International Journal of Computer Science & Applications

Vol. III, No. I, pp. 26 - 39

An Integrated System for Cancer-Related Genes Mining from Biomedical Literatures

Shih-Nung Chen[1] Kuo-Cheng Wen[2]

[1]Department of Computer Science and Information Engineering, Asia University, Taiwan 413, R.O.C.

[2]Department of Bioinformatics, Asia University, Taiwan 413, R.O.C.

Abstract

According to statistics, the rate of having cancer is relatively high for people in developing and developed countries. So cancer can be called as the enemy of human health. With the destruction of ecological environment and with the change of life style and diet, the rate of having cancer during the life of a general individual is one and a quarter and grows continuously by year. Thus, it is important to study the cause of cancer. However, while biomedical researchers search and retrieve biomedical literature, there is a problem of information overloading. Therefore, to survey thousands and hundreds of biomedical literatures by manual browsing is time wasting and difficult to make sure whether required information can be found. Besides, artificial inattention may lead to incorrect analysis information. The purpose of this study is to develop an integrated system for collecting and mining cancer-related gene. This system uses PubMed search engine of NCBI to search and retrieve biomedical literature and sequence of cancer-related gene. Users can combine cancer name and LOH (loss of heterozygosity), or cancer name and CGH (comparative genomic hybridization) to search, retrieve, collect, classify, and extract the biomedical literatures. This system can extract important information to accelerate the study and save plenty of time for biomedical researchers. Besides, this system can also be used on other diseases.

Keywords: cancer-related gene, loss of heterozygosity (LOH), comparative genomic hybridization (CGH), intelligent agent system, data mining

1. Introduction

According to the statistics made by the World Health Organization (WHO), in every year there are approximately 10 million people in the world suffering from all kinds of cancer and about 5 million people deceasing from it. Therefore, it is very important to speed up the study on the cause of cancer. Cancer is a disease arising from uncontrolled cell division and abnormal proliferation of cells. To observe from the aspect of molecular biology, cancer is a gene disease caused by gene's abnormal expression of translocation, amplification and deletion. These gene expressions trigger the instability of DNA and gene amplification and deletion are produced in large area of chromosome region, and lead to a serial of activation of oncogenes and inactivation of tumor suppressor genes. The detection of these two genes can be found out by two methods: loss of heterozygosity (LOH) and comparative genomic hybridization (CGH).

Bioinformatics has rapidly developed in several years. The finishing of genome sequence decoding in Human Genome Project is helpful for the discussion of gene identification, gene regulation, human variation, chromosome structural analysis, and disease heredity factors. Moreover, it is very anticipating for analyzing cancers and the cause of other diseases with a more sufficient method, and in further to use in the development of drugs. These related biological studies are usually stored in MEDLINE or various biomedical literature databases, and there are correlations and influences between various genes and diseases. Therefore, to find correlations from thousands and hundreds of biomedical literatures by manual browsing is time wasting and because of the big volume of information, it is difficult to make sure whether required information can be found. Besides, artificial inattention may lead to incorrect analysis information. This study proposes an integrated system to search, retrieve, collect, classify, and extract biomedical literature by an intelligent agent system. Furthermore, improved access to biomedical literatures and extraction of information from biomedical literatures so as to offer to biomedical researchers for the following analyses or experiments in order to find the cancer-related genes (oncogenes or tumor suppressor genes), and provides the correlations between cancers, between cancers and chromosome regions, between cancers and genes, and gene information analyzed from biomedical literatures.

2. Related Research

With the sequencing of human genome, there are more and more people involving in bioinformatics research. The following are some examples of related researches on biomedical literature:

1. Corney presented BioRAT, a new information extraction (IE) tool, specifically designed to perform biomedical IE. It can locate and analyse both abstracts and full-length papers, and incorporates document search ability with domain-specific IE. This study uses the proposed method to analyze the proportion of keyword in both abstracts and full-length papers [1].
2. Uramoto developed TAKMI, an interactive text mining system, and uses natural language techniques to extract deeper relationships among biomedical concepts from MEDLINE abstracts [2].
3. Daraselia presented MedScan, a completely automated natural language processing-based information extraction system. By using natural language processing, extracts interactions between human proteins from MEDLINE abstracts [3].
4. Schuemie analyzed a set of biomedical full-text articles. Different keyword measures indicate that information density is highest in abstracts, but that the information coverage in full texts is much greater than in abstracts. Analysis of five different standard sections of articles shows that the highest information coverage is located in the results section [4].
5. Chang developed GAPSCORE, to identify gene and protein names in text. GAPSCORE scores words based on a statistical model of gene names that quantifies their appearance, morphology and context [5].

