Sequence motif discovery with computational genome-wide analysis

Hirofumi Akashi[1,2], Fumio Aoki[3], Minoru Toyota[2,4], Reo Maruyama[2,4], Yasushi Sasaki[2,4], Hiroaki Mita[2,4], Hajime Tokura[1], Kohzoh Imai[2], Haruyuki Tatsumi[5] and Takashi Tokino[4]

[1]*Scholarly Communication Center, Sapporo Medical University*
[2]*First Department of Internal Medicine, Sapporo Medical University*
[3]*Department of Ecology and Evolution, State University of New York at Stony Brook*
[4]*Department of Molecular Biology, Cancer Research Institute, Sapporo Medical University*
[5]*First Department of Anatomy, Sapporo Medical University*

ABSTRACT

As a result of the human genome project and advancements in DNA sequencing technology, we can utilize a huge amount of nucleotide sequence data and can search DNA sequence motifs in whole human genome. However, searching motifs with the naked eye is an enormous task and searching throughout the whole genome is absolutely impossible. Therefore, we have developed a computational genome-wide analyzing system for detecting DNA sequence motifs with biological significance. We used a multi-parallel network computing system as a powerful computing engine. Furthermore, we improved the system to work as a background engine for web-based applications. The multi-parallel computing engine consists of a head processor, which issues control commands to data processing nodes for various kinds of jobs, such as retrieving arbitrary sequences, generating mapping images, and loading data sections from genome databases. We constructed the system to function as a flexible Client/Server structure connected over the network, and this system could be adapted to cope with increases in sequence data and to deal with algorithms for new investigation needs by slightly changing the control procedures and increasing the number of the processor node. We developed two additional tools to annotate the genome sequences. The first was the cDNA Reverse Splicing Tool, which divided cDNA sequences into exons and mapped them on the genomic sequence, and the second was DNA-Protein Translation Tool which showed open reading frames (ORFs) of whole genome. In order to examine the availability and efficiency of our system, we searched and identified p53RE (p53 response element) as a representative sequence motif on genomic sequences of chromosome 21 and 22. As a result, we detected 50,000 p53REs on fifty mega base genomic DNA sequences within 27 seconds.

Key words : Sequence motif, Bioinformatics, Algorithm, Genome annotation

---

Correspondence to : Fumio Aoki,
Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, NY 11794-5245
Phone : +1-31- 632-8539 ; Fax : +1-631-632-7626 ; E-mail : aoki @ life.bio.sunysb.edu.
Hirofumi Akashi and Fumio Aoki contributed equally to this work.

INTRODUCTION

The Human Genome Project was completed in 2003 [1] and we have been able to utilize a huge amount of DNA sequence data freely, and because of the growth of the Internet and construction of public biological databases such as GenBank, we have been able to get whole genome sequence data, analyze them and obtain significant information. An example of such information is DNA sequence motifs, which include a variety of cis-regulatory modules and are important for predicting the functions of several genes around them. However, searching motifs with the naked eye is extremely tedious work and searching throughout the whole human genome is fundamentally impossible. For the last several years, computational technology has also made great and rapid progress. Therefore, using the "computer" to analyze the genome sequence was a natural solution. However there were no tools to search sequence motifs through the whole genome in one instance. We therefore tried to develop a computational genome-wide analyzing system for detecting DNA sequence motifs. Such analysis required a powerful computing engine, so we therefore took advantage of the parallel processing architecture for VHP (visible human projects) Image Viewer, which had been successfully used in GIBN (Global Interoperability for Broadband Network), a G7 information Society Project, in 2000 [2-6].

First, we chose p53RE as a target sequence motif of our system, because transcriptional co-activator p53 protein binds to this element and regulates many cell-biological functions, such as cell cycle regulation, apoptosis, inhibition of angiogenesis, and DNA repair by inducing transcription of a variety of genes around p53REs[7]. Therefore, complete detection of all p53 regulated genes will lead to an understanding of the mechanism of carcinogenesis, the development of new treatments and the genomic drug discovery.

Here we report three bioinformatics tools developed for the Post-genome sequence era. We hope that our system will be a helpful tool for biological researchers to reveal the functions of genes.

MATERIAL AND METHODS

Software development

We developed three original softwares, Motif Detection Tool (MDT), cDNA Reverse Splicing Tool (cDRST) and Protein Translation Tool (PTT) for this investigation. The programs, including Web pages, CGI modules, and various kinds of processing modules and data communication modules were developed using general UNIX-C language in Linux OS.

