Bioinformatics Analysis and Validation of Related Genes in Human Colorectal Adenocarcinoma

Lichen Wang[1,2],Yao Chen[1]

1Department of Anatomy, Biomedical and Legal Medical Institute,West China Medical Center, Sichuan University, Chengdu, Sichuan 610041, China

2Department of Anatomy, North Sichuan Medical College, Nanchong 637129, China

Abstract **Objective** Studied the relationships of four expression sequence tags (Accession numbers: ES274071, ES274070, ES274084, ES274075)with colorectal adenocarcinoma(CRA). **Methods** Selected the EST of ES274071 as a seeded sequence, utilized bioinformatics method to clone and to analysis full-length cDNA, then validated with RT-PCR method and submitted to Genbank. Cloned ES274070 ORF by RT-PCR and validated by sequencing. Took ES274084 and ES274075 by bioinformatics analysis. **Results** As ES274071, by electronic clone and DNASTAR software, assembled the matching sequence together to form a 2790bp cDNA sequence, ORF is 834bp, its coded protein is SET translocation and there is a Nucleosome Assembly Protein(NAP) domain during 29-225bp. As ES274070, ORF is 336bp length and validated by DNA sequencing technology. As ES274084 and ES274075, coded protein is probably human collagen protein type. **Conclusion** SET is a suppressor of PP2A and NM23, its NAP activity enhance the affinity of chromatin. The study also validates its feasibility of electronic clone through experimental methods and offers new ways in cloning novel genes.

Key words Bioinformatics; Colorectal adenocarcinoma; Expression sequence tag; SET; NAP

Colorectal adenocarcinoma (CRA) is one of the most comman malignant cancer in our country. It has a high disease incidence. It is accounted for the second status of malignant cancer in alimentary canal, and it is said that the incidence is being grown year by year. The present studies indicated there are multi-gene and multi-stage mutations steps in the process of occurrence and development from normal colorectal epithelial cell to carcinomatous cell. So it is very important to study genes differential expression in the CRA. With development of human genome project, protein plan and biochip, related outcome and data is being increased, especially EST database. It results in some new changes in identifying and cloning genes, then a new subject-bioinformatics is being developed.

Correspondence to: Prof. Yao Chen. Department of Anatomy, Biomedical and Legal Medical Institute, West China Medical Center, Sichuan University, Chengdu, Sichuan 610041, China
Email: xmxfh@263.net

In this paper, we filter four EST segments that the expression is different between colorectal adenocarcinoma tissue and normal colorectal tissue through SSH combined with gene chips, accession numbers are: ES274071, ES274070, ES274084, ES274075 [1]. In order to study the relationships further with CRA, the experiments take some bioinformatics analysis and preliminary biomoleculer methods such as PCR-sequence to validate.

MATERALS AND METHODS

Internet sources and bioinformatics analysis software package

http://www.ncbi.nlm.nih.gov
http://www. expasy.org
http://www.cbs.dtu.dk
http://smart.embl-heidelberg.de
DNASTAR software
Primer Premier 5.0

Tumor sample

Rectum adenocarcinoma tissue is from patient in the west China Hospital of Sichuan University, the pathology diagnosis the staging is at B phase according to Dukes method.

Regents

Trizol reagent was purchased from TianGen Company, SuperScriptIII first strand synthesis kit was purchased from Invitrogen Company, LA −Taq DNA polymerase was purchased from TaKaRa Company, multifunction recovery kit was purchased from Biotech Company. The primers were synthesized by Invitrogen Company.

Bioinformatics analysis

Obtaining full–length cDNA

Draw one differentially expressed EST segment from the cDNA subtractive library (Accession number ES274071) then choose this EST as seeded sequence and found its matching sequence, assemble and extended it as long as possible through the extension method of matching sequences blast cycling until no more matching sequences could be found. Thus, a full−length cDNA was obtained.