The above-mentioned studies are to extract important messages from the abstract or the full-text article, but there is no any further analysis on these messages for acquiring useful information. Although semantic analysis of document is undergone in the TAKMI system developed by Uramoto, but the result is too complicated. The proposed system of this study can classify the collected biomedical literatures and extract information from them to find cancer-related genes. An interactive two-dimensional correlation assay is also used in the system to conduct analysis to information for offering simplified and important results.

3. Methodology

The purpose of this study is to demonstrate an integrated system for cancer-related gene research so as to help biomedical researchers rapidly and conveniently search and retrieve important information from huge biomedical literatures. According to established rules, the collection of related biomedical literature is conducted automatically from MEDLINE database, and through the optimizing classification, data mining is used to these biomedical literatures for extracting important information. Further, an interactive two-dimensional correlation assay is also used in the system to conduct analysis to information for offering simplified and important results. The simplified results enable biomedical researchers to operate conveniently, and rapidly acquire needed information for proceeding following researches and experiments. The following is to introduce the source of biomedical literature used in this study and study issues for reaching the above-mentioned purpose:

3.1 Biomedical Literature Database

PubMed, available via the NCBI Entrez retrieval system [6], was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), located at the National Institutes of Health (NIH). PubMed was designed to provide access to citations from biomedical literature, and most of the bibliographic information is from MEDLINE. MEDLINE is the NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. MEDLINE contains bibliographic citations and author abstracts from more than 4,800 biomedical journals published in the United States and 70 other countries. The database contains over 12 million citations dating back to the mid-1960's. Through PubMed, users can match keywords with Boolean operators, AND, OR or NOT, to search and retrieve related biomedical literatures in MEDLINE. For gene information, Entrez Gene is NCBI's database for gene-specific information. Entrez Gene offers integrated sequence, positioning, classification and a search system of structural data for NCBI users, and also offers the visualization of sequence and chromosome profile.

3.2 Intelligent Agent System

Intelligent agent is a software system with awareness, reactivity, mobility, cooperation, intelligence and autonomy. It is used to assist human to process information. According to users' requirements, it will automatically help users to collect information on the Internet [7] [8]. In this study, we have a large amount of biomedical literature need to search, retrieve, collect, classify, and extract useful information from these unorganized biomedical literatures for the front-end users rapidly and automatically. Therefore, regarding to the features of intelligent agent, an intelligent agent system responding to our demands is established. Its features and functions are respectively explicit as follows:

1. Automatic literature collection: According to users' query string, automatically search, retrieve, collect, and classify biomedical literatures.
2. Automatic data mining: Automatically extract useful information from retrieved biomedical literatures for the front-end users.
3. Correlation analysis: Provide a two-dimensional correlation assay for observing the correlation among cancer and cancer, cancer and chromosome region as well as cancer and gene.
4. Automatic literature renew: Automatically and regularly connect to PubMed and Entrez Gene database for obtaining the latest published biomedical literatures and information of chromosome and gene to renew the local database.

3.3 Biomedical Literature Preprocessing and Classification

There may exist many unnecessary words in retrieved biomedical literature, so this study use a parsing algorithm to remove these words so as to avoid affecting the classification and mining of biomedical literature. First, we build a word exclusion set. Based on the word exclusion set to delete some unmeaning words, such as for, the, of, he, she…etc. and differentiate the capital letter and small letter, removing punctuation, and –s at the end of words. Through preprocessing, a new literature will be generated, and according to the kinds of cancer to classify it.

In the part of literature classification, a decision tree is used in this study. A decision tree whereby convention the first or root node is displayed at the top,

connected by successive (directional) links or branches to other nodes. There are similarly connected until we reach terminal or leaf nodes, which have no further links.