System development

For our system, we constructed a parallel processing cluster with five low cost Personal Computers (PC) connected together by a 100 Mega Byte per second fast ether network. These computers had a 533 Mega Hertz CPU, 128 Mega Byte main memory, 20 Giga Byte hard drive, and were installed with RedHat6.2J, a Linux system software. In this investigation, the CPU power was more important than its memory space. Web server software, CGI (Common Gateway Interface) programs and task controller program were installed on the same PC, the other PCs were configured to work as Image Data Processors and Matching processors. Certainly, the Web server and task controller could be assigned to different PCs, and also it was not necessity for the task controller and parallel processing cluster PCs to be allocated to the same network. The system can be used through the Internet, as long as the user's computer is equipped with a widely used Web browser, such as Netscape or Internet Explorer.

Biological data

The human chromosome 21 (28 mega base) and 22 (23 mega base) genomic DNA sequence data released from GenBank were used to search for p53RE for validation of Motif Detection Tool and Protein Translation Tool. Also mRNA data released from GenBank were used

in cDNA mapping tool.

RESULTS

System Architecture

The system was composed in form of a Client/Server structure, including a web server, task controller, and multi-parallel computing engine. The task controller commanded the computing engine to perform various jobs, such as retrieving a particular portion, generating DNA mapping images, and loading data sections from the genomic sequence databases. The Web-based interface communicated with the task controller thorough a set of CGI programs. The task controller exchanged data with the multi-parallel computing engine via a TCP/IP socket connection which was widely used as the basic protocol over the Internet compliant and is a very promising system for the Next Generation Internet, a high performance network.

In dealing with the genome sequence, there were a lot of tasks such as loading chromosome sequences from the database, and generating images to visualize the nucleotide portion and their associated biological information. These functions are assumed under another two programs, cDNA Reverse Splicing Tool (cDRST) which divided cDNA sequences into exons and mapped them on the genomic sequence, and the second was Protein Translation Tool (PTT) which showed open reading frames (ORFs) of the whole human genome.

In our system, the Web server connected to the task controller as a client/server system, and the task controller connected to the data processing nodes in the computing engine in the same way. By using this flexible client/server architecture, we can modify the system architecture to correspond to increase in the database or new investigation algorithms.

Motif Detection Tool (MDT)

Our tool identifies sequence motifs with a two-step search. The first step is retrieving character strings matching with the query motif sequence. In this step, we had to detect the sites matching the motif sequence imperfectly as well as the perfectly matching sites, because each motif had redundancy and tolerated some levels of mismatch. The second step is analyzing positional relationships of motifs, because some motifs needed to be located near each other, such as in tandem arrangement, for their function.

We are going to describe the details of our system, taking p53RE as an example for a sequence motif. p53RE is defined by a 20-nucleotide palindrome sequence consisting of 10-nucleotide elements separated by an arbitrary spacer sequence[8]. This consensus binding sequence is allowed a mismatch of less than four in the 20-nucleotide and the number of nucleotides in the spacer is less than 12. p53RE is represented as shown bellow (m < 12).

[RRRCWWGYYY] + (N)m + [RRRCWWGYYY]

In the above expression, R represents a purine nucleotide, Adenine or Guanine, Y represents a pyrimidine nucleotide, Cytosine or Thymine, W represents a nucleotide that takes two intermolecular hydrogen bonds (binding force is Week), Adenine or Thymine.

In the first step of searching, we retrieved all the portions matched to any 10-nucleotide element of p53RE on the genetic DNA sequence, and in the second step, we collected the satisfying pairs from the records of the primary search by checking if the interval length was within (N)m (Fig. 1). To improve the response of the search process, the chromosome sequences were pre-loaded into the matching processors, and the matching rate checking operations for immediate stop when reaching maximum unmatched nucleic acid were performed during both the primary and secondary search. The operation significantly decreased the ineffective searching loops and then speeded up the retrieval process.

After collecting the satisfying records of the detection site, they were transferred from motif matching processors to both image data processors and expanding data processors via