ORF identification, chomosome location, expression pedigree analysis

Net to http://www.ncbi.nlm.nih.gov/orf to carry on

309 atgtcggcgccggcggccaaagtcagtaaaaaggagctcaactccaaccacgacggggccgacgagacct
M S A P A A K V S K K E L N S N H D G A D E T
379 cagaaaaagaacagcaagaagcgattgaacacattgatgaagtacaaaatgaaatagacagacttaatga
S E K E Q Q E A I E H I DE V Q N E I D R L N
449 acaagccagtgaggagattttgaaagtagaacagaaatataacaaactccgccaaccatttttttcagaag
E Q A S E E I L K V E Q K Y N K L R Q P F F Q K
519 aggtcagaattgatcgccaaaatcccaaatttttgggtaacaacatttgtcaaccatccacaagtgtctg
R S E L I A K I P N F W V T T F V N H P Q V S
589 cactgcttggggaggaagatgaagaggcactgcattatttgaccagagttgaagtgacagaatttgaaga
A L L G E E D E E A L H Y L T R V E V T E F E
659 tattaaatcaggttacagaatagatttttatttttgatgaaaatccttactttgaaaataaagttctctcc
D I K S G Y R I D F Y F D E N P Y F E N K V L S
729 aaagaatttcatctgaatgagagtggtgatccatcttcgaagtccaccgaaatcaaatggaaatctggaa
K E F H L N E S G D P S S K S T E I K W K S G
799 aggatttgacgaaacgttcgagtcaaacgcagaataaagccagcaggaagaggcagcatgaggaaccaga
D L T K R S S Q T Q N K A S R K R Q H E E P E
869 gagcttctttacctggtttactgaccattctgatgcaggtgctgatgagttaggagaggtcatcaaagat
S F F T W F T D H S D A G A D E L G E V I K D
939 gatatttggccaaacccattacagtactacttggttcccgatatggatgatgaagaaggagaaggagaag
D I W P N P L Q Y Y L V P D M D D E E G E G E
1009 aagatgatgatgatgatgaagaggaggaaggattagaagatattgacgaagaaggggatgaggatgaag
E D D D D D E E E E G L E D I D E E G D E D E
1078 gtgaagaagatgaagatgatgatgaaggggaggaaggagaggaggatgaaggagaagatgactaa 1142
G E E D E D D D E G E E G E E D E G E DD ⋆

Frame	from	to	Length	
+3	309	1142	834bp	277aa

Fig. 1 Result of translation of ORF

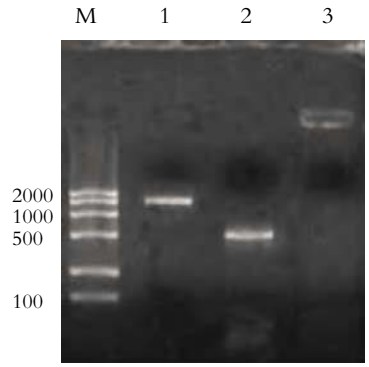

Fig. 2 RT−PCR results of ES274071 ORF
M: DL2000
lane1: ES274071 ORF fragment
lane2: β−actin
lane3: negative control

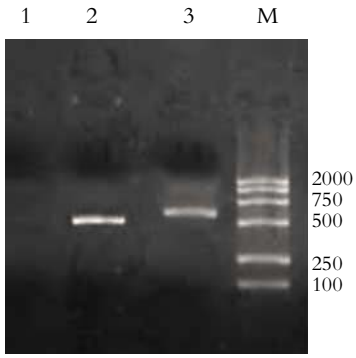

Fig. 3 RT−PCR results of ES274070 ORF
M: DL2000
lane1: negative control
lane2: β−actin
lane3: ES274070 ORF fragment

the open read−frame (ORF) identification of conting sequence;Net to http://www.ncbi.nlm.nih.gov/genome in order to acquire chomosome location; Took the assembled full−length cDNA sequence as a probe and used the Virtual Northern program (http://www.ncbi.nlm.nih.gov/SAGE/) to get expression pedigree.

Predict protein function

The similarity analysis of coded amino acid sequence was made with BlastP program provided by NCBI; the structure analysis was made with DNASTAR software; the prediction and analysis of protein´s secondary function was made with http://www.expasy.org.

Experimental validation

RNA isolation and Synthesis of first strand cDNA

Total RNA was isolated by using the Trizol reagent according to the manufacturer´s instruction. Then reverse transcription reaction was taken. Mixed 6ul ribonuclease−free water, 4μl dNTP (2.5mM), 1μl total RNA, 0.5μl Rnase Inhibitor and 2μl reverse transcription primer together, then incubated at 65℃ for 3min and cooled on ice for 2min. Then mixed 4μl 5×reaction buffer, 0.5μl Rnase Inhibitor, 1μl DTT and1μl SuperScriptIII reverse transcriptase (200U/μl)together, incubated at 50℃ for 60min. The reaction was terminated by heating at 70℃ for 15 min.

