Type 2 diabetes mellitus: phylogenetic motifs for predicting protein functional sites

Ashok Sharma*, Tanuja Rastogi, Meenakshi Bhartiya, A K Shasany and S P S Khanuja

Bioinformatics Division, Central Institute of Medicinal and Aromatic Plants, PO CIMAP, Lucknow 226 015, India

**Corresponding author (Email, ashoksharma@cimap.res.in)*

Diabetes mellitus, commonly referred to as diabetes, is a medical condition associated with abnormally high levels of glucose (or sugar) in the blood. Keeping this view, we demonstrate the phylogenetic motifs (PMs) identification in type 2 diabetes mellitus very likely corresponding to protein functional sites. In this article, we have identified PMs for all the candidate genes for type 2 diabetes mellitus. Glycine 310 remains conserved for glucokinase and potassium channel KCNJ11. Isoleucine 137 was conserved for insulin receptor and regulatory subunit of a phosphorylating enzyme. Whereas residues valine, leucine, methionine were highly conserved for insulin receptor. Occurrence of proline was very high for calpain 10 gene and glucose transporter.

[Sharma A, Rastogi T, Bhartiya M, Shasany A K and Khanuja S P S 2007 Type 2 diabetes mellitus: phylogenetic motifs for predicting protein functional sites; *J. Biosci.* **32** 999–1004]

1. Introduction

Diabetes is affecting nearly 10% of the global population above 20 years of age. Its most prevalent form is type 2 diabetes (T2D), with a share of 90-95%. By the year 2025, an estimated 300 million humans will be suffering from T2D and all its associated clinical complications. T2D is often part of a metabolic syndrome that includes obesity, elevated blood pressure and high levels of blood lipids. The identification of genes has led to an better understanding of the molecular pathways involved in the pathogenesis of T2D and, therefore, to advances in the prevention and control of T2D.

In rare forms of diabetes, mutations of one gene can result in disease. However, in T2D many genes are thought to be involved (table1). One method of finding the diabetes susceptibility genes is by whole genome linkage studies. On this basis two genes, Calpain10 and Hepatocyte nuclear factor 4 alpha (HNF4A) were identified and reported (Dean *et al* 2004).

The present analysis also concentrates on all the candidate genes including the receptor genes involved in T2D. In this report the phylogenetic motifs (PMs) for the prediction of functional sites were identified. PMs are generally conserved in sequence which implies that they can be considered motifs in traditional sense as well. However, there are instances where PMs are not (overall) well conserved in sequence (La *et al* 2005). This point is enticing, because it implies that are able to identify key sequence regions that traditional motif-based approaches may not. Across a structurally and functionally heterogenous dataset phylogenetic motifs have been demonstrated to correspond to a wide variety of functional site archetypes, including those defined by surface loops, active site clefts and less exposed regions. In this study the PMs identified for all the disease causing candidate genes are reported.

Keywords. Functional site prediction; phylogenetic motif; sequence-function relationship

Abbreviations used: HNF4A, Hepatocyte nuclear; factor 4 alpha; PMS, phylogenetic motifs; T2D, type 2 iabetes.

2. Methods

2.1 *Dataset*

Datasets consist of all the disease causing candidate genes with their potential homologues sequences for each gene (table 1). The dataset of homologus sequences corresponding to each gene were obtained using software Flower Power (*http://phylogenomics.berkeley.edu/cgi-bin/flowerpower/input_flowerpower.py*). Flower Power is a protein homology clustering algorithm similar to PSI-Blast in its iterated approach to search homolgs by searching the UniProt database, includes phylogenetic tree construction, subfamily HMM construction in clustering and alignment process using CLUSTALW.

