Bioinformatics Advance Access published April 21, 2006

BIOINFORMATICS

G8: a novel domain associated with polycystic kidney disease and non-syndromic hearing loss

Quan-yuan He[1], Xiang-hua Liu[2], Qiang Li[2], David J. Studholme[3], Xuan-wen Li[1] and Song-ping Liang[1*]

[1] Key Laboratory of Protein Chemistry and Developmental Biology of Education Committee, College of Life Sciences, Hunan Normal University, Changsha, P.R. China

[2] State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Handan Road 220, Shanghai 200433, P.R. China

[3] The Sainsbury Laboratory, Norwich, NR4 7UH, UK

Associate Editor: Alex Bateman

ABSTRACT

Summary: We report a novel protein domain – G8, which contains five repeated β-strand pairs and is present in some disease-related proteins such as PKHD1, KIAA1199, TMEM2 as well as other uncharacterized proteins. Most G8-containing proteins are predicted to be membrane-integral or secreted. The G8 domain may be involved in extracellular ligand binding and catalysis. It has been reported that missense mutations in the two G8 domains of human PKHD1 protein resulted in a less stable protein and are associated with autosomal-recessive polycystic kidney disease (ARPKD), indicating the importance of the domain structure. G8 is also present in the N-terminus of some non-syndromic hearing loss disease-related proteins such as KIAA1109 and TMEM2. Discovery of G8 domain will be important for the research of the structure/function of related proteins and beneficial for the development of novel therapeutics.

Contact: liangsp@hunnu.edu.cn

1 INTRODUCTION

Here we report a novel domain named G8, containing eight conserved glycine residues and consisting of five β-strand pairs. This novel domain is found in human disease-associated proteins PKHD1, KIAA1109, TMEM2 and some other uncharacterized proteins.

The PKHD1 protein (also known as fibrocystin and polyductin) is a large (447 kD) membrane protein involved in autosomal recessive polycystic kidney and hepatic disease. It is abundant in fetal-kidney collecting ducts but absent in the kidneys of some patients with autosomal recessive polycystic kidney disease. Its predicted structure suggests that it is an integral membrane receptor with extracellular protein-interaction sites and intracellular phosphorylation sites (Ward, et al., 2002) and may interact with extracellular protein ligands and transduce intracellular signals to the nucleus (Wilson, 2004).

KIAA1199, one of inner-ear-specific genes, is expressed in the cochlea and vestibule tissues. The KIAA1199 protein may be essential for auditory function and its mutated forms may cause non-syndromic hearing loss (Abe, et al., 2003). Recently, it was reported that upregulation of the KIAA1199 gene is associated with cellular mortality (Michishita, et al., 2005).

Human TMEM2 is expressed in cochlea and a variety of other tissues. It is located on the DFNB7-DFNB11 locus, a region linked to autosomal recessive non-syndromic hearing loss (ARNSHL), but no disease-causing mutations were found in TMEM2 coding region (Scott, et al., 2000).

Identification of the G8 domain should help our understanding of the structure/function of these related proteins and benefit the development of novel therapeutics.

2 METHODS

When analyzing the protein sequence of the KIAA1199 protein and its homologues, we found that they contain a glycine-rich region in the N terminus that did not match any entry in the Pfam 19.0 (Finn, et al., 2006) and SMART 5.0 (Letunic, et al., 2006) databases. Using PSI-BLAST (Altschul, et al., 1997) with an inclusion threshold of 0.05, we searched the NCBI non-redundant protein databases (http://www.ncbi.nih.gov/blast/) against the human KIAA1199 protein (amino acid residues 44-170 in gi|38638698). The search converged after five iterations and retrieved 98 non-redundant protein sequences in total. A multiple sequence alignment and phylogenetic tree of 26 distinct proteins (32 sequences) was generated using ClustalX (Thompson, et al., 1997) with manual adjustment. The alignment was colored using Chroma (Goodstadt and Ponting, 2001) (Fig. 1).

