The Design and Implementation of Updating and Monitoring Alternative Splicing Database

選擇性剪裁資料庫自動更新與監控程式之設計與實作

Chin-Feng Chen[1,] Wei-Chung Shia[2], Fang-Ming Hsu[3], Wei-Ru Yang[1], and Fang-Rong Hsu2*[2]

[1]Department of Computer Science and Information Engineering
Asia University, Taichung County, Taiwan, Republic of China

[2]Department of Information Engineering and Computer Science
Feng Chia University, Taichung, Taiwan, Republic of China

[3]Department of Information Management
Dong Hwa University, Hualien, Taiwan, Republic of China.

Abstract

Alternative splicing is a very important mechanism of gene expression in eukaryotic cells, to understand the forms of alternative splicing expression will be helpful to the research of medical science. However, since the huge amount of genomic sequence data, it costs huge computation time to construct an alternative splicing database. Since the genome sequencing project is still on going, an alternative splicing database needs to be updated from time to time. Traditionally, researchers just reconstruct the alternative splicing database when it needs to be updated. It will take huge computation time. Hence, we use the intelligence agent technology to help us analyzing the data and decrease the updating time. After the evaluation, our agent can decrease upper to 93% updating time compared to the traditional method of alternative splicing database updating.

Keywords: alternative splicing, database

Introduction

Alternative splicing is a very important mechanism of gene expression in eukaryotic cells. Now we know the reason that many diseases are highly related to the gene structure and its expression. Hence, to

understand the forms of alternative splicing expression will be helpful to the research of medical science. Through collecting and computing the huge amount of Expression Sequence Tag (EST) data, we are able to collect the information about the alternative splicing.

In resent years, following the development of bioinformatics and the draft complete in the Human Genome Project (HGP), we have a lot of type of sequence data for analyzing. Like EST, mRNA, Genome, etc. These biological sequences are the source of many related research. There are many famous of alternative splicing database, like ASAP [1], ASD [2], ASDB [3], ASG [4] and AVATAR[5]. However, following the better sequencing technology and many labs join to this research domain, sequence data also increase more and more. It also need more and more computation time.

In traditional method of alternative splicing detection, it needs to collect all sequences like EST, mRNA and genome to compute the result. When the sequence data changed or adding new data, these procedures will be run again to get the up to date result. For human genome, it needs five personal computers and 6 months of computation time. How to decrease the updating time and keep the accuracy of results is an important problem in bioinformatics.

In this research, we will analyze the sequence data and use the intelligence agent technology to monitor and update an alternative splicing database (AVATAR). This database collects the alternative splicing expression of humans. According to the difference of old and new version of the genomic sequence data, it can re-compute the variant data. Our method can decrease more than 93% of computation time, compared with the traditional methods.

Materials and Methods

Methods of Alternative Splicing Detection

In the transcription step of DNA to RNA, it is necessary to pass many transcript events, and alternative splicing is the very important one event in these. Alternative splicing is also the mechanism that a single gene can produce the different proteins because this mechanism can influence the gene structures, and it caused the same gene may have different expressions between two different individual [6].

If we want to know the forms of gene expression, we need to analyze the product after post transcript event. Hence, we have the expression sequence tag (EST) technology. EST is a type of cDNA sequences, and it is a product through reversing the transcription of the mRNA fragment. Because the mRNA means the coding region on genomic sequences, we can get original structure of genes by assemble EST sequences. When we collect many samples from different individuals, we can compute the different in these gene expressions.

When we acquire EST sequences, we need

to know the original position on genome. This step also called the EST to genome alignment. [7] We can get each exon and intron positions on genome after alignment, and sort out them according to the each exon positions on genome to establish the expression forms of alternative splicing. In Figure 1, it shows the detect method of alternative splicing by detect the exon and intron position on genome sequence.

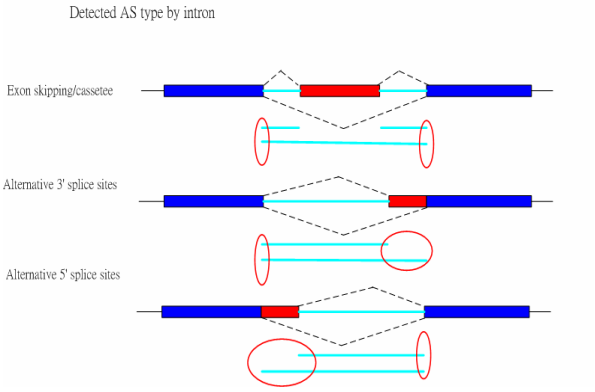

Figure 1: Method of alternative splicing site detection.

There are many expression forms in alternative splicing, and difference forms can generate difference protein. In Figure 2, it lists the most common expression forms of alternative splicing.

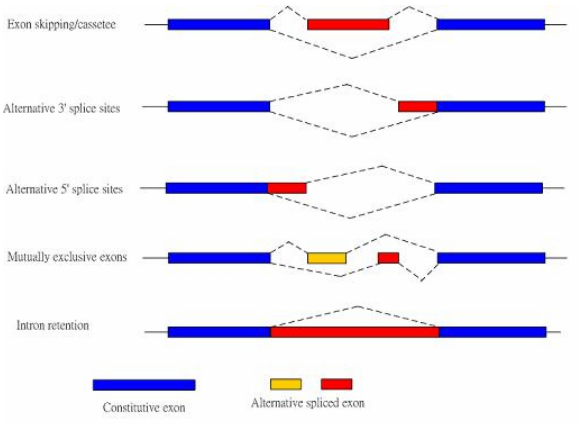

Figure 2: The most common types of alternative splicing.