Table 1. Type 2 diabetes mellitus genes

Gene Name	Gene symbol
Sulfonylurea receptor	ABCC8
Calpain10 enzyme	CAPN10
Glucagon receptor	GCGR
Glucokinase	GCK
Glucose transporter	SLC2A2
Transcription factor HNF4A	HNF4A
Insulin receptor	INSR
Potassium channel KCNJ11	KCNJ11
Enzyme lipoprotein lipase	LPL
Transcription factor PPARG	PPARG
Regulatory subunit of a phosphorylating enzyme	PIK3R1

Table 2. Number of PMs identified and their location

Genes	#Seq[a]	#PSZ[b]	# PMs[c]
Sulfonylurea receptor	124	-1.7	18
Calpain10 enzyme	14	-1.7	9
Glucagon receptor	88	-1.7	9
Glucokinase	192	-1.7	8
Glucose transporter	200	-1.7	6
Transcription factor HNF4A	65	-1.7	2
Insulin receptor	34	-1.7	32
Potassium channel KCNJ11	186	-1.7	11
Enzyme lipoprotein lipase	107	-1.7	14
Transcription factor PPARG	71	-1.7	2
Regulatory subunit of a phosphorylating enzyme	30	-1.7	13

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

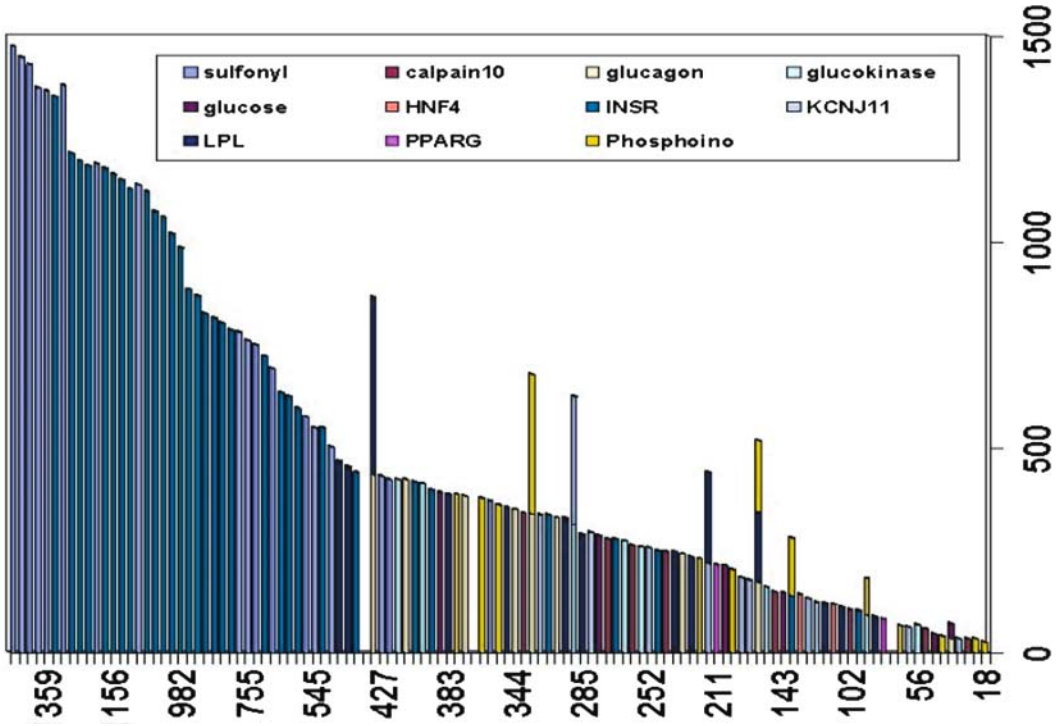
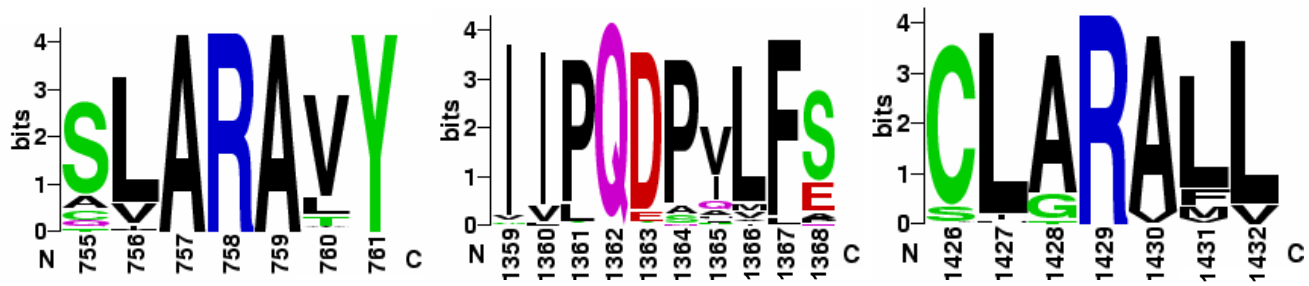
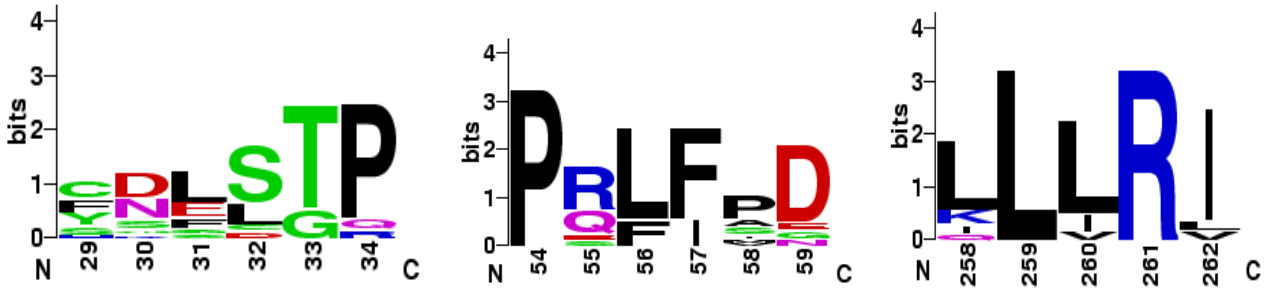