The classification of a particular pattern begins at the root node, which asks for the value of a particular property of the pattern. The different links from the root node correspond to the different possible values. Based on the answer, we follow the appropriate link to a subsequent or descendent node. In the tree, the links must be mutually distinct and exhaustive; that is, one and only one links will be followed. The next step is to make the decision at the appropriate subsequent node, which can be considered the root of a subtree. We continue this way until we reach a leaf node, which has no further question. Each leaf node bears a category label, and the test pattern is assigned the category of the leaf node reached [9].

3.4 Biomedical Literature Mining

There may exist correlation between genes, between disease and gene, or the appearing time of gene in the literature, and which also can be found or its weight can be calculated with the times of appearing. It can be taken as the importance of gene and can be used for further study of biomedical researchers. This study is conducted base on a standard formula proposed by Salton to offer a new two-dimensional correlation assay. Salton's standard formula is demonstrated below [10]:

$$new\_tf(i) = 0.5 + 0.5 \times \frac{tf(i)}{\max tf} \tag{1}$$

$$wt(i) = new\_tf(i) \times \log \frac{N}{n} \tag{2}$$

$$wq(i) = \frac{0.5 + 0.5 \times qf(i)}{\max tf} \times \log \frac{N}{n} \tag{3}$$

$$\text{Re}levance\ Score = \frac{\sum_{i=1}^{T} (wq(i) \times wt(i))}{\sqrt{\sum_{i=1}^{T} wq(i)^2 \times \sum_{i=1}^{T} wt(i)^2}} \tag{4}$$

where N = total number of literatures in the collection

T = total number of terms in the collection

n = total number of literatures that contains the term

$tf(i)$ = the frequency of term i in the literature

$qf(i)$ = the frequency of term i in the query

$maxtf$ = the maximum term frequency for any term in the collection

4. Implementation and Result

4.1 System Architecture

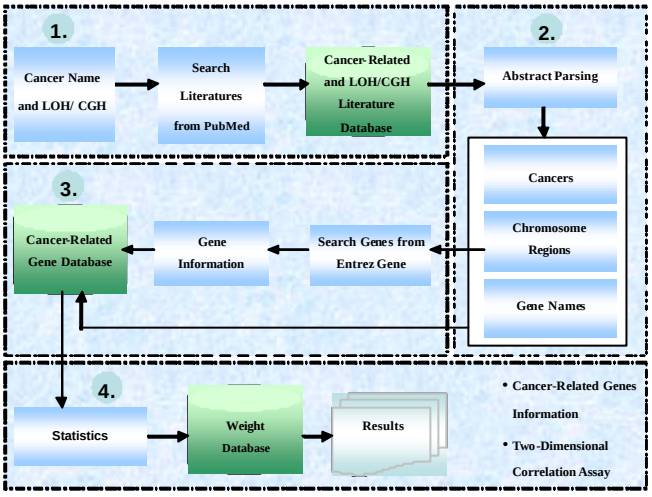


Figure 4.1 System Architecture

The purpose of this study is to demonstrate an integrated system for cancer-related gene research. This system is a bio medical literature mining system based on intelligent agent technology, can extract cancer-related genes from a large amount of bio medical literature, and examine its feasibility for proceeding following researches and experiments. This study designs several models to reach this purpose and demonstrate different results with web page, as shown in Figure 4.1. The function of each model is respectively explicated as follows:

1. Use cancer name and LOH, or cancer name and CGH as keywords to query PubMed. There is an index number in each bio medical literature, called PMID. When retrieved related biomedical literatures, PMID list is stored according to different keywords. The abstract of each bio medical literature is obtained and stored in the database automatically. When renewing, the system will compare

with PMID list in the database and requery PMID list through PubMed as a renewing reference.

2. After collecting biomedical literature automatically, this system uses a parsing algorithm to delete some meaningless words in biomedical literature. Then change capital letters into small letters, removing punctuation, and –s at the end of words. With the parsing algorithm, literature classification is proceed on the new generated literature with the decision tree, and stored in the database.
3. According to the chromosome region to query Entrez Gene, information of all genes in the chromosome region can be obtained. Then the system will store all the Gene ID, gets back related gene information and store in the database.
4. Literature mining is conducted to the processed biomedical literatures for finding cancer-related genes for all cancers. By calculating the relevance score as weights of every cancer, chromosome region, and gene of LOH and CGH in the collected abstracts. Afterwards, this information will be used for cross match in single cancer analysis.

