Iranian J Publ Health, Vol. 41, No.7, July 2012, pp.24-33

Original Article

Computational Prediction of Phylogenetically Conserved Sequence Motifs for Five Different Candidate Genes in Type II Diabetic Nephropathy

T Sindhu, S Rajamanikandan, *P Srinivasan

Dept. of Bioinformatics, Alagappa University, Karaikudi, India

(Received 11 Feb 2012; accepted 24 Apr 2012)

Abstract

Background: Computational identification of phylogenetic motifs helps to understand the knowledge about known functional features that includes catalytic site, substrate binding epitopes, and protein-protein interfaces. Furthermore, they are strongly conserved among orthologs, indicating their evolutionary importance. The study aimed to analyze five candidate genes involved in type II diabetic nephropathy and to predict phylogenetic motifs from their corresponding orthologous protein sequences.

Methods: AKR1B1, APOE, ENPP1, ELMO1 and IGFBP1 are the genes that have been identified as an important target for type II diabetic nephropathy through experimental studies. Their corresponding protein sequences, structures, orthologous sequences were retrieved from UniprotKB, PDB, and PHOG database respectively. Multiple sequence alignments were constructed using ClustalW and phylogenetic motifs were identified using MINER. The occurrence of amino acids in the obtained phylogenetic motifs was generated using WebLogo and false positive expectations were calculated against phylogenetic similarity.

Results: In total, 17 phylogenetic motifs were identified from the five proteins and the residues such as glycine, leucine, tryptophan, aspartic acid were found in appreciable frequency whereas arginine identified in all the predicted PMs. The result implies that these residues can be important to the functional and structural role of the proteins and calculated false positive expectations implies that they were generally conserved in traditional sense.

Conclusion: The prediction of phylogenetic motifs is an accurate method for detecting functionally important conserved residues. The conserved motifs can be used as a potential drug target for type II diabetic nephropathy.

Keywords: Diabetic nephropathy, Conserved regions, Phylogenetic motifs, PHOG1.0, MINER

Introduction

Diabetes mellitus is characterized by the metabolic disorders of carbohydrate, lipid, and protein. “Type II diabetes mellitus is one of the primary threats to human health due to increasing prevalence, chronic course and disabling complications” (1, 2). Diabetic nephropathy (DN) is a major microvascular complication that affects 30-40% of all diabetic patients and represents a major cause of morbidity and mortality, due to a serious gradual decline in renal function (3). Several genes, proteins, and environmental factors are likely to contribute to the onset of the disease DN (4). Several candidate genes have been identified for the association with DN using case-control studies. They were selected for their positional and/or functional characteristics and the contribution of the corresponding proteins in the pathophysiological axes (5).

***Corresponding Author:** Tel: +91-4565-230725, E-mail address: sri.bioinformatics@gmail.com

The expression of AKR1B1 gene has been seen in human kidneys. It catalyzes the reduction of glucose to sorbitol. In hyperglycaemic condition, this pathway becomes activated by excess amount of glucose, whereas in case of normal condition, it is relatively inactive. High levels of sorbitol accumulation disrupt osmoregulation in kidney cells, which leads to kidney damage (6). ELMO1 is promoting excess transcription growth factor-β, collagen type 1, fibronectin and integrin-linked kinase expression and inhibiting cell adhesion when it is over expressed. ELMO1 is expressed in the presence of high glucose and it has a potential role in the pathogenesis of diabetic nephropathy (7). Insulin like growth factor binding proteins plays a major role in cell growth and metabolism. It influences cell adhesion and migration and interacts with α5β1. Over expression of IGFBP1 is associated with many glomerular diseases, including diabetic nephropathy (8).

Ectonucleotidepyrophosphate/phosphodiesterase 1 is a candidate susceptibility gene for type 2 diabetes and obesity. It helps to catalyze the release of nucleoside 5-phosphatase from nucleotides and their products. ENPP1 is expressed in several tissues such as skeletal tissue, adipose tissue, liver and kidney tissues. This gene is a risk factor for the development of diabetic nephropathy in type 2 diabetic patients (9). Apolipoprotein E gene is associated with susceptibility of type 1 and 2 diabetic nephropathy. A polymeric protein consists of three isoforms defined by a single amino acid substitution at two sites. The affinity of ApoE for its receptors is altered by these amino acid substitutions, thereby influencing lipid metabolism. Several studies proved that the isoforms of ApoE is associated with diabetic nephropathy (10).