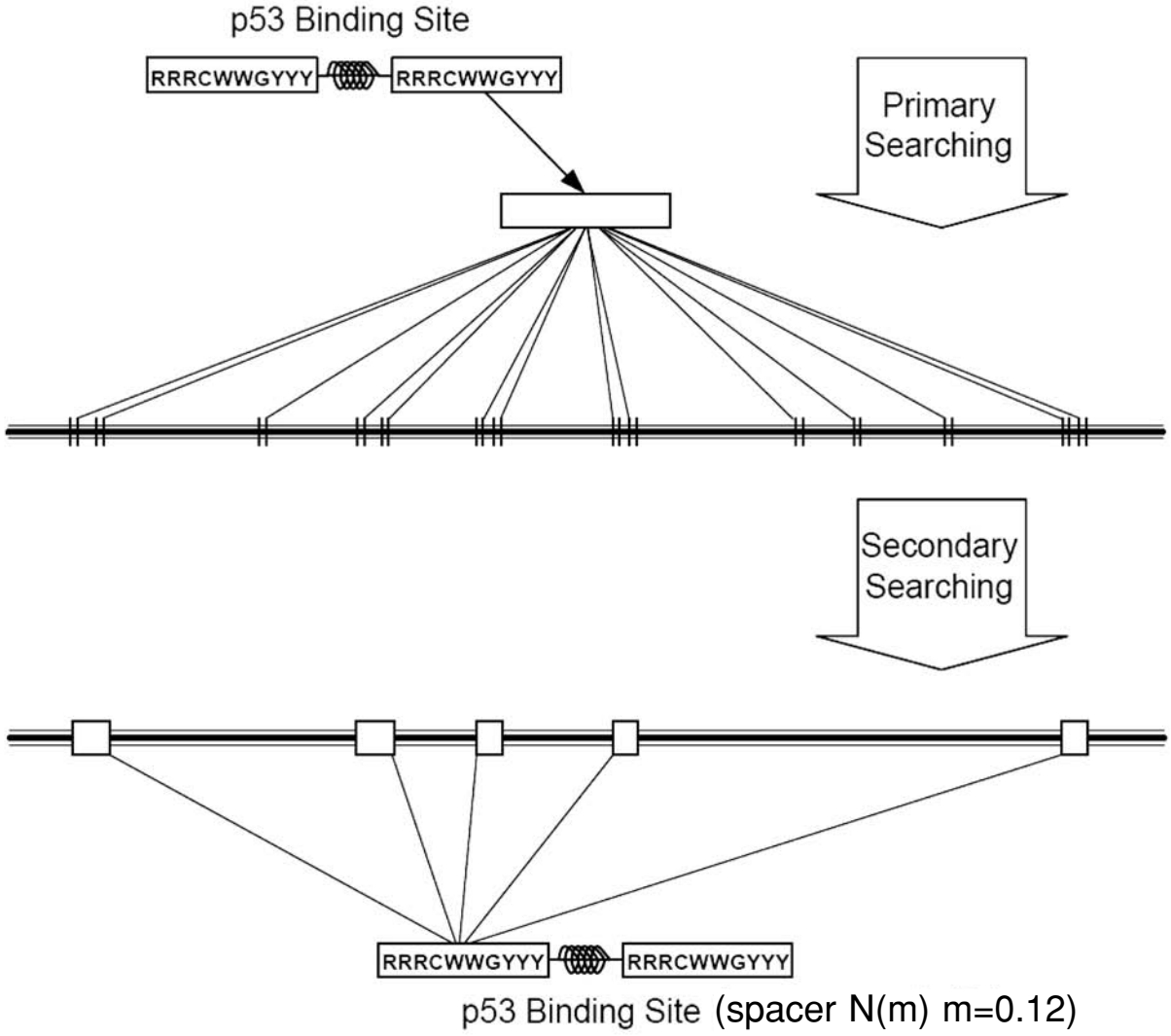

Fig.1 p53RE Detection Algorithm. In the first step, all the portions matched to any 10-nucleotide element of p53RE were retrieved, and in the second step, the pairs which met the requirement of interval length within (N)m were collected.

TCP socket connections, for the generation of mapping images and to expand the output. Here, 31 JPEG images were generated by a group of image processors, one was for the top image indicating the positions of all putative p53 binding sites on the chromosome sequence, and 30 images were for detail visualizations. Expanding portions nearby each p53RE could be retrieved and 200KB sequence output was achieved according to the user's query. The last task was to write the text data and image files onto the disk of the Web server, and to create an HTML file to the standard output (Fig. 2). In web implementation, the user could send the detection definitions for chromosome 21, 22 to the web server, and even arbitrary sequences pasted to the text field. Detection information, including chromosome length, motif definition, process time for each function, and the positions of binding sites with neighboring nucleotides were indicated in text form and motif locations were visualized on chromosome sequence by mapping. These mapping images were clickable images, and when one of the stripes was cricked an expanded mapping image was displayed, indicating the detail motif locations in that region.

