Analysis of the Subtractive Fragment of Human Rectal Adenocarcinoma cDNA by Means of Bioinformatics

Ying Kong[1], Yao Chen[1],

Department of Anatomy, Biomedical and legal medical Institute,
West China Medical Center of Sichuan University Chengdu, Sichuan 610041, China

Abstract **Objective** Studying the structure and function of the subtractive fragment of human rectum adenocarcinoma cDNA. **Method** From the cDNA subtractive library, collecting some fragment of suppression subtractive hybridization of human rectum adenocarcinoma cDNA and utilizing bioinformatics to extend EST sequence. **Result** Clone full-length cDNA and predict the function. **Conclusion** It offers new thought and new ways in cloning the pathogenic novel genes.

Key Words bioinformatics; expression sequence tag; contig; rectal adenocarcinoma

The initiation and development of rectum adenocarcinoma is a step-by-step process that involves multi-gene mutations. During the process, cell division and disintegration are closely related to gene cluster´s abnormal temporal expression and protein´s interactivity[1]. The identification of rectum adenocarcinoma´s special oncogene or suppression cancer gene has great significance on the study of carcinomatous change mechanism of rectum adenocarcinoma and its diagnosis, treatment and prevention. With the accomplishment of human genome sequence test, Genome research has entered into a new phase of gene abstraction and data analysis. Cloning novel genes by means of bioinformatics has become a new and developing strategy on gene cloning. This study aims to sift and clone the genes related to the rectal adenocarcinoma initiation. On the basis of the differentially expressed gene segment sequences drawn from the human rectal adenocarcinoma cDNA. Subtractive library which set up by Chen Yao[2]. The researcher makes this segment clone on Internet and tentatively predicts the segment´s chromosome location, tissue distribution and protein function.

MATERIAL AND METHODS

Cloning of gene segment related to human rectum adenocarcinoma

The differentially expressed gene segment drawn from the cDNA subtractive library is named 26 cDNA (Accession number: BM360875). Sequence is done by Shanghai Shenggong Company.

Internet sources and bioinformatics analysis software package.

Internet sources

http://www.ncbi.nlm.nih.gov/Blast
http://www.ncbi.nlm.nih.gov/Unigene/index.html
http://www.ncbi.nlm.nih.gov/SAGE/virtual northern
http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html
http://www.fruitfly.org/cgi-bin/seq_tools/promoter.pl

Analysis software package

Vector NTI Suite 7.1(informax Company)

Sifting in ″electronic″ cDNA library

Obtaining full-length cDNA

Draw one differentially expressed EST segment from the cDNA subtractive library (Accession number BM360875). The first step is to choose this EST as seed sequence and find its matching sequence. According to American Genetics Research Institute, if 2 ESTs´ alkali bases overlap beyond 40bp and share 95% of similar characters in their overlapping zone, they are matching sequences. The second step is to extend it as long as possible through the extension method of matching sequences with blast circling[5,6]. That is, assembling these matching sequences together to form a longer EST and then using the new EST to conduct blast retrieval to find more matching sequences. Repeat such process till no more matching sequences can be found. Thus, a contig is obtained.

ORF identification, chromosome location, analysis of genetic organization, and expression pedigree

analysis.

① Search with the ORF Finder program (http://www.ncbi.nlm.nih.gov/ORF Finder) offered by NCBI to carry on the open read-frame ORF identification of contig sequence. ②Make prediction on effective initiation codon of all obtained contigs. (http://125.itba.micnr.it/cgi-bin/wwwaug.pl); Select according to the principle that initiation codon in the KOZAK sequence is most suitable for translation initiation [5] Translation initiation locus (http://www.cbs.dtu.dk/services/NetStart/) prediction. PolyA prediction(http://125.itba.mi.cnr.it/cgi-bin/wwwHC_POLYA.pl. ③Directly use genome sequence to do genetic electron location. Net to: (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsBlast.html&&ORG=Hs) in order to acquire chomosome location. ④ Take the assembled full-length cDNA sequence as a probe and use the virtual Northern program (http://www.ncbi.nlm.nih.gov/SAGE/index.cgi? cmd=accsearch). Provided by CGAP (Cancer Genomic Anatomy Project) in NCBI as a tool, doing expression quantum analysis by drawing genetic representative tag from the SAGE (Serial analysis of gone expression) library.