4.2 Cancer-Related Gene Analysis

This part is mainly to offer access to information analyzed from the literature, including cancer name, chromosome region and gene information. Its detailed function is explicated as follows:

1. As shown in Figure 4.2, the system offers an information retrieval interface. Users can select cancer name and LOH, or cancer name and CGH to generate a query string. According to the query string, system will show the retrieved PMID, chromosome region appeared LOH or CGH in each literature, and calculate a weight to each chromosome region as its importance.

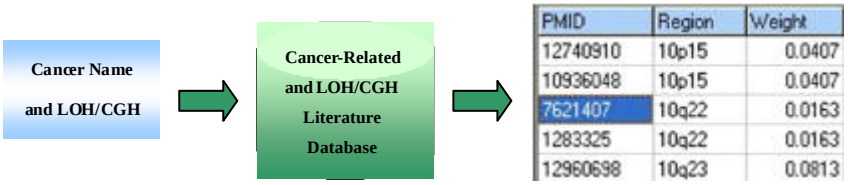
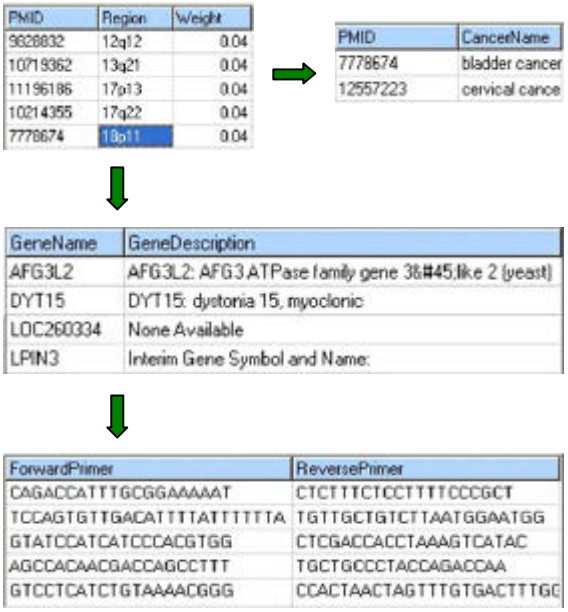


Figure 4.2 Information Retrieval Interface



PMID	Region	Weight
9828832	12q12	0.04
10719362	13q21	0.04
11196186	17p13	0.04
10214355	17q22	0.04
7778674	18p11	0.04

PMID	CancerName
7778674	bladder cancer
12557223	cervical cance

GeneName	GeneDescription
AFG3L2	AFG3L2: AFG3 ATPase family gene 3-like 2 (yeast)
DYT15	DYT15: dystonia 15, myoclonic
LOC260334	None Available
LPIN3	Interim Gene Symbol and Name:

ForwardPrimer	ReversePrimer
CAGACCATTTGCGGAAAAAT	CTCTTTCTCCTTTTCCCGCT
TCCAGTGTTGACATTTTATTTTTTA	TGTTGCTGTCTTAATGGAATGG
GTATCCATCATCCCACGTGG	CTCGACCACCTAAAGTCATAC
AGCCACAACGACCAGCCTTT	TGCTGCCCTACCAGACCAA
GTCCTCATCTGTAAAACGGG	CCACTAACTAGTTTGTGACTTTGC

Figure 4.3 Query Results

2. The query results divide into three parts, as shown in Figure 4.3:
   (a) indicate PMID and cancer name correlated with selected chromosome region
   (b) indicate all the gene name, gene annotation, and sequence in selected chromosome region
   (c) indicate forward primer and reverse primer in selected chromosome region