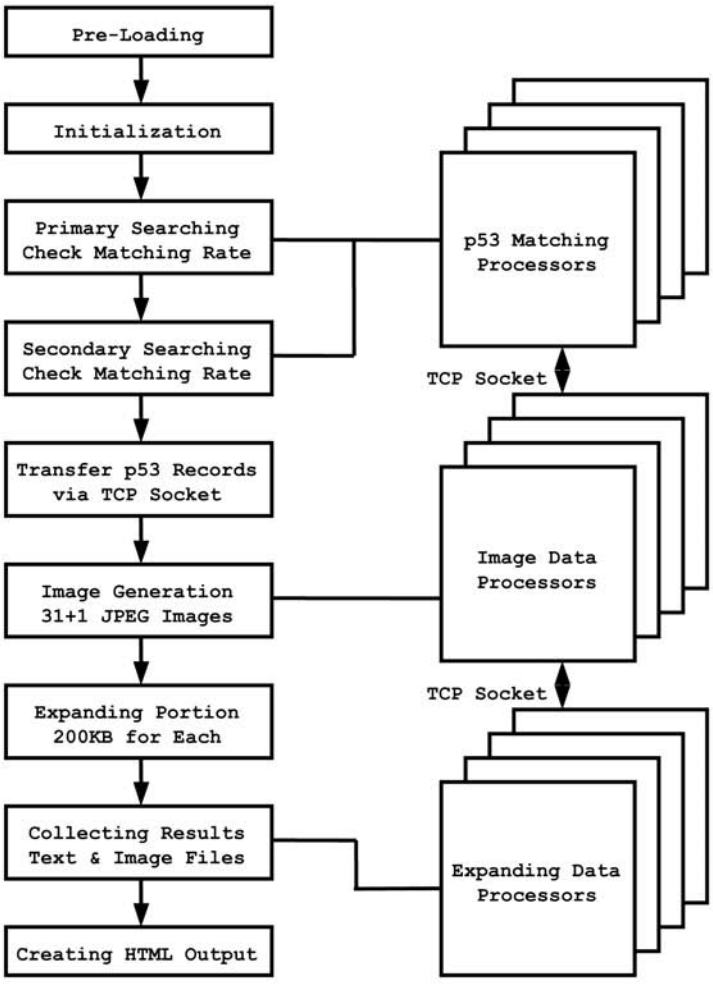

Fig.2 p53RE Detection Processing. The chromosome sequences were pre-loaded into the matching processors, and the primary and the secondary searches were performed until unmatched nucleic acid reached maximum rate. Next, records of the detection sites were transferred from motif matching processors to both image data processors and expanding data processors via TCP socket connections. Thirty-one JPEG images and 200KB sequence data were generated by a group of processors. The final task was to create HTML files to the standard output.

cDNA Reverse Splicing Tool (cDRST)

We also developed a tool to divide cDNA sequences into exons and map them on genomic sequences. This tool indicated that some genes existed around the detected motifs and lead to the discovery of some genes regulated by the detected motifs. There were some tools to predict exons based on genomic sequence, such as GenScan[9], but there was no typical algorithm to predict genomic structure reversely based on cDNA sequence. We therefore developed an adaptive expanding algorithm for re-splicing cDNA into exons reversely and mapping each exon onto the genetic sequence, adaptively. Usually, a cDNA only included the exons without any breaking marks between them. It was very difficult to divide cDNA into exon pieces at correct points, the only available knowledge was that the exon pieces exist on the genetic sequence with no more than a certain distance between them and they have a contextual order within the sequence. Our algorithm divided

cDNA into small portions with equal length to be matched onto the genetic sequence wihile taking account of the contextual relations. When a portion was matched somewhere on the genomic sequence, the algorithm started expanding along both ends as long as the corresponding nucleotide was identical between cDNA and genomic sequence. Then, the start and end pointers of this exon were stored as a matching record. After all the small portions had been examined, our tool eliminated redundant putative "exons" (Fig. 3).

By pre-loading the chromosome sequence into "Exon Matching Processors", the overhead for accessing tens of megabytes of data could be cut off. The cDNA portion was received from the Web browser when the user submitted a query. These were then pre-processed to check the content legality and to cut off the illegal bytes, while reading the parameters for minimum number of exon length and maximum number of unmatched nucleotides. The task controller received the information in the query to divide cDNA into small pieces of equal length, and sent them to Exon Matching Processors via TCP socket connections. Matching of

