures, feeding difficulties and slowing of mental and motor development [14]. Mutations in the GALC gene located on chromosome 14 (14q31) is greatly increase the risk of Krabbe disease which causes a deficiency of an enzyme called galactocerebrosidase [15]. The gene of nearly 60 kb, consist of 17 exons and 16 introns [16, 17]. At present most mutations in GALC gene have been identified to be point mutations or small insertions and deletions [18-20]. Even though our journalism review showed that there is a wide option of literature on the GALC gene related to Krabbe disease, there have been no computational studies undertaken for an in silico investigation of the nsSNP mutations in GALC. We undertook this work mainly to perform a computational analysis of the nsSNPs in the GALC gene, to see the possible mutations and offer a modeled structure for the mutant protein.

MATERIALS AND METHODS

Sequence data sets and polymorphism identification

The SNPs and their associated protein sequence for GALC gene were obtained from dbSNP [21] (http://www.ncbi.nlm.nih.gov/SNP/) for our computational analysis.

Analysis of functional consequences of coding nsSNPs by sequence-homology-based method (SIFT)

We used the program SIFT [22] available at http://sift.jcvi.org/www /SIFT_dbSNP.html to detect the deleterious coding nsSNPs. SIFT is a sequence-homology-based tool that presumes that important amino acids will be conserved in the protein family. Hence, changes at well-conserved positions tend to be predicted as deleterious. We submitted the query in the form of SNP IDs or as protein sequences. The underlying principle of this program is that SIFT takes a query sequence and uses multiple alignment information to predict tolerated and deleterious substitutions for every position of the query sequence. SIFT is a multistep procedure that, given a protein sequence, (a) searches for similar sequences, (b) chooses closely related sequences that may share similar functions, (c) obtains the multiple alignment of the chosen sequences, and (d) calculates normalized probabilities for all possible substitutions at each position from the alignment. Substitutions at each position which normalized probabilities less than a chosen cutoff are predicted to be deleterious and those greater than or equal to the cutoff are predicted to be tolerated [23]. The cutoff value in the SIFT program is a tolerance index of ≥ 0.05. The higher the tolerance index, the less function impact a particular amino acid substitution is likely to have.

Simulation for functional change in coding nsSNPs by structure-homology-based method (PolyPhen)

Analyzing the damaged coding nsSNPs at the structural level is considered to be very important to understand the functional activity of the protein of concern; we used the server PolyPhen [24], which is available at http://coot.embl.de/PolyPhen/. Input options for the PolyPhen server are protein sequence or SWALL database ID or accession number together with sequence position with two amino acid variants. We submitted the query in the form of protein sequence with mutational position and two amino acid variants. Sequence-based characterization of the substitution site, profile analysis of homologous sequences, and mapping of substitution site to a known protein three-dimensional structure are the parameters taken into account by the PolyPhen server to calculate the score. It calculates PSIC scores for each of the two variants and then computes the PSIC score difference between them. The higher the PSIC score difference is the higher is the functional impact a particular amino acid substitution is likely to have.

Modeling nsSNP locations on protein structure and their RMSD difference

Structure analysis was performed for evaluating the structural stability of native and mutant protein. We used the Web resource SAAPdb [25] to

identify the protein related to GALC gene and also confirmed the mutation positions and the mutation residues from this server. The mutation was performed by using the SWISSPDB. The deviation between the four structures is evaluated by their RMSD values.

RESULTS

SNP dataset from dbSNP

The GALC gene investigated in this work was retrieved from the dbSNP database [21]. It contain total of 1615 SNPs. These SNPs are submitted to the SIFT and PolyPhen server to detect the deleterious coding nsSNPs and to see the functional change in coding nsSNPs.

Deleterious nsSNP found by SIFT

The conservation level of particular position in a protein was determined by using a sequence homology-based tool, SIFT [22]. The protein sequences of 12 nsSNPs were submitted independently to the SIFT program to check its tolerance index. The higher the tolerance index, the less functional impact a particular amino acid substitution is likely to have, and vice versa. Among the 12 nsSNPs, 3 were found to be deleterious, having a tolerance index score of ≤ 0.05. The results are shown in Table 1.