Structure and function prediction of coded protein

Make similarity analysis of coded amino acid sequence with BlastP program (http://www.ncbi.nlm. nih. gov/Blast/Blastp) provided by NCBI; make amino acid composition analysis with http://molbiol. soton. ac. uk/cgi-bin/acomp.pl; make isoelectric point and molecular weight analysis with http://us. expasy. org/tools/pi-tool. html; make protein´s hydrophobic analysis and spatial distribution analysis of amino acid with http://www. expasy. org/cgi-bin/protscale. pl?; make prediction and analysis of protein´s secondary structure with GOR Secondary structure prediction (http:/molbiol. soton.ac. uk/compute/GOR.html); make prediction on O-linkage glycosylation locus, N-linkage glycosylation locus and phosphatase activity locus with http://www. cbs. dtu. Dk/services/Target P/.

RESULT

Genbank data base retrieval and EST sequence assembling

Refer to blastn (EST library) for the sequence BM360875 to find 19 matching sequences with great similarity; (ca453929 +bg399285 +ai366102 + aa724360+bq307348+bf971199+be047246+av709945+n47085 +bf037699 +aa971873 +bg261319 +ai796798 + h75565 +aa601927 +aa490213 +be465125 +bm789265 + bm830192=contig)

Software Vector NTI suite 7.1 assemble the matching sequences together to form a 1575bp cDNA sequence.

ORF identification

A 1575bp sequence is obtained by assembling together the matching EST sequence. Search with ORF Finder program to discover that initiation codon ATG exist in the location of 482-482 of this sequence; a termination codon TAA exists in the location of 1097-1099 of the low reaches, which constitutes the longest ORF at +2 482-1099, with a length of 618bp. and codes 205 amino acid residues; termination codon TAA exist in the location of 442-444 of the upper reaches; all the three codons are in the same read-frame. There is poly A site signal AATAAAA at the 1556bp. And further predict the effective initiation codon.

structure and homology analysis Coded protein.

Product of this genetic coded protein contains 205 amino acid with the comparative molecular weight of 23.53KD and PI value of 8.79. Search protein BLASTP in data base [All non-redundant GenBank CDS translations +PDB +SwissProt + PIR + PRF (1,476,168 sequences; 473,780,339 total letter)] to discover that it shares some similaries with TOB protein and the interplay of TOB protein and p185erb B2 has the function of restraining cell proliferation activity. So they may have similar function. Protein hydrophobic analysis (Fig.2) shows a piece of hydrophilic sequence exists in the interval of amino acid location 160--180, and a piece of hydrophobic sequence exists in the interval of amino acid location 10-20. Protein´s amino acid special distribution analysis shows that near the amino acid location 20 an amino acid segment lies in the molecular. The result of this distribution analysis furtherly proves protein´s hydrophobic analysis. The forecast of O-linkage glycosylation locus, N-linkage glycosylation locus and phosphates activity locus discovers that there aren´t O-linkage glycosylation locus and N-linkage glycosylation locus. Phosphates activity locuses exist at the amino acid locations of 67, 74, 75, 77, 78, 79, 81, 91, 97, 101, 118, 126, 129, 136, 154, 155, 156, 161. These locations are also in the beta-turn. The forecast of both phosphatase activity locus and be-

```
Length: 205 aa
   482 atgatgttcgtggcaatctgccacaggatcttagtgtttggatcg
       M  M  F  V  A  I  C  H  R  I  L  V  F  G  S
   527 acccatttgaggtttcttaccaaattggtgaacaagggaccagtg
       T  H  L  R  F  L  T  K  L  V  N  K  G  P  V
   572 aaggtgctttacgtggatgataataatgaaaatggatgtgagttg
       K  V  L  Y  V  D  D  N  N  E  N  G  C  E  L
   617 gataaggagatcaaaaacagctttaacccagaggcccaggttttt
       D  K  E  I  K  N  S  F  N  P  E  A  Q  V  F
   662 atgcccataagtgacccagcctcatcagtgtccagctctccatcg
       M  P  I  S  D  P  A  S  S  V  S  S  S  P  S
   707 cctccttttggtcactctgctgctgtaagccctaccttcatgccc
       P  P  F  G  H  S  A  A  V  S  P  T  F  M  P
   752 cggtccactcagcctttaacctttaccactgccacttttgctgcc
       R  S  T  Q  P  L  T  F  T  T  A  T  F  A  A
   797 accaagttcggctctaccaaaatgaagaatagtggccgtagcaac
       T  K  F  G  S  T  K  M  K  N  S  G  R  S  N
   842 aagggtgcacgtacttctcccatcaacctcggcttgaatgtgaat
       K  G  A  R  T  S  P  I  N  L  G  L  N  V  N
   887 gacctcttgaagcagaaagccatctcttcctcaatgcactctctg
       D  L  L  K  Q  K  A  I  S  S  S  M  H  S  L
   932 tatgggcttggcttgggtagccagcagcagccacagcaacagcag
       Y  G  L  G  L  G  S  Q  Q  Q  P  Q  Q  Q  Q
   977 cagccagcccagccgccaccgacaccaccaccaccacagcagcaa
       Q  P  A  Q  P  P  P  T  P  P  P  P  Q  Q  Q
  1022 caacagccgaaaacctctgctctttctcctaactgccaaggaact
       Q  Q  P  K  T  S  A  L  S  P  N  C  Q  G  T
  1067 ttattctttcctacatatgcagggtcaaggtag 1099
       L  F  F  P  T  Y  A  G  S  R  *
    Frame  from  to Length
    +2  482 .. 1099 618
```