Computational methods to predict the function of a protein from its amino acid sequence play a major role in guiding the experimental characterization of a genome (11). Although experimental methods exist to identify sequences bound by a specific protein, they have not been widely applied, and computational approaches to 'motif discovery' have proven to be a useful alternative (12). A sequence-based phylogenetic motif represents a promising functional site prediction strategy. Phylogenetic Motifs (PMs) are short sequence alignment fragments; consistently correspond to functional sites defined by surface loops, active site clefts and less exposed regions (13). Structural clusters of conserved positions and the trace residues, which are alignment positions that are conserved within the phylogenetic clusters, are used to identify functional regions (14).

The aim of the present study was to identify the functional region of proteins by building multiple sequence alignment between the target sequence and its sequence orthologous in order to find preferentially conserved residues. Structural verification was also done to check the accuracy of functional site prediction.

Materials and Methods

Datasets

Datasets consist of five candidate genes expressed in type II diabetic nephropathy such as aldose reductase, apolipoprotein E, engulfment and cell motility protein 1, ectonucleotide pyrophosphatase/ phosphodiesterase family member 1, Insulin-like growth factor binding protein 1 were obtained from the literature (15-19). Protein sequences corresponding to each gene were retrieved from UniprotKB database and their structures were downloaded from PDB.

Phylogenetic analysis

The dataset of orthologous protein sequences, those sharing a common ancestor were obtained using PHOG database, available at http://bioinf.fbb.msu.ru/phogs/index.html.

The PHOG database is used in various areas such as comparative genomics, proteomics, and evolutionary studies (20). The obtained orthologous proteins grouped together and their sequences were aligned to compare equivalent residues using the program ClustalW (21). A global dynamic programming al-

gorithm was used to construct an alignment for full length of the sequences. These multiple sequence alignments provide structural and functional information.

Phylogenetic motifs identification using MINER

The program MINER was used for PM identification, which is available at http://www.pmap.csupomona.edu/MINER/. It uses a sliding sequence window algorithm to evaluate comprehensively the phylogenetic similarity between each window and the complete alignment (22). The multiple sequence alignment file generated by ClustalW and three-dimensional structures of the master proteins were used as input file for the program MINER. For all the data sets, sequence window width of 5 was used.

Each protein family requires a unique value to identify correctly the functional regions. Phylogenetic similarity cutoff falls between 1.5 and 2.0 was used for accurate functional site predictions (23). Therefore, the threshold value was adjusted to -1.7 for all the datasets. Partition metric (PAM) clustering algorithm was used to evaluate the optimal range of thresholds. Phylogenetic similarity was calculated using the partition metric algorithm consequently resulting in phylogenetic similarity z scores.

Results

In our study, phylogenetic approach was used for the identification of the key functional residues. The five candidate genes obtained for this study and investigated PDB structures of the proteins involved and the number of identified PMs were listed in Table 1.

Table 1: Predicted phylogenetic motif parameters of the discussed protein data set

Genes	HUGO symbol	OMIM reference	UniprotKB ID	Seq[a]	PSZ[b]	PMs[c]	Structure investigated
Aldose reductase	AKR1B1	103880	P15121	19	-1.7	7	1ABN
Apolipoprotein E	APoE	107741	P02649	37	-1.7	5	1B68
Engulfment and cell motility gene 1	ELMO1	606420	Q92556	7	-1.7	1	2VSZ
Ectonucleotide pyrophosphatase/ Phosphodiesterase family member 1	ENPP1	173335	P22413	12	-1.7	1	2YS0
Insulin-like growth factor binding protein 1	IGFBP1	146730	P08833	10	-1.7	3	1ZT5

[a]Number of sequences in the alignment
[b]Phylogenetic similarity z-score threshold used in identification of the phylogenetic motifs
[c]Number of phylogenetic motifs identified

Neighbor-Joining (NJ) method in ClustalW was used for the construction of phylogenetic tree of five proteins with its orthologs was shown in Fig. 1. The functional importance of the proteins was verified through the three-dimensional structures to highlight better PM regions. Predicted conserved regions within the structure of the five proteins were shown in Fig. 2.