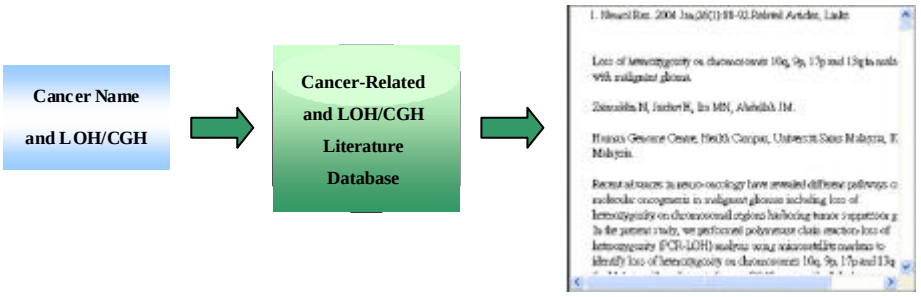


Figure 4.4 Biomedical Literature Viewer

Figure 4.5 Built-in Search Engine

3. As shown in Figure 4.4, the system offers a biomedical literature viewer for all the abstracts.
4. As shown in Figure 4.5, the system built-in a search engine. The search engine can accept any keywords to query numbers appeared and position of the keyword in all the abstracts.

In Figure 4.3, we take chromosome region 18p11 as the example. First, there will be a weight as the importance for the chromosome region. After clicking on it, the system will retrieve the information related to the chromosome region such as other cancers cited the same chromosome region and the literature's PMID list, all the genes in the chromosome region, forward primers, and reverse primers to researchers with different needs.

4.3 Two-Dimensional Correlation Assay

By using relevance score of cancer, chromosome region, and gene derived from different cancers in the statistics, this study adds the whole relevance score into a raking list, and designs a following ranking list formula based on the formula proposed by [11]:

$$TCRS = \sum_{i=1}^{T} RS(i) \tag{5}$$

$$TTRS = \sum_{j=1}^{C} RS(j) \tag{6}$$

$$TRS = \sum_{i=1}^{T} TTRS \tag{7}$$

$$Weight = F_1 \times P + F_2 \times RS + F_3 \times \frac{TRS}{TCRS + TTRS} \tag{8}$$

$$F_1 + F_2 + F_3 = 1 \tag{9}$$

where T = total number of terms in the collection

C = total number of cancers in the collection

P = the frequency of LOH or CGH of a chromosome region in the single literature

RS (Relevance Score) = the relevance score of single literature

TCRS (Total Cancer Relevance Score) = the relevance score for all term in single cancer literature

TTRS (Total Term Relevance Score) = the relevance score for any term in all cancer literature

TRS (Total Relevance Score) = the relevance score of total literature

Formula (5) is the relevance score for all cancer, chromosome region and gene in single cancer literature. Formula (6) is the relevance score of any cancer, chromosome region and gene in all cancer literature. Among them, *RS* is calculated from Formula (4). Formula (7) is the relevance score of total literature. Formula (8) is to calculate the weight of certain cancer, chromosome region and gene as its importance. F_1, F_2, F_3 is coefficient. Furthermore, Formula (9) allows users to adjust the coefficient base on its importance. Default value F_1=0.4, F_2=0.3, F_3=0.3. At present, collect amounts of cancer, chromosome region and gene respectively are 59, 3270 and 18692. Two-dimensional correlation assay is conducted to cancer, chromosome region and gene, and analysis results are shown as Table 4.1, Table 4.2 and Table 4.3. Because of the exceeding volume of information, only 5 items of information are randomly taken.

The result of two-dimensional correlation assay can be used in further analysis by researchers. They can focus on certain cancer to observe the relevance weight with other cancers. If weight is high, the relevance is also high. The relevance between cancers and chromosome regions and between cancers and genes can be examined in the same method.

