Original Article **Asian J Agri Biol, 2014, 2(1):28-33.**

IN SILICO ANALYSIS OF SMAD4 INVOLVED IN HEPATOCELLULAR CARCINOMA

Muhammad Asif Rasheed*, Sara Afzal, Fatima Tariq, Shazia Mannan
COMSATS Institute of Information Technology, Sahiwal

ABSTRACT

Hepatocellular carcinoma (HCCa, also called malignant hepatoma) is a primary malignancy of the liver. Most cases of HCCa are secondary to either a viral hepatitide infection (hepatitis B or C) or cirrhosis (alcoholism) being the most common cause of hepatic cirrhosis. HCCa is the fifth most common cancer and the third most common cause of cancer death worldwide. Many different proteins are involved in HCCa including insulin growth factor (IGF) II **,** signal transducers and activators of transcription (STAT) 3**,** STAT4, mothers against decapentaplegic homolog 4 (SMAD 4)**,** fragile histidine triad (FHIT), SIRT etc. The present study includes the bioinformatics analysis of SMAD 4 protein in order to understand the proteomic aspect and improvement of the diagnosis of the disease. Different information related to proteins were gathered from different databases e.g. official Symbol and chromosome location were taken from National Center for Biotechnology Information (NCBI) Gene database. Structural weight (mass of the protein) was taken from Uniprot database. Length of amino acids chain was taken from NCBI Protein database. Different domains of protein were taken from InterProScan. The proteins details related to diseases were checked from Online Mendelian Inheritance in Man (OMIM) database. The multiple sequence alignments of top eight closest sequences to these proteins were performed by Claustalw2. Structure of the protein and evaluation of the quality of the structures were included from Easy modeller and Pymol programs. This analysis not only helped to gather the information related to the protein at one place but also analyzed the structure and quality of the protein. Moreover, the present study also included the phylogenetic analysis among different species based on the protein.

Keywords: Hepatocellular carcinoma, NCBI, Uniprot, InterProScan, OMIM, Pymol

INTRODUCTION

Proteomics is the study of structures and functions of proteins which are the main components of the physiological and metabolic pathways of cells. In fact, genome expression can be described by the proteins which are responsible for regulation and normal functioning of the cell (Petriction EF et al., 2002). Different cells and tissues of the body contain different proteins which may vary due to certain environmental and disease conditions. Most of the functions in an organism are performed by the macromolecular protein complexes for example DNA replication machinery, transcription machinery, protein folding apparatus and many other protein complexes (Rai AJ et al., 2004).

The cell processes, and the effect of proteins on the cell processes can be well understood through proteomics studies. It can also be helpful in studying the effect of the environment and the cell processes on the proteins. The differences between the different conditions of an organism can be highlighted and analysed by the proteomics study. Through applying the proteomics tool and identifying the proteins in disease and the healthy individuals' samples, we can identify the biomarkers to differentiate the both classes. Such discoveries can lead to protein-based diagnostic tools and better understanding of a disease state. For example, protein expression pattern or profile during cancer is different compared to a normal healthy situation. Such unique proteins which are present in the diseased condition and not in the healthy cells can be used as markers for the disease study as well as target during the disease treatment. The analysis of the proteins using a proteomic approach is very useful and has different applications. For example, proteins isolated from body tissues gives a complete idea of the tissue situation as well as provided a good basis for to study other processes such as post translational modifications, protein

*Corresponding author: e-mail: arkmsksh@hotmail.com

functionality or protein complexes (Görg A et al., 2004).

Bioinformatics introducing new algorithms in the field of proteomics to handle large and heterogeneous data sets. With the integrated use of "-omics" disciplines, the identification and the characterization of candidate genes, proteins and molecules involved in a given disease will probably represent one of the milestones of future health care (Tanke HJ., 2007). Proteomics deals with the identification of the proteins produced by cells in normal and diseased conditions, while metabolomics monitors the role of small molecules (lipids, sugars and amino acids) involved in daily cellular function (Garcia I and Tabak LA., 2008). There are approximately 30,000 genes in the human genome and the number of proteins is likely at least three times higher, as resulting from alternative splicing and posttranslational modifications (Wright JT and Hart TC., 2002).