Table 1: List of nsSNPs that were predicted to have functional significance by SIFT

SNP ID	Nucleotide change	Amino acid change	Tolerance index
rs115869593	G/T	T177N	0.27
rs111887056	C/G	A21P	0.31
rs78774548	G/T	P75Q	0.08
rs74887188	C/T	I305V	0.31
rs73312829	A/G	R79C	0.00
rs34362748	C/T	D248N	0.19
rs34134328	C/G	T468S	0.1
rs17687109	A/G	L400P	0.13
rs1805078	C/T	R184C	0.02
rs421262	C/T	T641A	0.67
rs398607	A/G	I562T	0.07
rs11623	G/T	G57C	0.00

We observed that, of 3 deleterious nsSNPs, 2 showed a highly deleterious tolerance index score of 0.00 and 1 showed a tolerance index score of 0.02. Three nsSNPs showed a nucleotide change of G□T, two nsSNPs showed a change of C□G, 4 nsSNPs C□T, 3 nsSNPs A□G. C□T nucleotide changes occurred the maximum number of time and C□G nucleotide changes occurred in a minimum number of times, as can be seen from Table 1. The nucleotide change A□G and G□T accounted for the high number of deleterious nsSNPs, with a SIFT tolerance index of 0.00. This was closely followed by the nucleotide change C□T, which showed a tolerance index of 0.02.

Damaged nsSNP found by the PolyPhen

The structural levels of alteration were determined by applying the PolyPhen program [24]. Twelve protein sequences of nsSNPs investigated in this work were submitted as input to the PolyPhen server and the results are shown in Table 2.

Table 2: List of nsSNPs that were predicted to be functionally significant by PolyPhen

SNP ID	Nucleotide change	Amino acid change	PSIC SD
rs73312829	A/G	R79C	2.792
rs1805078	C/T	R184C	1.639
rs398607	A/G	I562T	1.677
rs11623	G/T	G57C	2.592

A position-specific independent count (PSIC) score difference of 1.1 and above is considered to be damaging. It can be seen that, of 12 nsSNPs, 4 were considered to be damaging. All 4 nsSNPs exhibited a PSIC score difference in the range 1.639 to 2.792. Three nsSNPs that were observed to be deleterious by the SIFT program also were damaging according to PolyPhen. Hence, we could infer that the results obtained on the basis of sequence details (SIFT) were in good correlation with the results obtained for structural details (PolyPhen), as can be seen from Tables 1 and 2. It can be seen from Tables 1 and 2 that 2 nsSNPs (rs73312829 and rs11623) had a SIFT tolerance index of 0.00 and PSIC score difference ≥2.00. Hence the mutations occurring with these 2 nsSNPs would be of prime importance in the identification of Krabbe disease caused by the GALC gene, according to SIFT and PolyPhen results.

Modeling of mutant structure

Mapping the deleterious nsSNPs into protein structure information was performed through the Single Amino Acid Polymorphism database (SAAPdb) [25]. The available structure of galactocerebrosidase from mouse has the PDB ID 3ZR5. According to this resource, we modeled the human protein by SWISS model. The mutation occurred in human protein is in 4 SNP Ids, namely, rs73312829, rs1805078, rs398607 and rs11623. The mutations were at residue position 79 (R□C), 184 (R□C), 562 (I□T) and 57 (G□C). The mutations were at position 79, 184,562 and 57 were performed by SWISSPDB viewer independently to get 4 modeled structures. Table 3 also shows that the RMSD values between the native type and the mutant type for R79C - 1.57 Å, R184C - 1.27 Å, I562T – 1.29 Å and G57C – 1.40 Å.

Table 3: RMSD native-structure and mutant models

Parameter	R79C mutant (rs73312829) With native type GALC	R184C mutant (rs1805078) With native type GALC	I562T mutant (rs398607) With native type GALC	G57C mutant (rs11623) With native type GALC
RMSD of entire structure	1.57 Å	1.27 Å	1.29 Å	1.40 Å

The higher the RMSD value is more the deviation between the two structures is, which in turn changes their functional activity. Since the RMSD values are higher for 4 mutant type structures compared to the native type structure, these 4 nsSNPs could be believed to affect the structure of the proteins. 3 nsSNPs were also shown to be deleterious according to the SIFT program and 4 of the nsSNPs was shown to be damaging according to the Poly-Phen server. The superimposed structures of native protein with 4 mutant-type proteins (R79C, R184C, I562T and G57C) are shown in Figure 1.