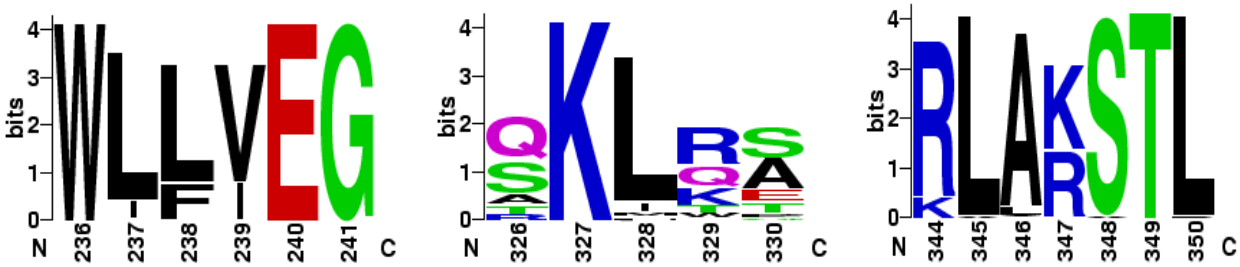
Figure 1. Graph showing identified PM range.



(a)Sulfonylurea Receptor



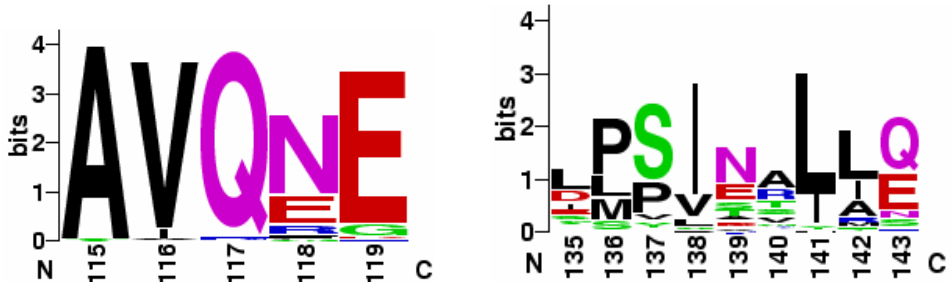
(b) Calpain 10



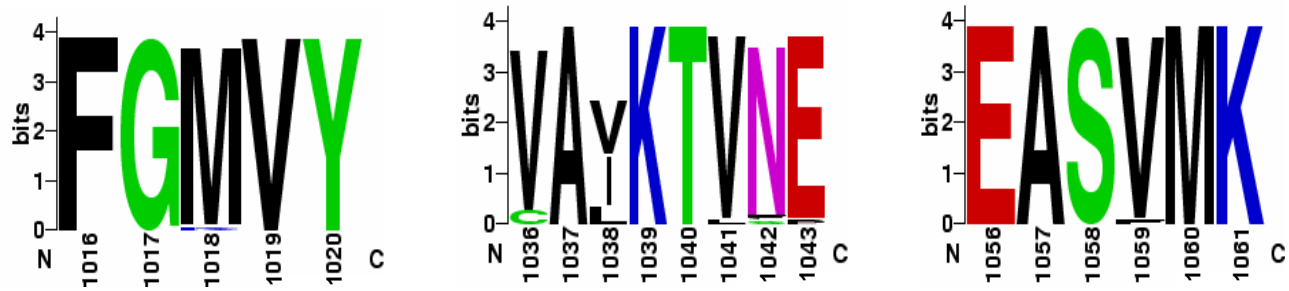
(c)Glucagon Receptor

(d)Glucokinase gene

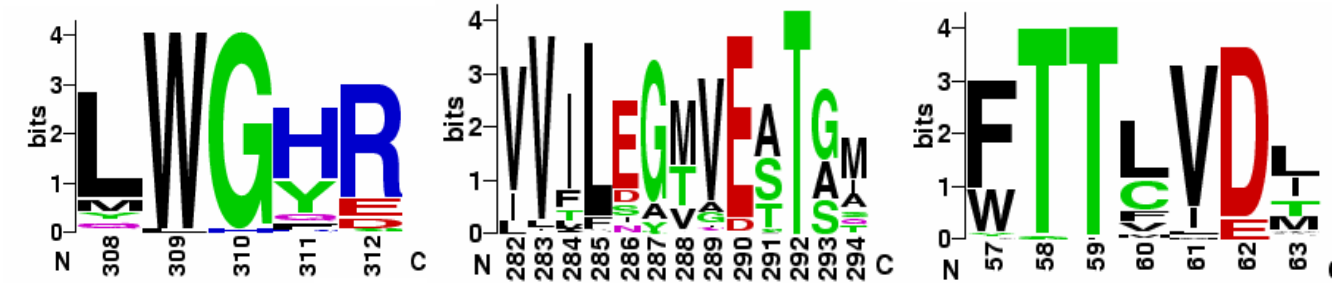
(e)Glucose Transporter



(f)Transcription Factor HNF4A



(g)Insulin Receptor

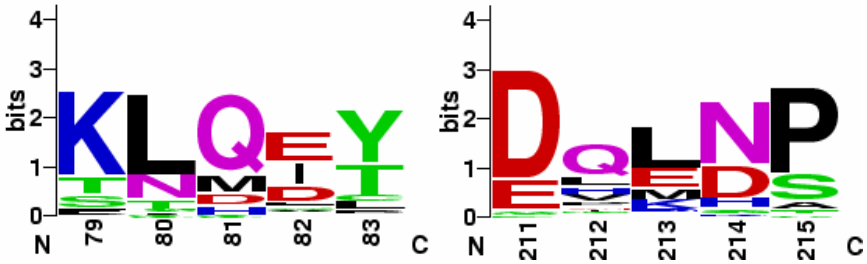


(h)Potassium Channel KCNJ11

2.2 *Phylogenetic motif identification*

Phylogenetic motifs were obtained using software MINER (La *et al* 2005). The aligned sequences generated through Flower Power were uploaded to MINER. In this report a window width of five, which has previously been shown to be most sensitive for identifying functional sites was used throughout (La and Livesay *et al* 2005). Phylogenetic similarity was quantified using PSZ value (lower PSZ values, indicate higher phylogenetic similarity).

(i)Enzyme Lipoprotein Lipase



(j)Transcription Factor PPARG

(k)Regulatory subunit of a Phosphorylating Enzyme

Figure 2(a–k). Sequence logos (*weblogo.berkaley.edu*) showing frequency of occurrences of amino acid residues in the obtained phylogenetic motifs.