PCR amplification and sequencing

The primer pairs 5´CCTTCGCCTTCCCTTCTC and 5´CGACTGAGCACAAGAGGGA were used in clone ES274071 to amplify its ORF. The primer pairs 5´TTGGAGCCCTGAGTATCTGTG and 5´TAATGGAACCTGGTGCTAAGTC were used in clone ES274070 to amplify its ORF. Two pairs of primer were designed according to the full−length cDNA.

ES274071: PCR amplification cycles involved initial denaturation at 94℃ for 4 min; then 94℃ for 30s, 52.5℃ for 30s, 72℃ for 2min, 30 cycles; 72℃ for 7min.

ES274070: PCR amplification cycles involved initial denaturation at 94℃ for 3 min; then 94℃ for 30s, 54℃ for 30s, 72℃ for1min, 30 cycles; 72℃ for 5min. The PCR products were separated on 1.0% agarose gel and analyzed. Then recovered accordant fragment with kit and sequencing.

RESULTS

Result of Bioinformatics analysis

ES274071: ORF identification

Genbank database retrieval and DNAStar Software assembled the matching sequences together to form a 2790bp cDNA sequence. Searched with ORF Finder program to discover that initiation codon ATG existing at the location of 309−311 of the sequence; a termination codon TAA existing at the location of 1140−1142 of the low reaches,which constitutes the longest ORF at 309−1142, with a length of 834bp. (Fig.1).

Fig. 4 Result of EST ES274071 sequencing Panes express upper and lower primers

Fig. 5 Result of EST ES274070 sequencing Panes express upper and lower primers

Genetic organization expression pedigree analysis

Search of genome sequence shows that the gene Chromosome location exist in 9q34. SAGE pedigree shows that this gene has its expression the tissues of oligodendroglioma cortex, ovarian adenocarcinoma, mammary gland duct carcinoma and so on.

Struction and homology analysis encoded protein

Product of this genetic coded protein contains 277aa with the comparative molecular weight of 32101.98KD, PI: 3.97. According to Chou-Fasman principle with DNASTAR software to find that α-helix may exist during 1-14, 18-57, 91-118, 135-145, 200-212, 226-235, 239-262, 267-275. β-folding may 82-97, 103-114, 123-126, 189-196, 216-221; others are β turn and coil. Searched protein with BlastP to discover that it shares similarities with SET translocation, and there is a Nucleosome assembly Protein (NAP) domain. Protein hydrophobic analysis shows a piece of hydrophobic sequence may exist in the interval of amino acid location 75-86, 90-96, 106-108.

ES274070: Search coded protein with BlastP indata base (All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF)to discover that it shares similarities with foregone protein CAJ02118, identitie is 39%, positives is 51%, it is a hypothetical protein,

maybe a kind of tentative oxydoreductase Fe−S subunit.

ES274084: Searched with BlastP programe that offers by NCBI to discover there is a foregone NM000088 has high homology with it. Its full length is 5927bp. It is being found that the EST of ES274084 is at the location 5528 −5840 by Blast2. According to some principles to determine whether two sequences are matching[2], it is being found that ES274084 is probably a part of NM000088 or pertained to the same gene family. So we could acquired some bioinformatics information through analysis of NM000088. Its ORF is127−4521, encoded 1464aa, product is collagen1. ES274084 is located in its lower untranslated region. Chromosome location is in 17q22 (Fig. 5).

SAGE pedigree shows that this gene has its expression in the tissues of stomach, lymphoid node, pancreas, ovary, liver, neurocanal cytoma, mammary gland duct carcinoma, ovarian serous carcinoma and so on. Net to http://smart.embl −heidelberg.de/smart/show_motifs.pl to discover there is a collagen, type1, alpha1, (COL1A1) domain in1228−1464.

Results of RT−PCR and sequencing

The PCR products were separated on 1.0% agarose gel (Fig. 7, Fig. 8) PCR Recovered product were sequenced.(Fig. 2, Fig. 3)

DISCUSSION

Nowadays, the relevant research achievement and data was being increased quickly, with the accomplishment of human genome plan, the development of protein group plan and chip biotechnology, especially the quick expansion of EST. It is much more convenient obtaining and using of data due to the development of internet. It becomes a popular method to do scientific research with database, software and network source as tool. A new subject−bioinformatics comes into being and starts to play an important role in cloning novel genes and finding the function of genes.This has caused revolutionary changes to the strategy of identifying and cloning genes[3,4].