Hepatocellular carcinoma

Hepatocellular carcinoma (HCCa) is a primary malignancy of the liver. Most cases of HCCa are secondary to either a viral hepatitide infection (hepatitis B or C) or cirrhosis (alcoholism) being the most common cause of hepatic cirrhosis (Kumar V et al., 2003). In countries where hepatitis is not endemic, most malignant cancers in the liver are not primary HCCa but metastasis of cancer from elsewhere in the body, e.g., the colon. HCCa is the fifth most common cancer and the third most common cause of cancer death worldwide, and there are few effective therapeutic options available for those suffering from advanced disease (Parkin DM et al., 2005). HCCa poses a major challenge because of its clinical heterogeneity and lack of good diagnostic markers and treatment strategies. This type of cancer occurs more often in men than women. It is usually seen in people with age of 50 or older (American Cancer Society). However, the age varies in different parts of the world. The disease is more common in parts of Africa and Asia than in North or South America and Europe.

Different up-regulated proteins during the disease include insulin growth factor (IGF) II**,** a disintegrin and metalloproteases (ADAM) 9**,** signal transducers and activators of transcription (STAT) 3**,** suppressors of cytokine signaling (SOCS) 3**,** and cyclin D1 while the down-regulated proteins during the disease include collagen I**,** SMAD 4**,** fragile histidine triad (FHIT), and SOCS1 (Tannapfel A et al., 2003). Other proteins include OCT4**,** BIRC5, CCND1, BCRP, Sox2, GST-Nck1-SH2**,** HLA-DQ**,** miR-106b**,** c-Myc, Ki67 and SIRT.

Analysis of data by different databases and software tools

The bioinformatics analysis of SMAD4 protein was performed by using different tools, software and databases related to bioinformatics. Different details related to the proteins were gathered from different databases and software including databases of National Center of Biotechnology Information (NCBI), UniProt, Online Mendelian Inheritance in Man (OMIM), Basic Local Alignment Search Tool (BLAST), Easy modeller, Pymol, ClustalW, ProtParam, and InterProScan.

FASTA sequence of the protein

The FASTA sequence of SMAD4 protein is as follow.

>gi|13603414|dbj|BAB40977.1| SMAD4 [Homo sapiens]

MDNMSITNTPTSNDACLSIVHSLMCHRQG
GESETFAKRAIESLVKKLKEKKDELDSLIT
AITTNGAHPSKCVTIQRTLDGRLQVAGRK
GFPHVIYARLWRWPDLHKNELKHVKYCQ
YAFDLKCDSVCVNPYHYERVVSPGIDLSG
LTLQSNAPSSMMVKDEYVHDFEGQPSLST
EGHSIQTIQHPPSNRASTETYSTPALLAPSE
SNATSTANFPNIPVASTSQPASILGGSHSE
GLLQIASGPQPGQQQNGFTGQPATYHHNS
TTTWTGSRTAPYTPNLPHHQNGHLQHHP
PMPPHPGHYWPVHNELAFQPPISNHPAPE
YWCSIAYFEMDVQVGETFKVPSSCPIVTV
DGYVDPSGGDRFCLGQLSNVHRTEAIERA
RLHIGKGVQLECKGEGDVWVRCLSDHAV
FVQSYLDREAGRAPGDAVHKIYPSAYIK

Chromosome location

The chromosome location of SMAD4 is 18q21.1 while the length of the amino acid chain of SMAD4 is 436 amino acids. Moreover, the molecular weight of the protein is 48085. Furthermore, different domains of SMAD4 proteins include SMAD domain, dwarfin type, MAD homology 1, dwarfin type, SMAD/FHA domain, MAD homology, MH1, dwarfin, SMAD domain like and integrated as shown in figure 1.

Basic local alignment search tool (BLAST) was used to collect eight closest sequences compare to the query FASTA sequence of SMAD4 protein in protein databank (PDB) database. Following eight closest proteins compare to query SMAD4 protein were found by using BLAST and selecting protein databank (PDB) database. The PDB codes of selected proteins with chain names were as follow:

1. 1G88_A
2. 1DD1_A
3. 3QSV_A
4. 1U7V_B
5. 1MR1_A
6. 1YGS_A
7. 1KHX_A
8. 1DEV_A

The multiple sequence alignment of the FASTA sequence of SMAD4 protein was performed by using Clustalw2 program with the sequences of eight species in order to check the relationship among them by constructing a phylogenetic tree as shown in figure 2. The sequences of these eight species were taken from NCBI protein database. The included species were *Homo sapiens (*GI:13603414), *Mus musculus (*GI:28201436), *Susscrofa (*GI:12083759), *Trichinellaspiralis (*GI:358440819), *Neovisonvison (*GI:17887367*), Ctenopharyngodonidella (*GI:313766706*), Branchiostomabelcheritsingtauense (*GI:146335604) and *Aedes aegypti (*GI:157137978).