Predicted highly conserved residues by WebLogo

The conserved residues were observed from a cursory examination of sequence logos. The sequence logos of the predicted motifs were shown in Fig. 3. In total, 17 conserved regions were identified for the five proteins with its orthologs. The seven PMs were identified in the protein aldose reductase and amino acid residues such as Lys21, Trp20, Thr19, Gly18, Leu17, Ala30, Glu29, Gln26, Val27, Thr28, Ile58, Gln59, Leu62, Lys61, Glu60, Gly90, Lys89, Val88, Leu87, Gly86, Asp230, Ile233, Arg232, Pro231, Val259, Ile260, Pro261, Lys262,

Ser263, Val 264, Thr265, Pro266, Leu301, Ser302, Cys303, Thr304, Ser305, His306 were identified from the predicted PMs. Among these residues, lysine and threonine remains highly conserved residues whereas histidine, tryptophan, aspartic acid, and alanine identified only one of the PMs.

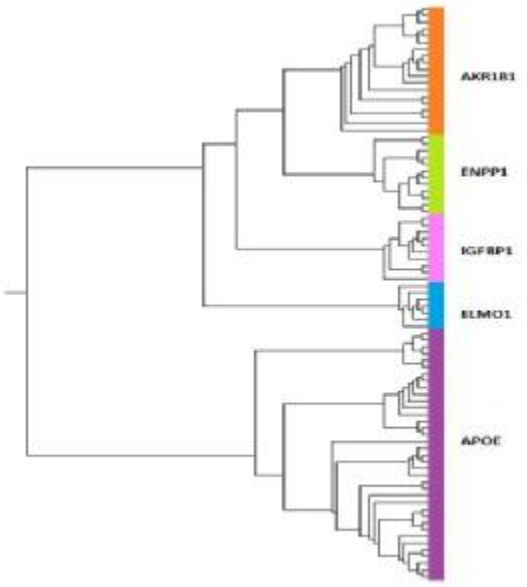


Fig. 1: The unrooted phylogenetic tree is composed of selected five-protein family with their orthologs. Colour differences within the phylogenetic tree correspond to the discussed proteins such as aldose reductase (orange), ectonucleotide pyrophosphatase/phosphodiesterase family member 1(light green), Insulin-like growth factor-binding protein 1(light pink), engulfment and cell motility protein 1 (blue) and apolipoprotein E (purple) with their orthologs

The three amino acid residues such as glutamic acid, leucine and alanine residues were observed as highly conserved residues among the five identified PMs in apolipoprotein E. Glycine, aspartic acid, arginine, valine, methonine, glutamine, and threonine were observed in appreciable frequency whereas serine observed in only one of the PMs. For Insulin-like growth factor binding protein 1, glutamic acid was commonly occurred as highly conserved residue in the identified 3 PMs.

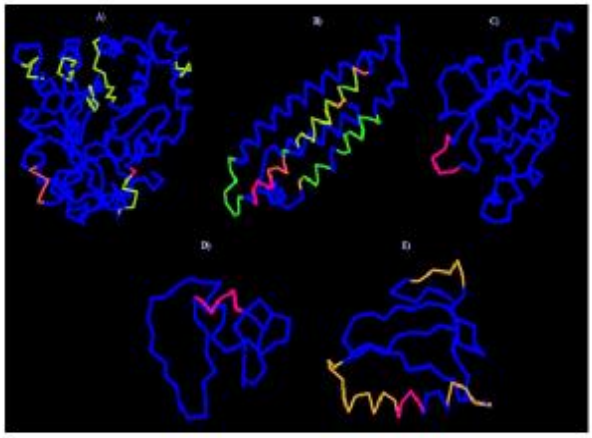
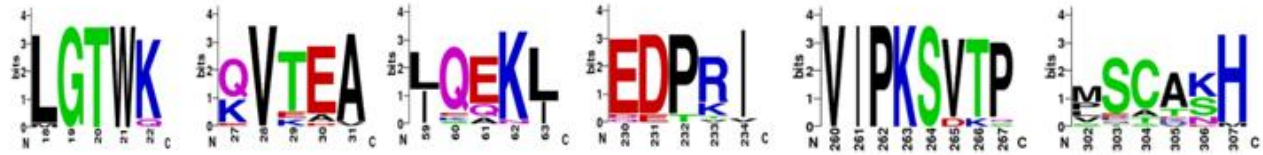