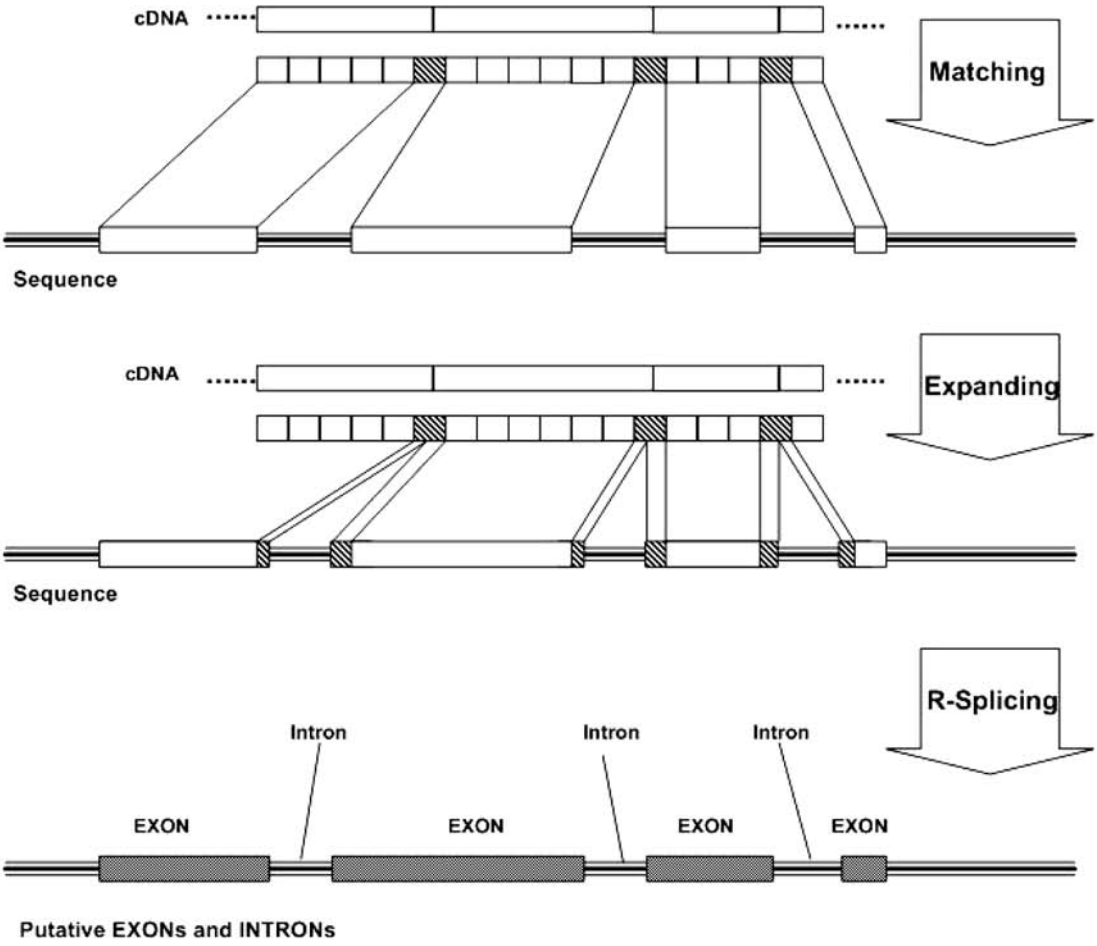

Fig.3 cDNA Reverse Splicing Algorithm. At first, cDNA was divided into small portions of equal length to be matched onto the genetic sequence while taking account of the contextual relations. When a portion was matched somewhere on the genomic sequence, the algorithm started expanding along both ends as long as the corresponding nucleotide was identical between cDNA and genomic sequence. Then, the start and end pointers of this exon were stored as a matching record. Finally, redundant putative exons were eliminated.

the sequences was performed simultaneously to check the matching rate based on maximum unmatched nucleotides, exon context and the distance between first and last exon. After collecting the records from the matching processors, the task controller picked out the overlapping records which had the same start and end pointers and merged them in one exon. Here, only one piece of the image was generated to indicate the cDNA location on the genetic sequence. The last task was to write the text data and image file onto the disk of the Web server, and to send the HTML content to the standard output. In web implementation, the user could send a query cDNA sequence and some parameters, such as minimum length of exon, maximum number of unmatched nucleotides and some output options. The result pages for this application indicated the text form and mapping showing the start and end positions of exons matched on the chromosome.

Protein Translation Tool (PTT)

DNA-Protein translation tool shows open reading frames (ORFs) of mapped cDNA around sequence motifs detected by our system. For