The structure of SMAD4 protein was predicted by using protein databank 1G88A protein as template. This closest protein was selected from protein databank by using BLAST search tool. Easy modeler program was used to predict the structure of SMAD4 protein shown in figure 3.

The Ramachandran plot was drawn by using Easy modeller in order to evaluate the structure produced by the program as shown in figure 4.

OMIM Details

Zhou et al. (1998) founded that mutational inactivation of SMAD4 caused TGF-beta unresponsiveness and gave a basis for understanding the physiologic role of this gene in tumorigenesis. SMAD4 plays a pivotal role in signal transduction of the transforming growth factor beta superfamily cytokines by mediating transcriptional activation of target genes. Zawel et al. (1998) founded that human SMAD3 and SMAD4 proteins could specifically recognize an identical 8 base pair (bp) palindromic sequence (GTCTAGAC). Tandem repeats of this palindrome conferred striking TGF-beta responsiveness to a minimal promoter. This responsiveness was abrogated by targeted deletion of the cellular SMAD4 gene. These results showed that SMAD proteins were involved in the biologic responses to TGF-beta and related ligands.

Bornstein et al. (2009) founded that expression of SMAD4 was down regulated in both malignant human head and neck squamous cell carcinomas and in grossly normal adjacent buccal mucosa. Deletion of SMAD4 specifically in mouse head and neck epithelia resulted in spontaneous head and neck squamous cell carcinomas with evidence of increased genomic instability and inflammation. Davis et al. (2008) founded that SMAD4, the common SMAD required for most transcriptional responses to BMP and TGFB signaling, is not required for processing of miR21 by BMP4 in primary pulmonary artery smooth muscle cells. Ding et al. (2011) founded SMAD4 as a key regulator of prostate cancer progression in mice and humans.

DISCUSSION

Proteomics is the study of structures and functions of proteins which are the main components of the physiological and metabolic pathways of cells. The analysis of the proteins using a proteomic approach is very useful and has different applications. Despite the fact that proteomics over the time have demonstrated a very good method in the analysis of proteins, but it has some limiting factors that should be considered before to run any assay. Bioinformatics introducing new algorithms in the field of proteomics to handle large and heterogeneous data sets. The bioinformatics analysis of SMAD4 protein involved in hepatocellular carcinoma may lead to better diagnosis of the disease as well as the treatment of the disease.

The different information collected related to SMAD4 protein include the cytogenetic location, FASTA sequence, length of its amino acid chains, domains, phylogenetic analysis of the protein with eight closest homologs, modelling of 3 dimensional structure of the protein using a suitable template and evaluation of the structure by using Ramachandran plot.

Different databases, programs and software tools were used for the purpose including NCBI, BLAST, ProtParam, InterProScan, Easy Modeller, ClustalW2 and Chimera.

Protein microarrays analysis showed that the expression of SMAD4 protein was down-regulated (Tannapfel A et al., 2003) while, one analysis reported that SMAD4 protein was up-regulated in hepatocellular carcinoma (Torbenson et al., 2002). Hence, the roles of TGF-ß1-Smad4 signaling pathway in HCCa need further studies. The BLAST of the SMAD4 was performed and template of 1G88A protein from protein databank with maximum query coverage of 34% and lowest e-value of 2e-110 was selected for homology modeling of the protein. The query coverage of the template was although the maximum among the available templates but not so good for the structure to be modeled perfectly. Hence, further studies are required in order to get batter template than the template used in current studies.

The phylogenetic analysis of the human SMAD4 protein (GI:13603414) was performed with the SMAD4 proteins of different other species including *Mus musculus* (GI:28201436), *Susscrofa* (GI:12083759), *Trichinellaspiralis* (GI:358440819), *Neovisonvison* (GI:17887367), *Ctenopharyngodonidella* (GI:313766706), *Branchiostomabelcheritsingtauense* (GI:146335604) and *Aedes aegypti* (GI:157137978) as shown in figure 2. By taking *Susscrofa* as out group, the closest relationship among *Trichinellaspiralis* and *Aedes aegypti* can be found by the analysis. This may show the speciation event among them occurred closely.