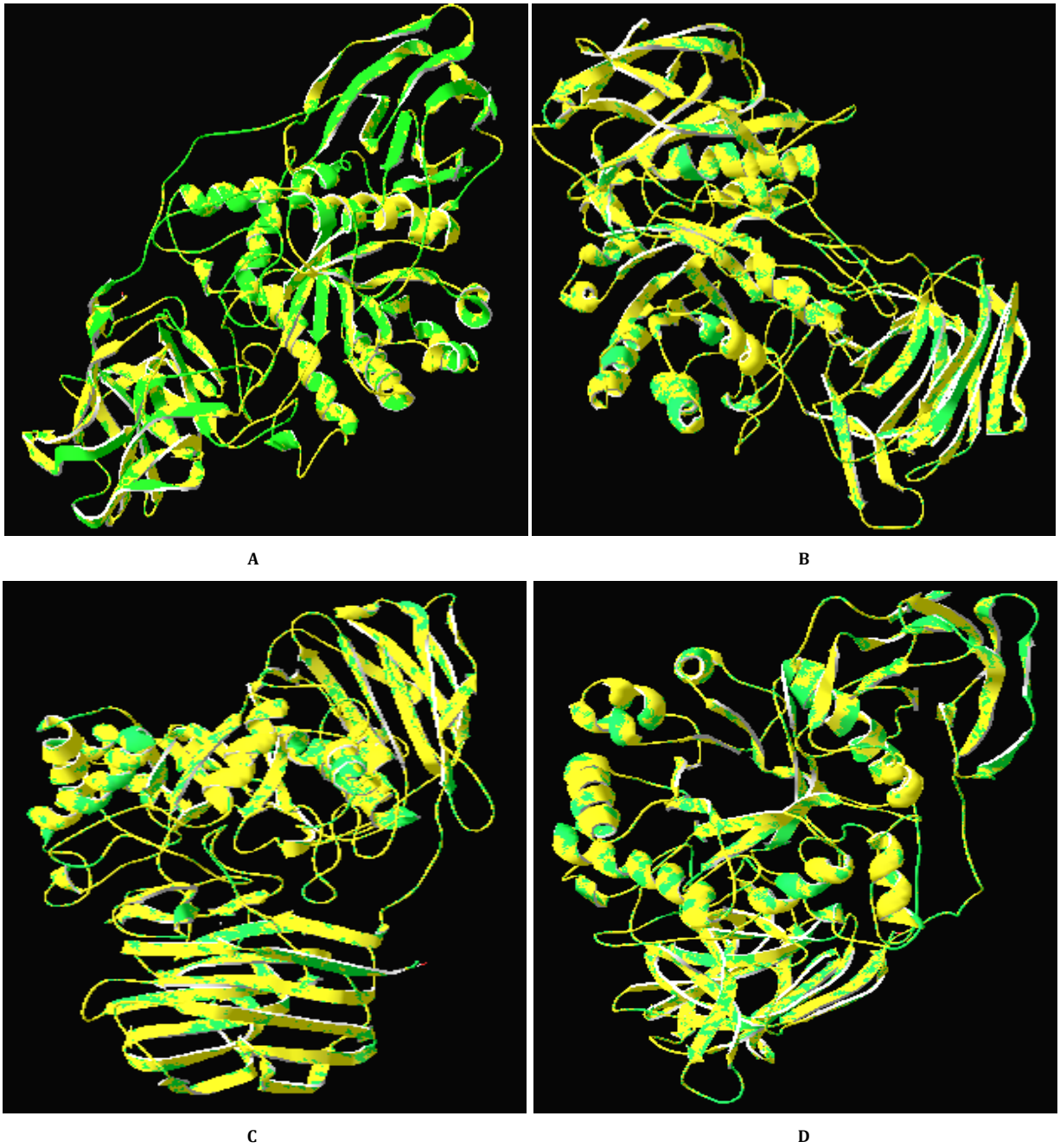

Fig. 1: (A) Superimposed structure of native protein (green color) with mutant protein G57C (yellow color). (B) Superimposed structure of native protein (green color) with mutant protein I562C (yellow color). (C) Superimposed structure of native protein (green color) with mutant protein R79C (yellow color). (D) Superimposed structure of native protein (green color) with mutant protein R184C (yellow color).

CONCLUSION

The GALC gene was investigated in this work by evaluating the influence of functional SNPs through computation methods. Of the total of 1615 SNPs in the GALC gene, 12 were found to be nonsynonymous SNPs in that 3 are deleterious found by SIFT and 4 were damaging as per the PolyPhen server. 2 nsSNPs were found to be common in both SIFT and PolyPhen server.

Higher the RMSD value is the more deviation between the two structures, which in turn changes their functional activity. It was found that the major mutation in the native protein of the GALC gene

was from Arginine to Cystine. So we conclude that rs73312829 with a mutation of Arginine to Cystine at position 79 in the native protein could be the main target mutation for the Krabbe disease caused by the GALC gene.

ACKNOWLEDGMENT

The authors thank the management of Vellore Institute of Technology, for providing the facilities to carry out this work.