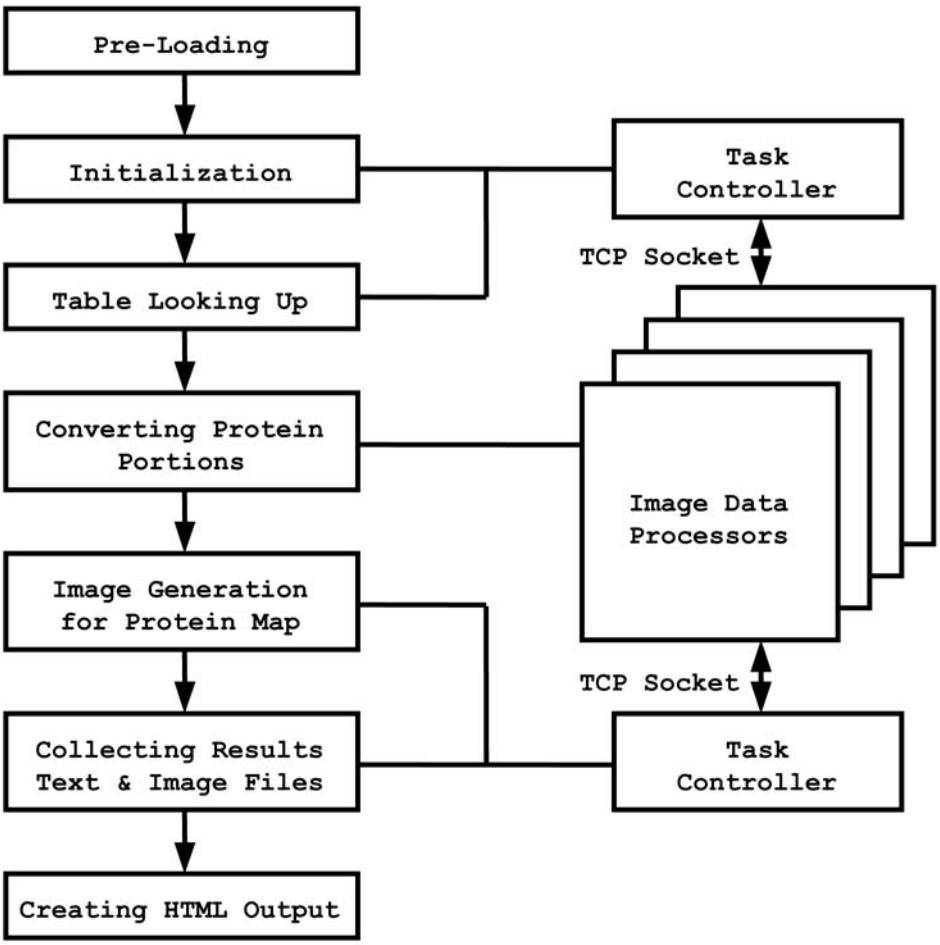

Fig.4 Protein Translation Processing. A table looking up algorithms was used and the given DNA sequences were translated according to the codon table. Image data processor was used only for generating protein mapping image of JPEG format for Web browser. Finally, all the text data and image files were written on the Web server's disk, and HTML content was created and send to the standard output.

the protein translation of a given DNA sequence, a table looking up algorithm was applied. The CGI program received query DNA sequence from Web-based GUI, checked its validity and converted it into internal format to be translated according to the codon table. On the Web browser, the user could obtain both the amino acid sequence and the image marked with red lines as stop codon by starting translations with three frames. The open reading frame (ORF) was automatically determined by detecting the maximum length of non-stop codon region and was shown in the final output JPEG image. The protein translation process was very simple. It simply converted every chromosome code in any given DNA portion into a protein description by looking up the codon table installed in the system. The process has a pre-processing step to check the content legality and cut off the illegal bytes, getting the text output switch from the user's query. We did not use parallel processors for the translation because this task was easily done by computer. An image data processor was used only to generate a protein mapping image of JPEG format for the Web browser. Finally, all the text data and image files were written on the Web server's disk, and HTML files were created and sent to the standard output (Fig. 4). In web implementation, DNA sequences were pasted in the text field and the translation result gave an image with 3 reading frames. The last stripe with the longest non-stop codon region is detected as the proper result. Three sets of protein sequences in text form corresponding to the original DNA portion were also printed.

System performance

The p53RE detection processing time and the matching records were measured for both chromosome 21 and 22 sequence, corresponding to the primary search and secondary search, and the perfectly matching and partially matching sequences from 19/20 to 16/20 with the spacer fixed to m=12. And we also counted the detection results of p53 binding site with changing the spacer changed from m=0 to m=12, and each p53 binding site number for the longer spacer included the binding site number with shorter spacer (Table 1, 2). The primary search took time more than the secondary search, and distributing the searching process to parallel processors was very effective for improving the response speed. cDNA Reverse Splicing Tool could divide all cDNA sequences examined into each exon and map them on the chromosome sequence exactly (Table 3).