In the Ramachandran plot for the evaluation of the modeled structure, many of the residues were not in the most favorite regions as shown in figure 4 showing that the modeling of the structure is not too accurate. This may be the reason because of the low query coverage of the template or the bad quality of the template. Moreover, It is concluded that bioinformatics has really made the proteomics research easy and efficient by introducing new algorithms in the field of proteomics to handle large and heterogeneous data sets. Proteins and molecules involved in a given disease will probably represent one of the milestones of future health care.

Figure 1: Different domains of SMAD4 protein: Different domains of SMAD4 proteins include SMAD domain, dwarfin type, MAD homology 1, dwarfin type, SMAD/FHA domain, MAD homology, MH1, dwarfin, SMAD domain like and integrated.

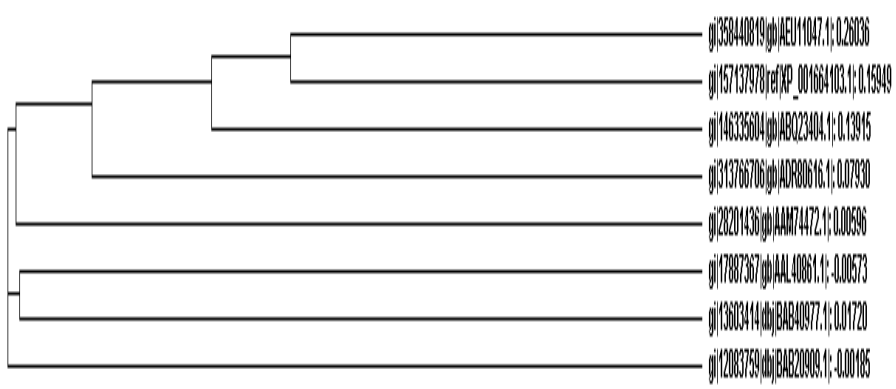

Figure 2: Cladogram: The phylogenetic analysis of the SMAD4 proteins of different species showing relationship among them. Closest relationship can be seen among *Trichinellaspiralis* and *Aedes aegypti*

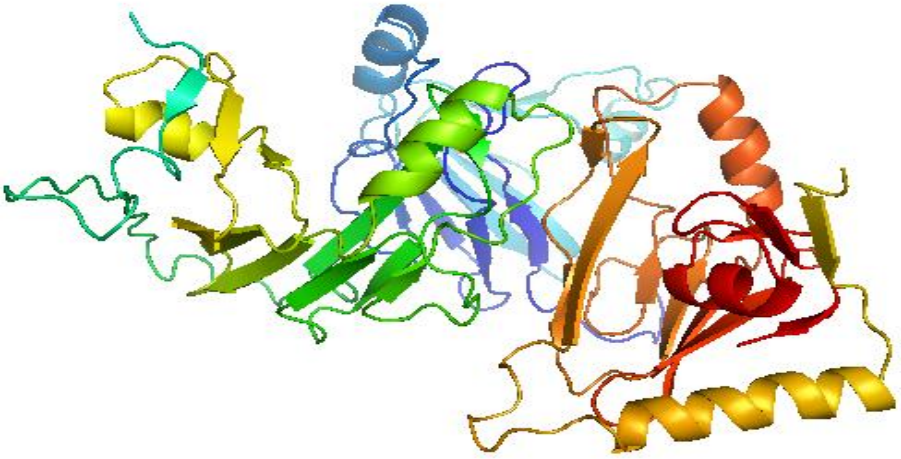

Figure 3: Easy Modeller structure prediction of SMAD4 showing alpha helices, beta sheets, loops and turns of the predicted protein