Fig. 2: The figure shows the predicted PMS in the five proteins obtained for this study (A) aldose reductase, (B) apolipoprotein E, (C) engulfment and cell motility protein 1, (D) ectonucleotide pyrophosphatase/phosphodiesterase family member 1, (E) Insulin-like growth factor-binding protein 1 and the each identified PM within them was highlighted with different colours

Proline, serine, valine, arginine and lysine were also found in appreciable frequency whereas cysteine found only one of the PMs.

Only one PM was identified for engulfment and cell motility protein 1 and the amino acid residues occurred in the PM was Arg570, Arg569, Arg568, Ala567 and Asn566. Among these residues, arginine remains conserved although asparagine and alanine were also observed. Similarly, one PM was predicted in ectonucleotide pyrophosphatase/ phosphodiesterase family member 1 and the highly conserved residues like, Phe13, Arg14, Glu17, Gly16 and Cys15 were occurred in the identified PM.

(A) Aldose reductase



(B) Apolipoprotein E

(C) Engulfment and cell motility protein 1

(D) Ectonucleotide pyrophosphatase/phosphodiesterase family member 1

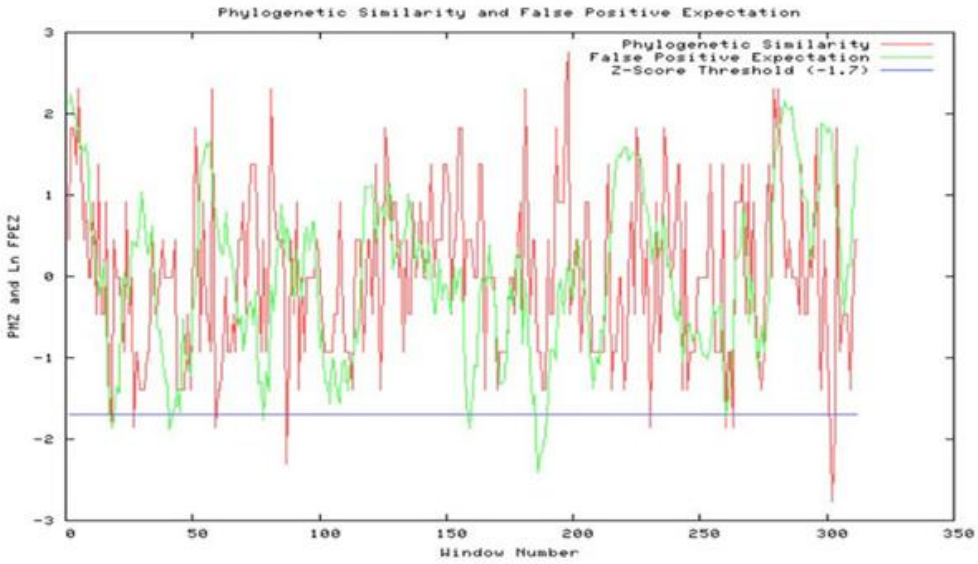
(E) Insulin-like growth factor-binding protein 1

Fig. 3: Sequence logos of **(A-E)** visually highlight the occurrence of conserved amino acid residues in the identified PMs