DISCUSSION

Herein, we report the development of three

Table 1 Secondary searching results of p53RE detection on chromosome 21 with various length of Spacer

Match Rate	20/20	19/20	18/20	17/20	16/20
Spacer = 0	7	43	190	939	2,796
Spacer = 1	12	80	381	1,675	4,755
Spacer = 2	14	109	562	2,399	6,768
Spacer = 3	16	143	736	3,151	8,957
Spacer = 4	20	170	895	3,833	11,159
Spacer = 5	24	201	1,048	4,491	13,017
Spacer = 6	28	225	1,233	5,133	14,841
Spacer = 7	31	257	1,407	5,864	16,859
Spacer = 8	34	285	1,593	6,678	18,639
Spacer = 9	38	310	1,750	7,258	20,104
Spacer = 10	40	349	1,945	7,895	21,683
Spacer = 11	40	370	2,110	8,583	23,400
Spacer = 12	45	395	2,255	9,094	24,782

Table 2 Secondary searching results of p53RE detection on chromosome 22with various length of Spacer

Match Rate	20/20	19/20	18/20	17/20	16/20
Spacer = 0	10	32	173	1,333	3,604
Spacer = 1	14	54	372	2,068	5,684
Spacer = 2	17	82	558	2,914	8,081
Spacer = 3	20	114	752	3,812	10,676
Spacer = 4	21	137	925	4,526	13,216
Spacer = 5	23	161	1,120	5,207	15,144
Spacer = 6	27	181	1,272	5,774	16,950
Spacer = 7	30	211	1,461	6,497	18,683
Spacer = 8	31	239	1,719	7,602	20,496
Spacer = 9	32	273	1,939	8,294	22,059
Spacer = 10	35	297	2,111	8,924	23,623
Spacer = 11	36	316	2,277	9,537	25,093
Spacer = 12	42	349	2,444	10,109	26,390

Table 3 Results of cDNA Reverse Splicing Tool

No.	cDNA Name	Length	Start on Ch. 22	End on Ch. 22	Length	Exon	Max Exon	Min Exon
1	HCF2	2,182	520,977	529,402	8,426	4	907	147
2	CRKL	1,881	659,108	691,986	32,876	3	822	470
3	LZTR-1	4,227	724,099	740,721	16,623	27	705	54
4	HP2XM	3,552	756,843	770,513	13,671	14	934	71
5	MAPK1	1,476	1,505,733	1,609,317	103,585	9	318	112
6	MAPE-c	2,148	2,277,510	2,289,023	11,514	6	946	39
7	GNAZ	2,676	2,825,262	2,854,601	29,340	4	829	327
8	MMP11	2,260	3,502,366	3,513,832	11,467	8	906	131
9	MIF	523	3,623,957	3,624,738	782	3	177	143
10	DDT	413	3,691,018	3,697,080	6,063	3	177	103
11	ADORA2A	2,363	4,150,048	4,159,256	9,209	2	1,751	615
12	ADRBK2	3,628	5,281,900	5,440,870	158,971	21	1,660	54
13	MN1-c	7,556	7,432,458	7,485,679	53,222	2	4,736	2,822
14	ADTB1-c	3,845	9,011,971	9,107,307	95,337	21	1,040	70
15	NEFH	3,750	9,164,375	9,175,502	11,128	4	2,545	129
16	LIF-c	3,870	9,888,693	9,894,998	6,306	3	3,608	84
17	SMTN	1,576	10,739,585	10,752,860	13,276	11	309	63
18	LIMK2	3,807	10,896,638	10,928,314	31,677	15	1,784	57
19	SLC5A1	2,449	11,691,508	11,758,894	67,387	15	671	62
20	TIMP3	5,468	12,449,056	12,511,275	62,220	5	3,861	84
21	HMOX1	1,550	15,029,334	15,042,441	13,108	5	736	102
22	MB-c	1,067	15,212,715	15,212,881	10,563	3	681	166
23	MYH9-c	5,882	15,878,220	15,944,787	66,568	39	333	64
24	MPST	1,225	16,590,966	16,596,595	5,630	2	619	608
25	LGALS2-c	429	17,136,998	17,146,670	9,673	4	161	33
26	LGALS1	522	17,242,375	17,246,535	4,161	4	196	77
27	GCAT	1,442	17,374,704	17,383,636	8,933	9	333	85
28	KCNJ4-c	1,903	17,993,063	18,010,764	17,702	2	1,846	59
29	GTPBP1	2,123	18,272,794	18,297,655	24,862	12	355	90
30	DNAL4-c	1,480	18,345,244	18,360,885	15,642	4	1,108	82
31	MGAT3	2,327	19,024,056	19,056,170	32,115	2	2,085	243
32	ATF4	2,016	19,087,300	19,089,416	2,117	2	1,108	909
33	ADSL	1,692	19,913,294	19,933,483	20,190	13	316	45
34	P300	9,046	20,658,521	20,746,276	87,756	37	2,764	43
35	ACO2	2,458	21,035,862	21,095,444	59,583	18	261	56
36	G22P1	2,096	21,188,726	21,230,773	42,048	12	416	97
37	BZRP	636	22,706,187	22,717,881	11,695	3	440	60