The region was named the G8 domain, since it contains eight conserved glycine residues. To predict the secondary structure of G8, the profile of the alignment was submitted to Jpred server (http://www.compbio.dundee.ac.uk/~www-jpred/submit.html) (Cuff, et al., 1998). Taxon distribution was determined by searches against all available genome and protein database at GenBank using TBLAST (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi).

3 RESULTS AND DISCUSSION

The G8 domain is about 120 amino acid residues in length. The secondary structure prediction of the G8 domain suggests that it

*To whom correspondence should be addressed.

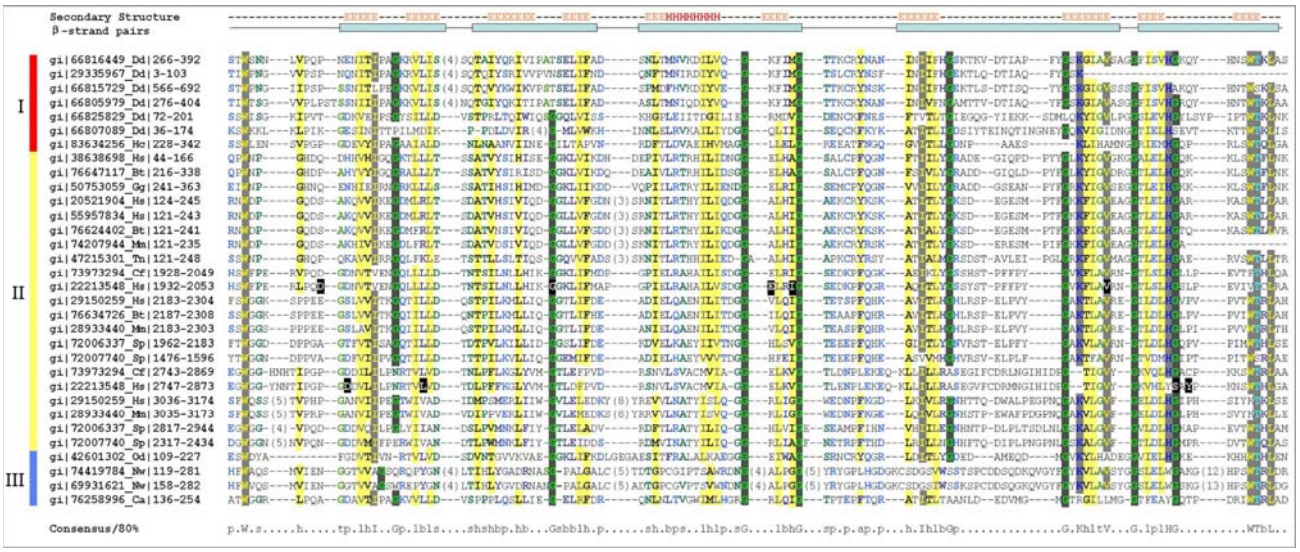

Fig. 1. Representative alignment of G8 domain, obtained by ClustalX and Chroma software. The secondary structure was predicted by JPRED tool. All of ARPKD related mutation sites in the two G8 domains of human PKHD1 protein (gi|22213548) were high lighted by black shadows. Different taxonomic groups are shown by colored lines on the left of the alignment: I, lower eukayrote; II, animal and III, bacterial. The sequences are: gi|66816449, 266-392, hypothetical protein DDB0204534, Dictyostelium discoideum; gi|29335967, 3-103, ComF, Dictyostelium discoideum; gi|66815729, 566-692, hypothetical protein DDB0215273, Dictyostelium discoideum; gi|66805979, 276-404, hypothetical protein DDB0219347, Dictyostelium discoideum; gi|66825829, 72-201, hypothetical protein DDB0201847, Dictyostelium discoideum; gi|66807089, 36-174, hypothetical protein DDB0187448, Dictyostelium discoideum; gi|83634256, 228-342, conserved hypothetical protein, Hahella chejuensis KCTC 2396; gi|38638698, 44-166, KIAA1199, Homo sapiens; gi|76647117, 216-338, similar to KIAA1199, Bos taurus; gi|50753059, 241-363, similar to KIAA1199, Gallus gallus; gi|20521904, 124-245, KIAA1412 protein, Homo sapiens; gi|55957834, 121-243, transmembrane protein 2, Homo sapiens; gi|76624402, 121-241, similar to transmembrane protein 2, Bos taurus; gi|74207944, 121-235, unnamed protein product, Mus musculus; gi|47215301, 121-248, unnamed protein product, Tetraodon nigroviridis; gi|73973294, 1928-2049|2743-2869, PKHD1 precursor, Canis familiaris; gi|22213548, 1932-2053|2747-2873, polycystic kidney and hepatic disease 1, Homo sapiens; gi|29150259, 2183-2304|3036-3174, fibrocystin L, Homo sapiens; gi|76634726, 2187-2308, PREDICTED: similar to fibrocystin L, partial, Bos taurus; gi|28933440, 2183-2303|3035-3173, fibrocystin L, Mus musculus; gi|72006337, 1962-2183|2817-2944, similar to fibrocystin L, Strongylocentrotus purpuratus; gi|72007740, 1476-1596|2317-2434 similar to fibrocystin L, Strongylocentrotus purpuratus; gi|42601302, 109-227, similar to transmembrane protein 2, Oikopleura dioica; gi|74419784, 119-281, hypothetical protein Nwi_0717, Nitrobacter winogradskyi Nb-255; gi|69931621, 158-282, hypothetical protein NhamDRAFT_0056, Nitrobacter hamburgensis X14. gi|76258996, 136-254, Blue (type 1) copper domain, Chloroflexus aurantiacus J-10-fl. This multiple sequence alignment has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/) with the accession number (ALIGN_000989).