Colorectal adenocarcinoma(CRA)is one of the most common malignant cancer in the world, which is harmful to human health, but its pathogenesis is still not completely clear until nowdays. Generally thinking, the etiopathogenisis of CRA results from intrinsic factor that is the hereditary susceptibility of carcinoma of large intestine and extrinsic factor such as food and drink or environmental factor and so on[5]. It is very important to study the molecule mechanism of CRA, which could lead prevention and treatment for CRA. Meanwhile,the combination of biotechnology and information resource can speed up the process of selecting and testing novel genes that cause CRA.

In order to get full length cDNA, this study use bioinformatics method, and take the four EST fragments as seeded sequence that are selected through suppression subtractive hybridization combined with cDNA chip technique; to carry out a series of analysis ,then do experimental validation.

ES274071: Search product protein with BlastP to discover this genetic encoded protein of assembled sequence is SET translocation. SET is a suppressor of protein phosphatase 2A (PP2A)and non−metastatic gene (NM23). PP2A is a kind of phosphorylase, pertaining to tumor suppressor protein, (its overexpression may result in cellule cycle disorder [6]).It can also adjust corpuscular increment, vegetation and differentiation. NM23 is a kind of candidate tumor suppressor protein; It can restrain cell activity, repress the genesis and metastasis of tumor in situ. Its lower−expression plays an important role in infiltration and tissue differentiation in carcinoma of large intestine[7,8].

By making use of Simple Modular Architecture Research Tool (SMART) to do analysis on protein sequence [9]. The result shows there is a Nucleosome Assembly Protein (NAP) domain.NAP can make histone H2A and H2B deposit to DNA, while the compact binding of histone and double−halix DNA plays a key point role in nucleosome formation, location and the stabilization of chromatic high grade structure[10].

This study also cloned the ORF and sequenced, then submitted it to Genbank, accession number is EF534308. Compared with an existing sequence NM003011, it extends16bp at 5′.

ES274070: Predict its full length1721bp, encoded protein is a hypothetical protein, maybe a kind of ten-

tative oxydoreductase Fe−S subunit and concerned with energizing.

ES274084 and ES274075: Through analysis, it is being found that these two sequences are complementary. Its encoded protein product is collagen type1, pertained to extracellular matrix. The characteristic structure of collagen assign tissue counteract tension potential. It can also effect corpuscular morphous, provoke epithelial cell differentiation and increment. Collagen and other extracellular matrixs play a regulatory part in corpuscular morphous, increment, differentiation and locomotion [11].

This paper use bioinformatics and molecular experiment method to study the four EST segments, it will support some references to explore molecular mechanism of CRA. Besides, it also offers a new way in cloning novel genes that is the combination of bioinformatics and experimentation.

REFERENCES

1 Yao Chen (correspondent), Yi−Zeng Zhang, Zong−Guang Zhou, *et al.* Identification of differentially expressed genes In human colorectal adenocarcinoma. World of Gastroenterol, 2006, 12:1025−1032.

2 Zhang chenggang, He fuchu(chief editor). Methods and practices of Bioinformatics. The first edition. Beijing Science Press, 2002:64−142.

3 Allen J F. Bioinformatics and discovery: Induction beckons again. Bioessays, 2001, 23: 104−107.

4 Board P, Tetlow N, *et al.* Database analysis and gene discovery in pharmacogenetics. Clin Chen Lab Med, 2000, 38: 863−867.

5 Tang zhaoqiu (chief editor). Contemporary phymatology. The second edition. Shanghai Fudan University Press, 2000: 809−810.

6 Gu yanyun. The new progress of PP2A in structure and function. Overseas medicine: molecular biology, 2003, 25: 228−231.

7 Steeg PS, Bevilacqua G.Kopper L, *et al.* Evidence for a novel gene associated with liw tumor metastatic potential.[J] JNCI, 1998, 80: 200.

8 Huang zhaoquan.The expression and relationship with metaptosis of nm23 in carcinoma of large intestine. LiuZhou medical science,2003, 15: 184−185.

9 Schultz J, Milpetz F, *et al.* SMART, a Simple Modular Architecture Research Tool: Identification of signalling domains. Acda.Sci, USA, 1998, 95: 5857−5864.

10 Liu bingwen, Chen junjie (chief editor). Medical Molecular Biology. The second edition, Beijing China XieHe Medical University Press, 2005: 235−255.

11 Guo P. Up−regulation of angiopoietin−2, matrix metalloproteinase −2, membrane type −1 metalloproteinase, and laminin5γ2 correlates with the invasiven of human glioma[J]. Am J Pathol, 2005, 116: 877−890.