bioinformatics tools, Motif Detection Tool, cDNA Reverse Splicing Tool and Protein Translation Tool. We used a multi-parallel network computing system constructed by low cost PCs instead of expensive super computers. The prototype system was implemented to work for p53RE Detection with ambiguous searching, cDNA mapping on genomic sequence with adaptive exon matching, and DNA translation to protein sequence. A multi-step comparison algorithm with partial compatibility was proposed for motif detection, an adaptive expanding algorithm was proposed for cDNA Reverse Splicing Tool, and both algorithms proved efficient and effective in our prototype system. A Web-based interface was used for easy access, the user can send queries to the system and get results, including text data and DNA mapping images from the Web browser. This system used the parallel processing cluster as its background engine consisting of a task controller and data processing nodes. This flexible client/server structure allows efficient modification of the system's architecture to deal with expanding databases or new investigation algorithms.

In searching and identifying p53RE as a representative sequence motif, our system detected a larger-than-expected number of potential p53 binding sites. Because, it is hardly possible that all the detected potential binding sites

are functional, we have to narrow down the candidates of p53RE according to their biological meaningfulness and significance. It is expected that the other sequence motifs, as well as p53RE, are also not defined by sequence only. Now, we are developing advanced systems to specify functional motifs by using comparative genomics and merging other annotation information sequentially.

Today, The Human Genome Project has been completed and other species genomes are being rapidly revealed. Moreover, a lot of biological databases including nucleotide or protein sequences, 3D molecular structures, Pathway data and so on, are made available to the public. We are able to utilize a huge amount of biological data but it is impossible to search the databases and to find the required information by the naked eye. On the other hand computer and network technologies are advancing quickly. In this context, the new academic field of "Bioinformatics", which fuses informatics and biology, is now established and making progress, and our project is a part of Bioinformatics. In Bioinformatics research, it is important to reflect on the ideas of biological researchers regarding bioinformatics tools, to validate the results of computational analysis biologically and to relay validated results back to informatics researchers. In this project, we organized well communicated research teams consisted with informatics and biological researchers, and this allowed us to perform successful for bioinformatics research.

It is expected that future tasks will include the release of the Web site on our findings for public service, the implementation of more algorithms, the improvement of the control procedure corresponding to the increasing parallel processing nodes and genome sequence database. A detection system of sequence features based on artificial intelligence is another important future project.

ACKNOWLEDGEMENT

This study was supported partly by Industrial Technology Research Grant Program in 2000 from New Energy and Industrial Technology Development Organization (NEDO) of Japan, and Research for the Future Program of Japan Society for the Promotion of Science under the Project "Integrated Network Architecture for Advanced Multimedia Application Systems" (BJSPS-RFTF97R16301).

REFERENCES

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431: 931-945.
2. Aoki F, Tastumi H, Nogawa H, Akashi H, Nakahashi N, Guo X. A Parallel Approach for VHP Image Viewer. IWS2000 Proceedings on Medical Session-I; 2000: 209-214.
3. Aoki F, Nogawa H, Tatsumi H, Akashi H, Nakahashi N, Guo X, Maeda T. Distributed Computing Approach for High Resolution Medical Images. 16th World Computer Congress 2000 Proceedings on Software Theory and Practice; 2000: 611-618.
4. Nogawa H, Tatsumi H, Nakamura M, Kato S, Takaoki M. An Application of An End-User Computing Environment for the Visible Human Project. The Second Visible Human Project Conference Proceedings; 1988: 99-100.
5. Aoki F, Akashi H, Goudge M, Toyota M, Sasaki Y, Guo X, Li SJ, Tokino T, Tatsumi H. Post-Genome Applications Based on Multi-Parallel Computing over High Performance Network. IWS2001 Proceedings on Bio-Medical 2001: 61-67.
6. Nishinaga N, Tatsumi H, Gill M, Akashi H, Nogawa H, Reategi I.Trans-Pacific Demonstration of Visible Human (TPD-VH). Space Communications 2001; 17:303-311.
7. Maruyama R, Aoki F, Toyota M, Sasaki Y, Akashi H, Mita H, Suzuki H, Akino K, Ohe-Toyota M, Maruyama Y, Tatsumi H, Imai K, Shinomura Y, Tokino T. Comparative genome analysis identifies the vitamin D receptor gene as a direct target of p53-mediated transcriptional activation. Cancer Res

2006; 66: 4574-4583.
8. Tokino T, Thiagalingam S, el-Deiry WS, Waldman T, Kinzler KW, Vogelstein B. p53 tagged sites from human genomic DNA. Hum Mol Genet 1994; 3: 1537-1542
9. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997; 268: 78-94.

(Accepted for publication, Jan. 19, 2007)