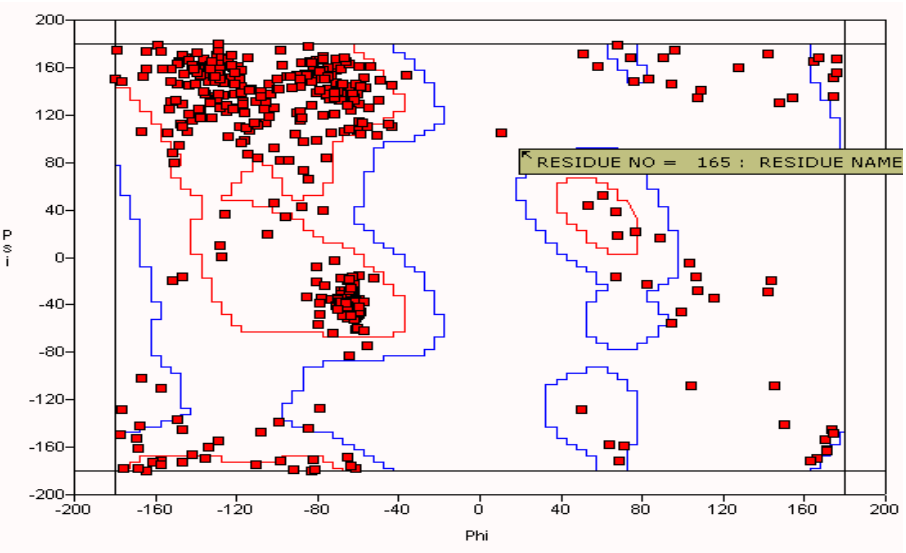

Figure 4: Ramachandran plot for SMAD4 protein: The structure evaluation of SMAD4 was performed through this plot. Many of the residues are not present in the most favorite regions showing that the quality of the protein is not so good.

REFERENCES

American Cancer Society. Cancer Facts & Figures 2013. Atlanta, Ga: American Cancer Society; 2013.

Bornstein, S., White, R., Malkoski, S., Oka, M., Han, G., Cleaver, T., Reh, D., Andersen, P., Gross, N., Olson, S., Deng, C., Lu, S.-L., Wang, X.-J. Smad4 loss in mice causes spontaneous head and neck cancer with increased genomic instability and inflammation. J. Clin. Invest. 119: 3408-3419, 2009.

Davis, B. N., Hilyard, A. C., Lagna, G., Hata, A. SMAD proteins control DROSHA-mediated microRNA maturation. Nature 454: 56-61, 2008.

Ding, Z., Wu, C.-J., Chu, G. C., Xiao, Y., Ho, D., Zhang, J., Perry, S. R., Labrot, E. S., Wu, X., Lis, R., Hoshida, Y., Hiller, D., and 16 others. SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. Nature 470: 269-273, 2011

Garcia I, Tabak LA (2008). Beyond the "omics": Translating science into improved health. J Am Dent Assoc; 139, 392-5.

Görg A, Weiss W, Dunn MJ (2004). Current two-dimensional electrophoresis technology for proteomics. Proteomics 4 (12), 3665–85

Kumar V, Fausto N, Abbas A (editors) (2003). Robbins & Cotran Pathologic Basis of Disease (7th ed.). Saunders. 914–7.

Parkin DM, Bray F, Ferlay J, Pisani P CA Cancer J Clin. (2005). Global cancer statistics, 2002. Mar-Apr; 55(2):74-108

Petriction EF, Zoon KC, Kohn EC (2002). Clinical proteomics: Translating benchside promise into beside reality. Nat. Rev. Drug discov. 1: 683-695.

Rai AJ, Chan DW. (2004). Cancer proteomics: Serum diagnostics for tumor marker discovery. Ann N Y Acad Sci: Jun; 1022: 286-94.

Tanke HJ (2007). Genomics and proteomics: The potential role of oral diagnostics. Ann N Y Acad Sci; 1098, 330-4.

Tannapfel A., Anhalt K., Hausermann P., Sommerer F., Benicke M., Uhlmann D., Witzigmann H., Hauss J., and Wittekind C. (2003). Identification of novel proteins associated with hepatocellular carcinomas using protein microarrays. J Pathol. 201: 238-249.

Torbenson M, Marinopoulos S, Dang DT, et al: Smad4 overexpression in hepatocellular carcinoma is strongly associated with transforming growth factor beta II receptor immunolabeling. Hum Pathol 33: 871-876, 2002.

Wright JT, Hart TC (2002). The genome projects: Implications for dental practice and education. J Dent Educ; 66, 659-71.

Zawel, L., Dai, J. L., Buckhaults, P., Zhou, S., Kinzler, K. W., Vogelstein, B., Kern, S. E. Human Smad3 and Smad4 are sequence-specific transcription activators. Molec. Cell 1: 611-617, 1998

Zhou, S., Buckhaults, P., Zawel, L., Bunz, F., Riggins, G., Le Dai, J., Kern, S. E., Kinzler, K. W., Vogelstein, B. Targeted deletion of Smad4 shows it is required for transforming growth factor beta and activin signaling in colorectal cancer cells. Proc. Nat. Acad. Sci. 95: 2412-2416, 1998.