After all phylogenetic comparisons are made, the PSZ threshold was adjusted to alter what constitutes a "hit". The threshold can be adjusted to be more stringent or more accommodating. Relaxing the z-score threshold identifies regions with no obvious functional relevance. On the other hand, setting the threshold too stringently will result in a large number of false positives. The effect of adjusting the z-score threshold was empirically examined. The threshold value of – 1.7 was taken for all the datasets.

3. Results and discussion

In this report, phylogenetic approach was used for the identification of the key functional residues in type 2 diabetes mellitus causing genes. For all the genes for diabetes mellitus type 2 software MINER[1] identifies PMs.

For Sulfonylurea receptor 18 PMs were identified (table 2). Among all the PMs two amino acid residues leucine and alanine remains highly conserved. Phenylalanine, alanine, glutamic acid, lysine, valine, threonine, isoleucine were also found in appreciable frequency. Arginine occurred only in one of the PMs. Whereas the occurence of methionine was very less (figure 2a).

[1]The software MINER is freely available at *http://www.pmap.csupomona.edu/MINER/*. Source code is available to academic community on request.

Similarly, the method was applied for 14 other protein sequences for Calpain10 gene and various PMs with some of the highly conserved residues like proline, tyrosine, phenylalanine (figure 2b).

In case of glucagon receptor leucine was one of the most conserved amino acid. Among all the PMs identified (table 2) for glucokinase gene, occurence of glycine remained conserved although phenyl alanine, arginine, aspartic acid, tryptophan were also observed (figure 2d). Three amino acid residues glutamate, glycine and proline were conserved for glucose transporters (figure 2e). Moreover serine, methionine were present less in number.

For potassium channel KCNJ11 other then glycine the occurrence of threonine, arginine was observed to be higher (figure 2h). Phenylalanine and serine were also present in fair number. In case of transcription factor HNF4A, valine, alanine, leucine isoleucine residues all were present fairly in similar frequency (figure 2f). Aspartic acid, lysine, leucine occurred equally in transcription factor PPARG. Glutamate and lysine were highly conserved for regulatory subunit of a phosphorylating enzyme. The frequency of occurrence of tyrosine and aspartate was also appreciable. For insulin receptor valine, leucine, methionine remains highly conserved whereas histidine and glutamine were less frequent (figure 2g).

Also some of the identified PMs for different genes fall in the same range (figure 1) with some of the amino acid residues conserved reflecting their significance. Glycine 310 remains conserved for glucokinase and potassium channel KCNJ11. Similarly, isoleucine 137 was conserved for insulin receptor and regulatory subunit of a phosphorylating enzyme. Thus, it can also be considered that the identified conserved residues in PMs are functional.

On the basis of fortuitous results it can be concluded that conserved residues obtained from PM's identification in T2D can be used for further investigation of future drug targets.

References

Babbitt P C, Hasson M S, Wedekind J E, Palmer D R, Barrett W C, Reed G H, Rayment I, Ringe D, Kenyon G L, Gerlt J A 1996 The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids; *Biochemistry* **35** 16489–16501

Bailey T L, Gribskov M 1998 Methods and statistics for combining motif match scores. *J Comput Biol* **5** 211–221

Dean L and McEntyre J 2004 *The genetic landscape of diabetes* (Bethesda, Md: NCBI, NIDDK) 1–57

La D and Livesay D R 2005 MINER: software for phylogenetic motif identification; *Nucleic Acids Res.* **33** W267–W270

La D, Sutch B and Livesay D R 2005 Predicting proten functional sites with phylogenetic motifs; *Proteins* **58** 309–320

Lichtarge O, Bourne H R and Cohen F E 1996 An evolutionary trace method defines binding surfaces common to protein families; *J. Mol. Biol*. **257** 342–358

Mihalek I, Res I and Lichtarge O 2004 A family of evolution–entropy hybrid methods for ranking protein residues by importance; *J. Mol. Biol.* **336** 1265–1282

Sol M A, Pazos F and Valencia A 2003 Automatic methods for predicting functionally important residues; *J. Mol. Biol*. **326** 1289–1302

*e*Publication: 28 June 2007