Fig.1 Results of translation of ORF

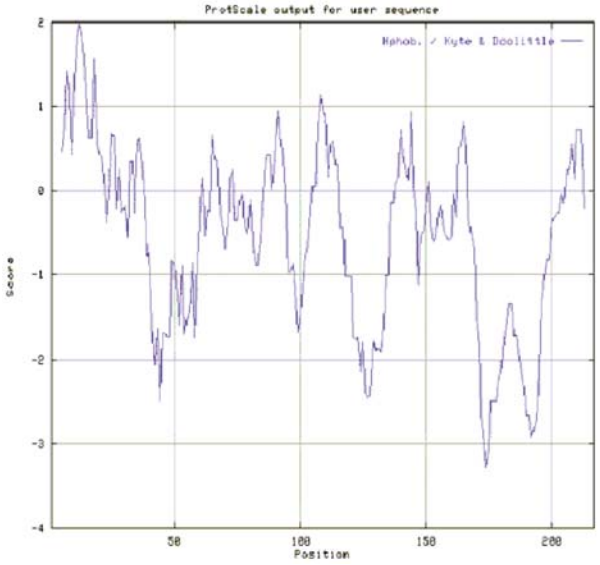

Fig.2 hydrophobic analysis

ta-turn is in accordance with the principle that in the beta-turn, phosphates activity locuses usually exist [10].

Beta–turn forecast: Fig.3 shows that beta–turn may exist near the amino acid locations of 45, 75, 125,185.

Prediction and analysis of protein´s structure

According to chou-Fasman´s prediction on protein´s substructure, 41 amino can form 2-helix. The