Table 4.1 Two-Dimensional Correlation Assay between Cancer and Cancer

	astrocytoma	bladder cancer	brain tumor	breast cancer	carcinoid tumor	cervical cancer
astrocytoma	75.72%	00.00%	64.56%	00.00%	00.00%	00.00%
Ependymoma	74.07%	00.00%	80.60%	00.00%	00.00%	00.00%
cholangiocarcinoma	00.00%	00.00%	00.00%	00.00%	00.00%	00.00%
melanoma	00.00%	00.00%	80.23%	80.52%	00.00%	00.00%
papilloma	00.00%	00.00%	7.99%	7.85%	00.00%	12.96%

Table 4.2 Two-Dimensional Correlation Assay between Cancer and Chromosome Region

	astrocytoma	bladder cancer	brain tumor	breast cancer	carcinoid tumor	cervical cancer
10p11	00.00%	31.20%	00.00%	00.00%	00.00%	31.58%
11q13	30.86%	36.00%	31.32%	34.38%	00.00%	34.74%
11q21	00.00%	00.00%	00.00%	00.00%	00.00%	34.74%
11q1	00.00%	31.20%	00.00%	00.00%	00.00%	00.00%
10q25	30.86%	31.20%	30.44%	30.22%	00.00%	00.00%

Table 4.3 Two-Dimensional Correlation Assay between Cancer and Gene

	astrocytoma	bladder cancer	brain tumor	breast cancer	carcinoid tumor	cervical cancer
CDK4	34.29%	00.00%	33.09%	00.00%	00.00%	00.00%
EGFR	38.57%	00.00%	34.41%	33.72%	00.00%	00.00%
CCND3	30.86%	00.00%	30.44%	00.00%	00.00%	00.00%
SIL	00.00%	00.00%	00.00%	00.00%	00.00%	31.58%
JUN	00.00%	31.20%	00.00%	30.22%	00.00%	00.00%

5. Conclusion

This study proposes an integrated system for cancer-related gene research with two-dimensional correlation assay based on bio medical literature mining. This system can find out chromosome regions that cancer possibly produced and cancer-related genes (oncogenes or tumor suppressor genes). And the system will show the other

cancer name that chromosome region appeared LOH or CGH, and calculate a weight to each chromosome region as its importance. Therefore, different gene information can be analyzed with two-dimensional correlation assay. The methods can find out important information in large amount of biomedical literature and conduct analysis to information for biomedical researchers. It is helpful for the future studies or experiments. It can not only accelerate the study and also greatly reduce the time for integrating and analyzing biomedical literature.

In the future work, there will be more and more various types of biomedical literature and the overall information volume will therefore increase continuously. We consider porting this system to distributed environment for accelerating the information processing. In this study, we use PubMed and Entrez Gene for abstract retrieval. In the future, we will add BioMed Central and PubMed Central, the two full-text article databases as the resource of biomedical literature for refining a more precise analysis result.

References

[1] D. P. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "BioRAT: extracting biological information from full-length papers," *Bioinformatics*, Vol. 20, No. 17, pp. 3206-3213, July 2004.

[2] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda, "A text-mining system for knowledge discovery from biomedical documents," *IBM Systems Journal*, Vol. 43, No. 3, pp. 516-533, July 2004.

[3] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and l. Mazo, "Extracting human protein interactions from MEDLINE using a full-sentence parser," *Bioinformatics*, Vol. 20, No. 5, pp. 604-611, 2004.

[4] M. J. Schuemie, M. Weeber, B. J. A. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jellier, B. Mons, and J. A. Kors, "Distribution of information in biomedical abstracts and full-text publications," *Bioinformatics*, Vol. 20, No. 16, pp. 2597-2604, November 2004.

[5] J. T. Chang, H. Schutze, and R. B. Altman, "GAPSCORE: Finding gene and protein names one word at a time," *Bioinformatics*, Vol. 20, No. 2, pp. 216-225, January 2004.

[6] NCBI HomePage, http://www.ncbi.nlm.nih.gov, July 2005.

[7] R. H. Guttman, A. G. Moukas, and P. Maes, "Agent-mediated electronic commerce: A survey," *The Knowledge Engineering Review*, Vol. 13, No. 2, pp. 147-159, July 1998.

[8] S. Franklin and A. Graesser, "Is it an agent, or just a program?: A taxonomy for autonomous agents," in *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, 1996.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2001.

[10] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, Vol. 24, No. 5, pp. 513-523, 1988.

[11] Shih-Nung Chen, Jeffrey J. P. Tsai, and Wei-Hao Chen, "An intelligent agent-based biomedical literature mining system for cancer-related genes," in *Proceedings of the IEEE 5th International Symposium on Multimedia Software Engineering (MSE 2003)*, pp. 279-286, December 2003.