contains ten β-strands and one helix. These strands are separated by conserved glycine residues and contain some conserved hydrophobic residues. After further examining the alignment, we found that the G8 domain is actually composed of five β-strand pairs (Fig.1). Each repeat has a sequence resembling $hX_{(0-3)}hX_{(1-3)}GX_{(1-11)}hX_{(1-3)}h$, where X is any residue and h is a hydrophobic residue. Based on the structural prediction, the conserved glycine residues and hydrophobic residues might be important for correct folding of G8 domains, the glycine residues allowing rotation in the backbone, and hydrophobic interactions among hydrophobic residues on β-strands/helix contributing to structural stabilization. The alignment also indicates some potential functionally important residues such as K2038, H2040 and T2048 in human PKHD1 protein (gi|22213548). These highly conserved polar residues cluster on the C terminus of the G8 domain and may comprise the core of its active site.

The G8 domain is widely distributed, being found in proteins from various animal (from *Strongylocentrotus purpuratus* to *Homo sapiens*), lower eukayrotes (such as *Dictyostelium discoideum* and *Tetrahymena thermophila*) and bacteria (such as the alpha-proteobacteria *Nitrobacter hamburgensis*, gamma-proteobacterium *Hahella chejuensis and the green non-sulfur bacterium Chloroflexusaurantiacus*) but absent in plants, viruses and archaea. Many G8-containing proteins are integral membrane proteins with signal peptides and/or transmembrane segments, and others lacking TM domain may be secreted (Fig. 2).

Several other protein domains ferquently co-occurr in proteins with a G8 domain. These include the IPT/TIG domain (SMART: SM00429, ig-like, plexins, transcription factors domain), the GG domain (domain in KIAA1199, FAM3, POMGnT1 and TMEM2 proteins, with two well-conserved glycine residues)(Guo, et al., 2006) and the PbH1 domain (SMART: SM00710, Parallel beta-helix repeats domain). IPT/TIG domains are found in cell surface receptors such as Met and Ron as well as in intracellular transcription factors and take a role in the control of cell dissociation, motility, invasion of extracellular matrices as well as DNA binding (Collesi C, 1997). The GG domain is widely present in eukaryotic proteins and T4 phage gp35 proteins. It was predicted to be structurally important in long tail fibers of T4 (Guo, et al., 2006), which is responsible for host cell recognition and infection and initial attachment to susceptible bacteria (Dickson RC, 1973). Known functions of the PbH1 domain include binding extracellular proteins and catalysis of polycaccharide hydrolysis (Bedford and Leder, 1999). Based on the functions of G8-associated domains