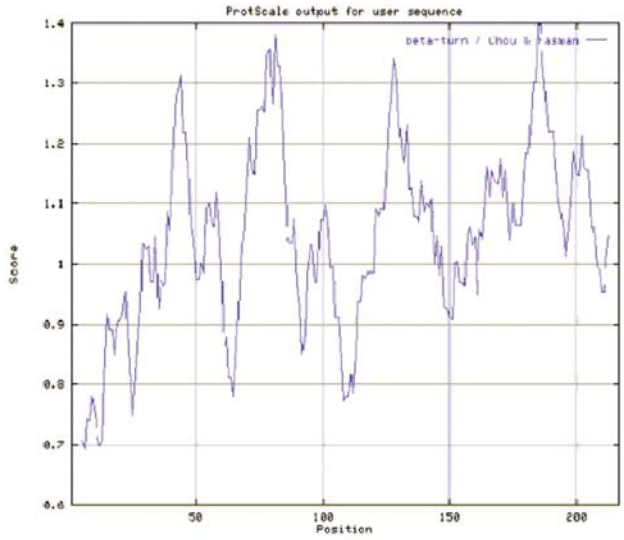

Fig.3 beta–turn forecast

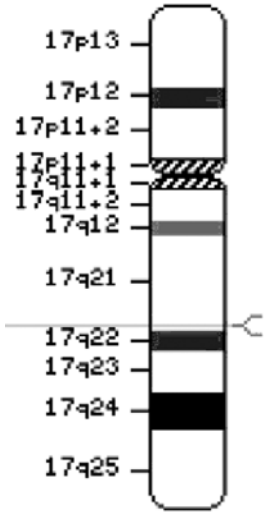

Fig.4 The position of 26 cDNA

protein´s secondary structure was forecast and analysis according to Chou-Fasman´s forecast of protein´s secondary structure. The result is that, in this protein, 41 amino acids compose α-helix, which are the 1–8, 18–21, 42–50, 101–109, 134–143 amino acid residues respectively, with a proportion of 21.7%; 47 amio acid residues compose β-sheet structure, which are the 9-12, 30-35, 97-100, 124-133, 144-152, 18-184, 195-200 amino acid residues respectively, with a proportion of 24.9%; the remaining are β –angles and irregular coiled coil.

Protein sequence profile and structure-function area analysis.

Make analysis by netting to (http://www.isrec.isb−sib.ch/software/PFSCAN-form.html) and the result suggests that an area rich in glutanic acid in the protein sequence locates at the amino acid location of 170-178.

Make use of simple Modular Architecture Research Tool (SMART) and net with (http:// smart.embl-heidelberg. de)to do analysis on protein sequence. The result shows that a PFAM: Anti-Proliferate structure area locates in the interval of amino acid location 1-83, with the similar function of Anti-proliferate protein.

Chromosome location

Search of genome sequence shows that the gene exists in 17q 21-22. As is shown by Fig 4:

Genetic organization expression pedigree analysis

SAGE pedigree shows that this gene has its expression the tissues of mammary gland carcinoma; prostate carcinoma, pancreas carcinoma, gastric, colon carcinoma, liver, white blood cells, ovary carcinoma and so on as is shown by Fig.5.