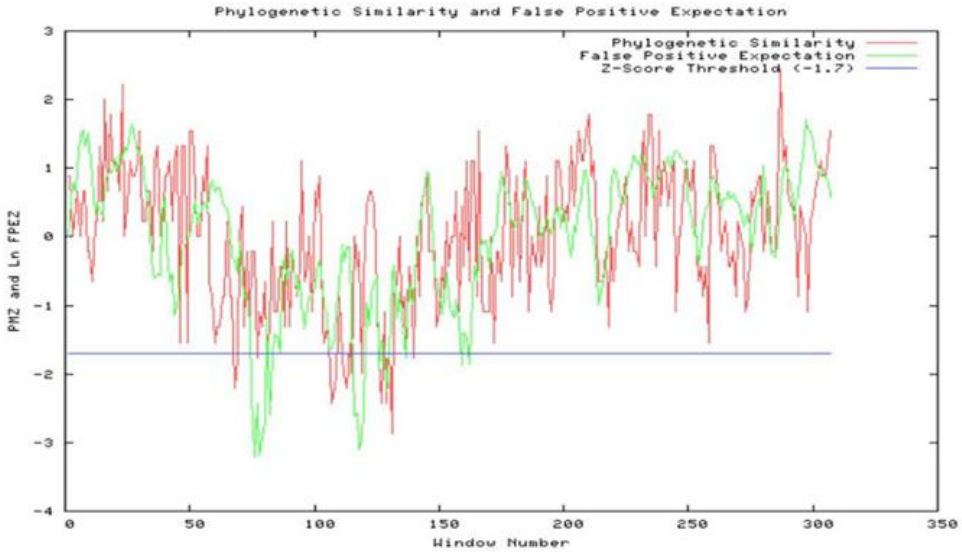
Phylogenetic similarity z scores (PSZs) vs. False positive Expectations (FPEs)

To determine the probability of randomly encountering each motif, FPEs were calculated. FPE approach was used to identify the traditional motifs, which are low sequence entropy regions. Too many false positives predicted by using conservation-based approaches, were said to be satisfactory. The identified phylogenetic similarities versus FPE for the five proteins were demonstrated in Fig. 4. From the result, too many FPEs were occurred in ectonucleotide pyrophosphatase/ phosphodiesterase family member 1, aldose reductase and apolipoprotein E. So, the conserved residues predicted from these proteins can be considered as traditional motifs.

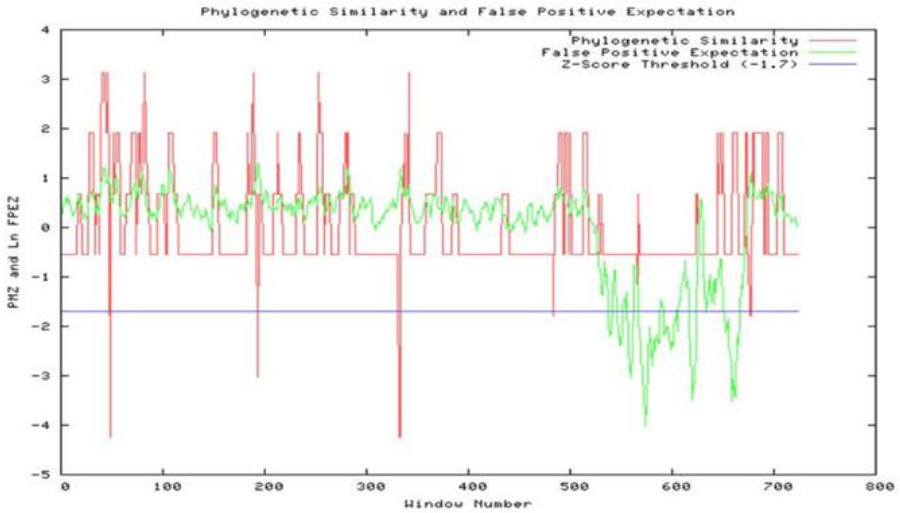
(A)


Phylogenetic Similarity and False Positive Expectation
Phylogenetic Similarity
False Positive Expectation
Z-Score Threshold (-1.7)
PHZ and Ln FPEZ
Window Number