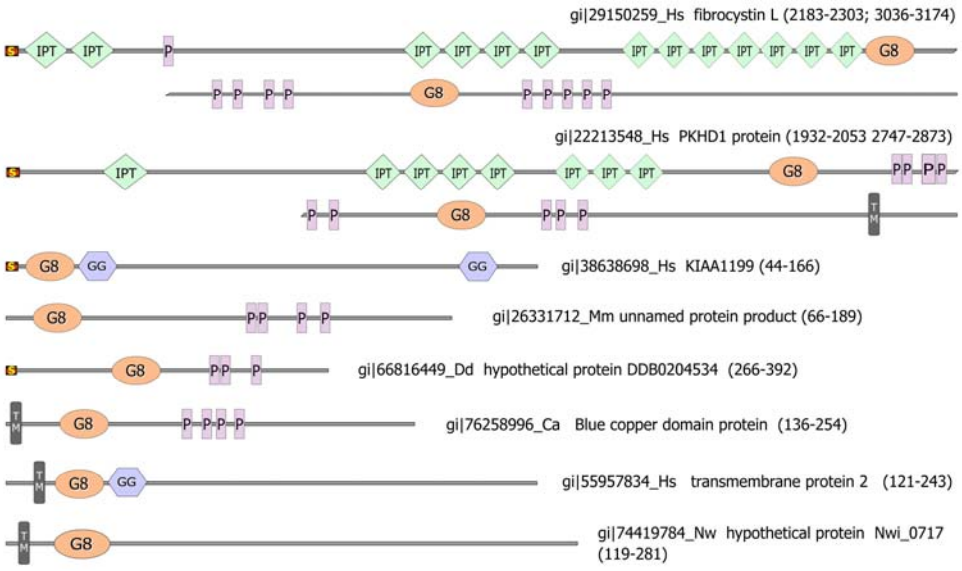

Fig. 2. Domain architecture of representative proteins with G8 domain. There are two classes of G8 containing proteins (I and II). S indicates signal peptide, TM indicates transmembrane segments, IPT indicates the IPT/TIG domain (SMART: SM00429, ig-like, plexins, transcription factors domain), P indicates the PbH1 domain (SMART: SM00710, Parallel beta-helix repeats domain) and GG indicates the GG domain (domain in KIAA1199, FAM3, POMGnT1 and TMEM2 proteins, with two well-conserved glycine residues)(Guo, et al., 2006).

and proteins, it is reasonable to predict that G8 may involve in extracellular ligand binding and catalysis processing.

Many G8-containing proteins have been associated with diseases such as polycystic kidney disease and non-syndromic hearing loss. The PKHD1 protein contains two G8 domains that, based on their high degree of sequence identity (28%) between the tandem copies of the G8 domain, probably originated from tandem duplication. Nine missense mutations in the G8 domains in human PKHD1 protein (gi|22213548) are reported to be associated with autosomal-recessive polycystic kidney disease (ARPKD) including D1942G, G1971D, E1995G, I1998T and V2032L in the first G8 domain; D2761Y, L2772P, S2861G and Y2863C in the second one (Fig. 1)(Bergmann, et al., 2004; Rossetti, et al., 2003; Ward, et al., 2002). These results show that substitution of a conserved glycine residue (G1971D) and hydrophobic residues (I1998T, V2032L and L2772P) might disrupt the proper conformation and thus lead to the depletion of the normal function of G8. Until now, no disease-causing mutation in G8 domain of KIAA1109 and TMEM2 proteins has been observed. The role of the G8 domain in non-syndromic hearing loss disease is still unknown.

CONCLUSIONS

In summary, the G8 domain is widely distributed, presenting in both animal and bacterial proteins including some hereditary disease related protein such as PKHD1, KIAA1109 and TMEM2 proteins. It contains five repeated β-strand pairs. Structural and domain architecture analysis indicates that G8 domain may be involved in extracellular ligand binding and progress of catalysis. Mutations of G8 domain in human PKHD1 protein are associated with ARPKD. Discovery of G8 domain would be important for the research of the structure/function of related proteins and benefit for the development of novel therapeutics.