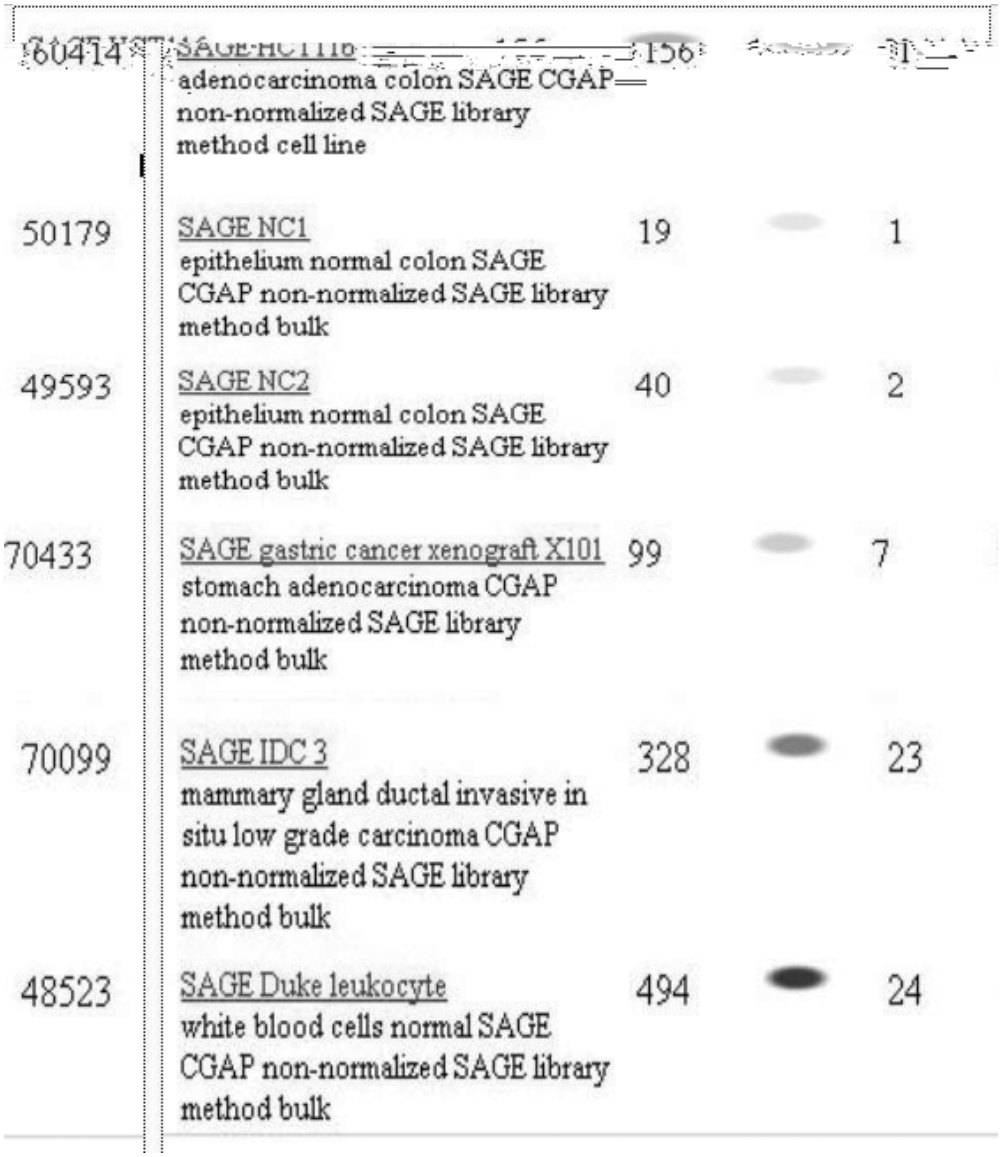

Fig.5 The expression pattern of SAGE tag of 26 cDNA

DISCUSSIONS

With the accomplishment of human genome plan, the development of protein group plan and chip biotechnology, more and more genome information is released; relevant research achievement and data increase quickly; various data bank analysis methods have been making improvements and progress; the development of internet makes the obtaining and using of data much more convenient. It is becoming a popular method to make scientific research with database as its experimental materials and software and network source as tool. A new subject comes into being and starts to play an increasingly important role in cloning novel genes and function research for its characters of rapid development. Especially the quick expansion of EST has caused revolutionary changes to the strategy of identifying and cloning novel genes.

There are many elements and steps that bring forth rectum adenocarcinoma. Although it has been confirmed that there must be some abnormal changes in the case of rectum adenocarcinoma, little further information can be detected till now. Tests on abnormal genes are helpful to the diagnosis and treatment of rectum adenocarcinoma. Mean while, the combination of biotechnology and information resource can speed up the process of selecting and testing novel genes that cause rectum adenocarcinoma. Based on means of bioinformatics, this research obtain the full-length cDNA by taking the EST segment (Accession number: BM360875) that comes from the cDNA subtractive library of human rectum adenocarcinoma as seed sequence and assembling them together.