REFERENCE

1. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Natural Genetics 1999; 22:231–238.
2. Nadeau JH. Single nucleotide polymorphisms: tackling complexity. Nature, 2002; 420:517– 518.
3. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Research, 2002; 30:3894–3900.
4. Renell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 Lysozyme. Journal of Molecular Biology, 1991; 222:67–87
5. Barroso I, Gurnell M, Croeley VE, Agostini M, Schwabe JW, Soos MA, LI Maslen G, Williams TDM, Lewis H, Schafer AJ, Chatterjee VKK & O'Rahilly S. Dominant negative mutations in human PPAR gamma associated with severe insulin resistance diabetes mellitus and hypertension. Nature 1999; 402:880–883.
6. Thomas R, McConnell R, Whittacker J, Kirkpatrick P, Bradley J, Stanford R. Identification of mutations in the repeated part of the autosomal dominant polycystic kidney disease type 1 gene PKD1 by long range PCR. American Journal of Human Genetics 1999; 65:39–49.
7. Dryja TP, Mcgee TL, Halu LB, Conley GS, Olsson JE, Reichel E. Mutations within the rhodopsin gene in patients with autosomal dominant retinitis pigmentosa. The New England Journal of Medicine, 1990; 323:1302–1307.
8. Smith EP, Boyd J, Frank GR, Takahashi M, Cohen RM, Specker B. Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man. The New England Journal of Medicine, 1994; 331:1056–1061.
9. Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. Human Gene Mutation Database: towards a comprehensive central mutation database. Journal of Medical Genetics, 2008; 45:124–126.
10. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS. Human Gene Mutation Database. Human Mutation 2003; 21:577–581.
11. Collins FS, Brooks LD. A DNA polymorphism discovery resource for research on human genetic variations. Genome Research, 1998; 8:1229–1231.
12. Saunders CT, Barker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. Journal of Molecular Biology, 2002; 322:891–901.
13. Lau AY, Chasman DI. Functional classification of proteins and protein variants. Proceedings of the National Academy of Sciences of the United States of America, 2004; 101:6576–6581.
14. Hussain SA, Zimmerman HH, Abdul-Rahman OA, Hussaini SM, Parker CC, Khan M. Optic Nerve Enlargement in Krabbe Disease: A Pathophysiologic and Clinical Perspective. Journal of Child Neurology, 2011; 26:642–644.
15. Cannizzaro LA, Chen YQ, Rafi MA, Wenger DA. Regional mapping of the human galactocerebrosidase gene (GALC) to 14q31 by in situ hybridization. Cytogenetic and Cell Genetics, 1994; 66:244–5.
16. Paola Luzi, Mohammad A. Rafi, and David A. Wenger. Structure and Organization of the human galactocerebrosidase (GALC) gene. Genomics 1995; 26:407-409.
17. Paola Luzi, Mohammad A. Rafi and David A. Wenger. Characterization of the large deletion in the GALC gene found in patients with Krabbe disease. Human Molecular Genetics 1995; 2335–2338.
18. Barbara Tappino, Roberta Biancheri, Matthew Mort, Stefano Regis, Fabio Corsolini, Andrea Rossi, Marina Stroppiano, Susanna Lualdi, Agata Fiumara, Bruno Bembi, Maja Di Rocco, David N Cooper, and Mirella Filocamo. Identification and Characterization of 15 Novel GALC Gene Mutations Causing Krabbe Disease. Human Mutation 2010; 31:1894–1915.
19. Xu C, Sakai N, Taniike M, Inui K, Ozono K. Six novel mutations detected in the GALC gene in 17 Japanese patients with Krabbe disease, and new genotype-phenotype correlation. Journal of Human Genetics. 2006; 51:548-54.
20. Lissens W, Arena A, Seneca S, Rafi M, Sorge G, Liebaers I, Wenger D, Fiumara A. A single mutation in the GALC gene is responsible for the majority of late onset Krabbe disease patients in the Catania (Sicily, Italy) region. Human Mutation. 2007; 28:742.
21. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski M, Sirotkin K, dbSNP: the NCBI database of genetic variation. Nucleic Acids Research, 2001; 29:308–311.
22. Ng CP, Henikoff S, SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research, 2003; 31:3812–3814.
23. Ng CP, Henikoff S, Predicting deleterious amino acid substitutions. Genome Research, 2001; 11:863–874.
24. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Research, 2002; 30:3894–3900.
25. Cavallo A, Martin AC. Mapping SNPs to protein sequence and structure data. Bioinformatics, 2005; 8:1443–1450.