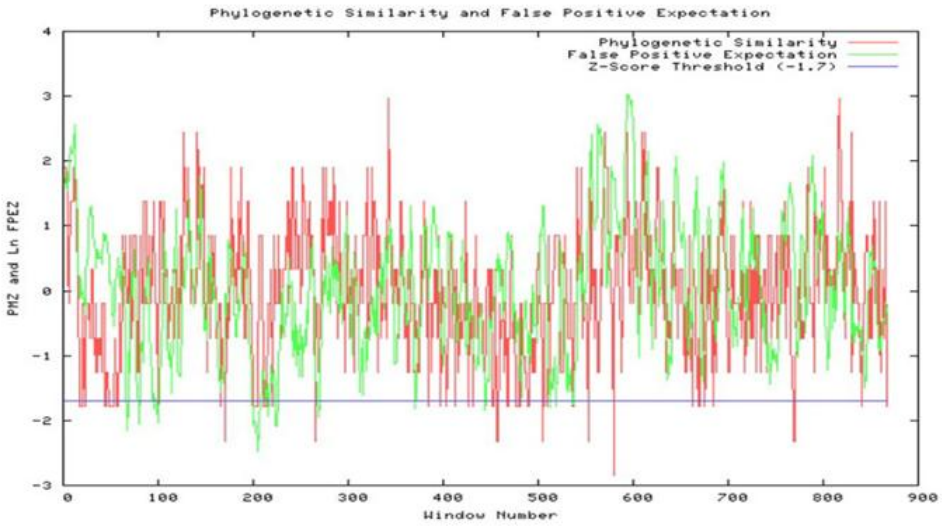
(B)


Phylogenetic Similarity and False Positive Expectation
Phylogenetic Similarity
False Positive Expectation
Z-Score Threshold (-1.7)
PHZ and Ln FPEZ
Window Number

(C)


Phylogenetic Similarity and False Positive Expectation
Phylogenetic Similarity
False Positive Expectation
Z-Score Threshold (-1.7)
PHZ and Ln FPE2
Window Number

(D)


Phylogenetic Similarity and False Positive Expectation
Phylogenetic Similarity
False Positive Expectation
Z-Score Threshold (-1.7)
PHZ and Ln FPE2
Window Number

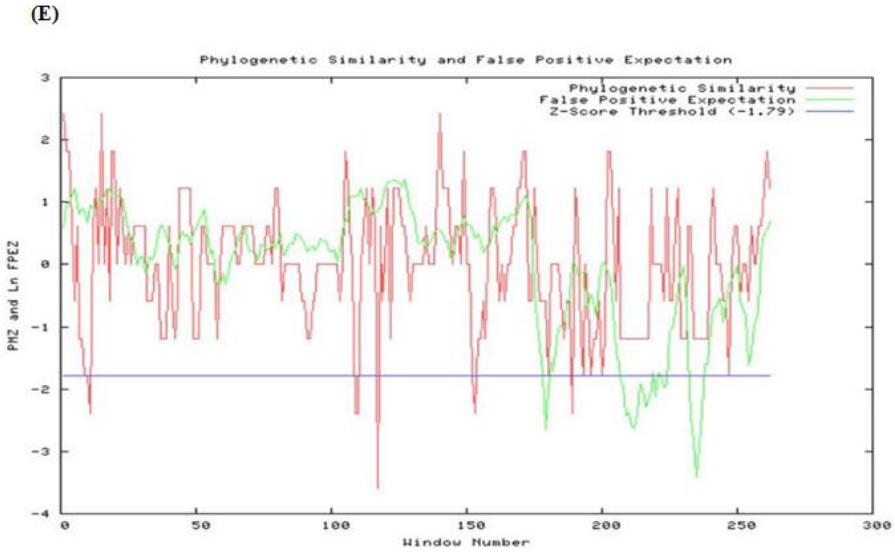


Fig. 4: The plot demonstrates the sequence correspondence of phylogenetic (red) and traditional motifs (green). **(A)** aldose reductase, **(B)** apolipoprotein E, **(C)** engulfment and cell motility protein 1 **(D)** ectonucleotide pyrophosphatase/ phosphodiesterase family member 1 **(E)** Insulin-like growth factor binding protein 1

Discussion

In the present study, computational analysis of available sequences, crystal structures of the five different expressed proteins in type II diabetic nephropathy were used to identify functionally important residues. Comparing the structures of homologous proteins and analyzing of large multiple sequence alignments can help to identify sequence, structural conservation, and conserved interactions that are crucial for protein stability and function (24). Here, the three dimensional structures of the proteins were used to establish the accuracy of functional site predictions. Ortholog detection is essential for functional annotation of genomes, with applications to predicting protein-protein interactions (25). Based on phylogenetic analysis, a group of evolutionary related ortholog sequences were identified and aligned to find the phylogenetically conserved regions of proteins.