There is a maximal efficient read frame which plays a vital role in deciding the completeness of contig. By 6 phasic-read-frame prediction together with initiation codon ATG, termination codon TAA, TGA, TAG, many read frames can be obtained by only searching one sequence. Determination of the efficient maximal reading frame principle is: ① The length usually is required to be larger than 100 lodons, and the largest is the usual choice. ② kozak sequence is A/GNNATGG. This sequence existence is most suitable for translation initiation. In this sequence, the third nucleotide in the 5′direction of ATG upper reaches usually is purine, and mostly is A (specifically-3A). The nucleotide closely following ATG usually is purine too, and in most cases is G(+4G). Experiments prove that the utilization ratio of ANNATGN and GNNATGpu is the highest. However, the nucleotide sequence near ATG without initiation function has no such conservativeness. ③ CpG islands exist in many genes′ upper reaches of vertebrate. They have a length of about 1kb. It′s GC content is higher than the average content of the whole gene. 56% of human genes have CpG island in the upper reaches. The appearance of such a special sequence in vertebrate usually indicates that most probably there is a gene in the down reaches. ④ poly A site signal is in existence. ⑤ whether there is termination codon in the upper reaches, if there is usually this reading–box is the maximal. Acording to these principles, analysis proves that contig is the complete cDNA, the genetic proofs are the following: ① Initiation codon is in the kozak sequence and the forecast value of ATTGATGAT is 75. ③ At the 1556bp, 3′UTR has polyA site signal AATAAA. ④ In the upper reaches of initiation codon ATG, there is the same-box termination codon TAA at 422-444bp, which may show the reading-box is the maximal. ⑤ Gene location is explia4, which is 17q21-22, Both the efficient initiation codon forecast and translation initiation locus forecast prove 482bp is the initiation of officient reading frame.

This paper makes an analysis on translated protein′s amino acid composition and conducts retrieval analysis on isoelectric point, molecular weight. hydrophobicity and Blast homology. Some tentative predictions are made on the protein secondary. Though the secondary prediction can′t fully consider protein′s middle and long range interactivity, this result still offers much structure information. What′s more, secondary prediction can reflect the structural trend of partial sequence segment. Secondary structure prediction still plays an important role in research of structural molecular biology [10]. Analysis proves that homologous protein shares the greatest similarity [107-109 (59%)] with anti-proliferation protein on their structure and function. Besides anti-proliferate structural zone has similar function with anti-proliferation protein. This implies that gene–coded protein has the anti–proliferation function. Decrease in cell expression of gene–coded protein may lead to improper cell proliferation[9].

As a new and developing discipline, bioinformatics is founded on the basis of network and database. Both network speed and database accuracy determine work efficiency and precision. Though having great function of synthesizing and analyzing,

softwares have their own limits in practical application. The same material could be analyzed with different softwares and thus bring forth somewhat different result, which makes experiments much more complicated. Meanwhile, not all the EST segments can be cloned and analyzed on internet. Despite all the disadvantages, this is a quick and efficient research method for its guide on planning experimental work. With further development of bioinformatics and improvement of database, prediction on structure and function of gene would become more and more accurate, and the relationship between bioinformatics and molecular biology would get closer and closer, which would give better guidance on experiments.

REFERENCE

1. Wu Naihu (chief editor). Genetic Engineering Principles (first volume). The second edition. Beijing: Science Press, 1998.
2. Chen Yao, Zhang Yizheng. The construction of cDNA suppression subtractive library of human rectum adenocarinomas. U.S Chinese Journal of lympholgy and oncology, 2002, 1(1): 9–14.
3. Diatchenko L, Lau YF, Campboll AP, et al; suppression subtiactive hybridization: a method for generating differentially regulated or tissue–specific cDNA probes and libraries. Proc Natl Aced Sci USA, 1996, 93(12): 6025–6030.
4. T K Attwood, D J party-smith. Translated by Jingchu luo, et al. Introduction to Bioinformatics, the first edition. Beijing: Peking University press, 2002. 32–112.
5. Zhang chenggang, He Fuchu (chief editors). Methods and practices of Bioinformatics. the first edition. Beijing: Science Press, 2002. 64–142.
6. Wang Zhe (chief editor). Introduction to Bioinformatics. the first edition. Xi´an: The fourth Militany University press, 2002. 28–155.
7. Bi Meixia, Cloning and Identifying the Esophagus Carcinoma Related Gene ECRG–4 in the Internet Laboratory. Biochemistry and Biophysics Journal, 2001, 33(3): 257–261.
8. T. A. Brown, transtated by Yuan Jiangang, et al. Geneme. the first edition. Beijing: science Bress, 2002. 98–294.
9. Maekawa M, Nishida E, Tanoue T. Identification of the Anti-proliferative protein Top as a MAPK substrate. J Boil Chem, 2002, 277(40): 37783–37787.
10. Yan Longfei (chief edifor). Molecule structures of Protein. The first edition. Beijing: TsingHua uniuersiy Press, 1999. 31–211.