ACKNOWLEDGEMENTS

The authors thank Dr Alex Bateman (Wellcome Trust Sanger Institute, UK) and Dr Jingchu Luo (Peking University, China) for suggestions and comments on the manuscript. This work was supported by the grants from National 973 project of China (2001CB5102), National Natural Science Foundation of China (30430170, 90408017) and a grant from Human Liver Proteomics Project.

REFERENCES

Abe, S., Usami, S. and Nakamura, Y. (2003) Mutations in the gene encoding KIAA1199 protein, an inner-ear protein expressed in Deiters' cells and the fibrocytes, as the cause of nonsyndromic hearing loss, *J Hum Genet*, **48**, 564-570.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Bedford, M.T. and Leder, P. (1999) The FF domain: a novel motif that often accompanies WW domains, *Trends Biochem Sci*, **24**, 264-265.

Bergmann, C., Senderek, J., Schneider, F., Dornia, C., Kupper, F., Eggermann, T., Rudnik-Schoneborn, S., Kirfel, J., Moser, M., Buttner, R. and Zerres, K. (2004) PKHD1 mutations in families requesting prenatal diagnosis for autosomal recessive polycystic kidney disease (ARPKD), *Hum Mutat*, **23**, 487-495.

Collesi C, S.M., Gaudino G, Comoglio PM. (1997) A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype, *Molecular and Cellular Biology*, **16**, 5518-5526.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server, *Bioinformatics*, **14**, 892-893.

Dickson, R.C. (1973) Assembly of bacteriophage T4 tail fibers. IV. Subunit composition of tail fibers and fiber precursors, *J. Mol. Biol*, **79**, 633 – 647.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L. and Bateman, A. (2006) Pfam: clans, web tools and services, *Nucleic Acids Res*, **34**, D247-251.

Goodstadt, L. and Ponting, C.P. (2001) CHROMA: consensus-based colouring of multiple alignments for publication, *Bioinformatics*, **17**, 845-846.

Guo, J., Cheng, H., Zhao, S. and Yu, L. (2006) GG: a domain involved in phage LTF apparatus and implicated in human MEB and non-syndromic hearing loss diseases, *FEBS Lett*, **580**, 581-584.

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J. and Bork, P. (2006) SMART 5: domains in the context of genomes and networks, *Nucleic Acids Res*, **34**, D257-260.

Michishita, E., Garces, G., Barrett, J.C. and Horikawa, I. (2005) Upregulation of the KIAA1199 gene is associated with cellular mortality, *Cancer Lett*，in Press.

RC, D. (1973) Assembly of bacteriophage T4 tail fibers. IV. Subunit composition of tail fibers and fiber precursors., *J Mol Biol.*, **79**, 633-647.

Rossetti, S., Torra, R., Coto, E., Consugar, M., Kubly, V., Malaga, S., Navarro, M., El-Youssef, M., Torres, V.E. and Harris, P.C. (2003) A complete mutation screen of PKHD1 in autosomal-recessive polycystic kidney disease (ARPKD) pedigrees, *Kidney Int*, **64**, 391-403.

Scott, D.A., Drury, S., Sundstrom, R.A., Bishop, J., Swiderski, R.E., Carmi, R., Ramesh, A., Elbedour, K., Srikumari Srisailapathy, C.R., Keats, B.J., Sheffield, V.C. and Smith, R.J. (2000) Refining the DFNB7-DFNB11 deafness locus using intragenic polymorphisms in a novel gene, TMEM2, *Gene*, **246**, 265-274.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res*, **25**, 4876-4882.

Ward, C.J., Hogan, M.C., Rossetti, S., Walker, D., Sneddon, T., Wang, X., Kubly, V., Cunningham, J.M., Bacallao, R., Ishibashi, M., Milliner, D.S., Torres, V.E. and Harris, P.C. (2002) The gene mutated in autosomal recessive polycystic kidney disease encodes a large, receptor-like protein, *Nat Genet*, **30**, 259-269.

Wilson, P.D. (2004) Polycystic kidney disease, *N Engl J Med*, **350**, 151-164.