In the present analysis, the occurrence of amino acid residues in the identified PMs was generated by WebLogo. The sequence logos of each identified PMs provides the highly conserved residues in the sequence. Each logo consists of stacks of letters and one stack for each position in the sequence. The overall height of each stack indicates the sequence conservation whereas the height of symbols within the stack reflects the relative frequency of the corresponding amino acid at that position (26). From the results, the amino acids such as glycine, tryptophan and aspartic acid were found in appreciable frequency whereas arginine was identified in all the predicted PMs. All the conserved positions are not related to function but some amino acids tend to have structural roles when conserved (e.g. Trp, Leu, Gly, Cys) while others (mainly a polar amino acids, or specific types e.g. Asp, Ser, Cys, His) tend to be part of binding and active sites (27). Here, the conserved residue leucine occurred in apolipoprotein E, can be responsible for the structural role of the protein. Therefore, the result suggests that these conserved positions

which were predicted, can be important to the functional diversity of the proteins.

Traditional motifs were predicted from the identified PMs by calculating FPE in the same sequence window, which was used to calculate PMs, and FPE describes the probability of encountering each sequence window. PMs were identified based on phylogenetic similarity, whereas FPEs were calculated based on sequence conservation. Motif-based approaches result in many false positives to be useful in large-scale analyses (28). The predicted maps clearly indicate the presence of traditional motifs by showing too many FPEs.

The focus of the present study was to map conserved positions to a representative structure and orthologous sequences of the five different candidate genes expressed in type II diabetic nephropathy. The calculated PSZs versus FPE motif identification shows that phylogenetic motifs can be considered as traditional motifs. Most of the identified conserved residues were expected critically related to the function of the protein. Further investigation on these functional sites would provide a potential drug target for type II diabetic nephropathy.

Ethical considerations

Ethical issues (Including plagiarism, Informed Consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely observed by the authors.

Acknowledgments

The authors like to thank the Department of Bioinformatics, Alagappa University, Karaikudi, India for the support and providing the facilities for this work. The authors declare that there is no conflict of interests.

References

1. Sharma B, Balomajumder C, Roy P (2008). Hypoglycemic and hypolipidaemic effects of flavonoid rich extract from *Eugenia jambolana* seeds on streptozotocin induced diabetic rats. *Food Chem Toxicol*, 46(7): 2376–83.
2. Leena AA, Jill PC (2010). Type 2 diabetes prevention: A review. *Clinical Diabetes*, 28(2): 53-9.
3. Vionnet N, Treqouet D, Kazeem G, Gut I, Groop PH, Tarnow L, Parving HH, Hadjadi S, Forsblom C, Farrall M, Gauquier D, Cox R, Matsuda F, Heath S, Thevard A, Rousseau R, Cambien F, Marre M, Lathrop M (2006). Analysis of 14 candidate genes for diabetic nephropathy on chromosome 3q in European populations: strongest evidence for association with a variant in the promoter region of the adiponectin gene. *Diabetes*, 55(11): 3166-74.
4. Ntemka A, IIiadis F, Papanikolaou NA, Grekas D (2011). Network-centric analysis of genetic predisposition in Diabetic Nephropathy. *Hippokratia*, 15(3): 232-7.
5. Granier C, Makni K, Molina L, Jardin-Waltelet B, Ayadi H, Jarraya F (2008). Gene and protein markers of diabetic nephropathy. *Nephrol Dial Transplant*, 23(3): 792-9.
6. Tang WH, Martin KA, Hwa J (2012). Aldose reductase, oxidative stress, and diabetic mellitus. *Frontiers in Pharmacology*, 3:87.
7. Pezzolesi MG, Katavetin P, Kure M, Poznik GD, Skupien J, Mychaleckyi JC, Rich SS, Warram JH, Krolewski AS (2009). Confirmation of genetic associations at ELMO1 in the GoKinD collection supports its role as a susceptibility gene in diabetic nephropathy. *Diabetes*, 58(11): 2698-702.
8. Vasylyeva TL, Ferry RJ Jr (2007). Novel roles of the IGF-IGFBP axis in etiopathophysiology of diabetic nephropathy. *Diabetes Res Clin Pract*, 76(2): 177-86.
9. Sortica DA, Crispim D, Zaffari GP, Friedman R, Canani LH (2011). The role of ectonucleotide pyrophosphatase/ phosphodiesterase 1 in diabetic nephropathy. *Arg Bras Endocrinol Metabol*, 55(9): 677-85.
10. Freedman BI, Bostrom M, Daeihagh P, Bowden DW (2007). Genetic factors in diabetic nephropathy. *Clin J Am Soc Nephrol*, 2: 1306-16.

11. Martin DM, Berriman M, Barton GJ (2004). Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5: 178.
12. Moses AM, Chiang DY, Eisen MB (2004). Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput*, 324-35.
13. Roshan U, Livesay DR, La D (2005). Improved Phylogenetic Motifs Detection using Parsimony. *Proceedings of the IEEE BIBE Meeting BIBE05*: 19-26.
14. Dukka BK, Livesay DR (2008). Improving position-specific predictions of protein functional sites using phylogenetic motifs. *Bioinformatics*, 24(20): 2308-16.
15. Wolford JK, Yeatts KA, Red Eagle AR. Nelson RG, Knowler WC, Hanson RL (2006). Variants in the gene encoding aldose reductase (AKR1B1) and diabetic nephropathy in American Indians. *Diabet Med*, 23(4): 367-76.
16. Guan J, Zhao HL, Baum L, Sui Y, He L, Wong H, Lai FM, Tong PC, Chan JC (2009). Apolipoprotein E polymorphism and expression in type 2 diabetic patients with nephropathy: clinicopathological correlation. *Dial Transplant Nephrol*, 24(6): 1889-95.
17. Leak TS, Perlegas PS, Smith SG, Keene KL, Hicks PJ, Langefeld CD, Mychaleckyi JC, Rich SS, Kirk JK, Freedman BI, Bowden DW, Sale MM (2009). Variants in intron 13 of the ELMO1 gene are associated with diabetic nephropathy in African Americans. *Ann Hum Genet*, 73(2): 152-9.
18. Keene KL, Mychaleckyi JC, Smith SG, Leak TS, Perlegas PS, Langefeld CD, Freedman BI, Rich SS, Bowden DW, Sale MM (2008). Association of the distal region of the ectonucleotide pyrophosphatase/ phosphodiesterase 1 gene with the type 2 diabetes in an African-American population enriched for nephropathy. *Diabetes*, 57(4): 1057-62.
19. Stephens RH, McElduff P, Heald AH, New JP, Worthington J, Ollier WE, Gibson JM (2005). Polymorphisms in IGF-binding protein 1 are associated with impaired renel function in type 2 diabetes. *Diabetes*, 54(12): 3547-53.
20. Merkeev IV, Novichkov PS, Mironov AA (2006). PHOG: a database of supergenomes built from proteome complements. *BMC Evol Biol*, 6: 52.
21. Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22): 4673-80.
22. La D, Livesay DR (2005). MINER: software for phylogenetic motif identification. *Nucleic Acids Res*, 33: W267-70.
23. La D, Livesay DR (2005). Predicting functional sites with an automated algorithm suitable for heterogeneous datasets. *BMC Bioinformatics*, 6: 116.
24. Grant BJ, McCammon JA, Caves LS, Cross RA (2007). Multivariate analysis of conserved sequence-structure relationships in kinesins: coupling of the active site and a tubulin-binding sub-domain. *J Mol Biol*, 368(5): 1231-48.
25. Datta RS, Meacham C, Samad B, Neyer C, Sjolander K (2009). Berkeley PHOG: Phylofacts orthology group prediction web server. *Nucleic Acids Res*, 37: W84-9.
26. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: A sequence logo generator. *Genome Res*, 14(6): 1188-90.
27. Pazos F, Bang JW (2006). Computational prediction of functionally important regions in proteins. *Current Bioinformatics*, 1: 15-23.
28. La D, Sutch B, Livesay DR (2005). Predicting protein functional sites with phylogenetic motifs. *Proteins*, 58(2): 309-20.