If we want to know about the more details of alternative splicing, we need to analyze more EST sequence data to get it. However, the EST to genome alignment usually takes huge computation time. If we align all EST sequences form NCBI, it may take a year to compute these data. Just to compute the data of Homo sapiens (Human) can also needs 6 months.

Therefore, we try to decrease the computation time by comparing the differences between the new and old version of genomic sequences, recomputed the variant data to reach the data synchronization and auto update the data of the database.

Our purpose in this research tries to implement an intelligence agent that can monitor the new data published by NCBI and can update database automatically. NCBI often releases the new sequence, and when we get the new data, we need to recompute the alternative splicing data, and insert the new genomic data to the database automatically. This system divided into five components: data collection, EST update, genomic update, data report and mail report. Each component will be triggered by specific events, and call the related component to do the procedure. Figure 3 shows the flowchart of this system.

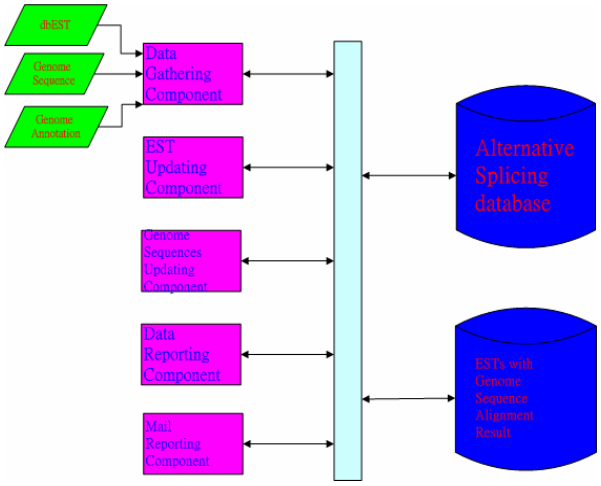

Figure 3: An Architecture diagram showing the components of our system

The Data Collection Component

This component is used to confirm the update of experimental material (including EST, genomic sequence and its annotation), and download these data automatically.

NCBI dbEST will update about thousands of human EST sequences monthly, and update the genomic sequence data for 4 months. This component will monitor the NCBI ftp server to get the new data and decompress these data to the specific folder for future use.

The EST Update Component

When the data collection component reporting to the data on NCBI is changed, EST update component will be base on the result of the differences in the new version of the genomic sequence we made previously to do the EST to genome alignment, and import the new result to the database. Finally, according to the EST sequence that belongs to the forward or reverse strand after align to compute the alternative splicing result. This agent can also read the annotation on genomic data to generate the differences between new and old genomic data automatically, and show the differences in graphs.

Genomic update component

The purpose of this component is to generate the difference list between the new and old genomic sequence after reading the annotation on it. This list is also the important reference data for other component when they update or re-compute sequence data. We will describe the processes of this component in detail.

First, we need to get the EST and Genome alignment result in pervious version. Next, we read the annotation on new genomic sequence, and get each EST's position on it. This is the most important step in all processes. The method to find the new EST's position is mainly relay on to compare the difference of gene position on new and old genomic. Finally, we according to the exon new position on EST and it belongs which one BAC to sort out it.

However, some EST may not belong to any known BAC. Because the quantity of BAC in new or old genomic maybe have some change, some BAC in old version maybe deleted in new version, and it caused some EST will be aligned again to get its position on new genomic sequence. Therefore, some EST may be included two or more BAC at the same time. To solve this problem, we

will check which each BAC if be changed or deleted, and according to this information to adjust the new position of ESTs. Figure 4 shows the position difference after the transfer of this component.

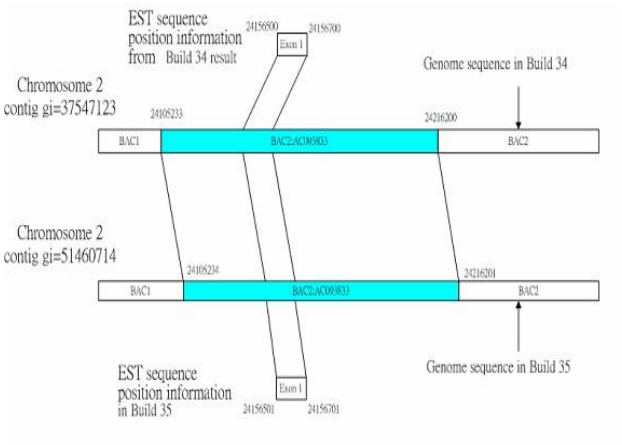

Figure 4: The transfer of EST position

The Report component

Following to the update of database, the last quantity of sequence data we have must be generate again. The source data of alternative splicing comes from many type of sequence, like mRNA and EST. To each gene or chromosome, how much of the EST or mRNA sequence be aligned to the same region concern with its accuracy. Difference data source need difference statistics, and the item of statistics is very much. Hence, we need to build up a report component to generate the newest statistics of data in database.

It has two ports of these statistics. First is the update of genome. We try to provide a visual interface to show the gene structure and the annotation on genomic and show the information of each BAC, gene, or contig.

When we added new EST data to the database, this component will generate the newest statistics of EST quantity according to the gene boundary. We also provide user can simply input the chromosome number, contig ID or gene name to query the quantity of EST that align in the same region.

Hardware and Software

We use an person computer have P4 2.6 GHz, 512MB RAM and 200GB hard disk to build up this database and compute sequence data. This database and other related programs are run on Windows 2000 SP4. We use Perl to develop our CGI program, and use IBM DB2 7.2.3 to our database management system.

Experimental Material

The source sequence data we use in this research released from NCBI human genomic sequence data build 34 (B34) and human expressed sequence tag sequence from NCBI dbEST. We implement an agent for AVATAR[5]. We update all sequence data to the build 35 (B35) data to evaluate the efficient of our method. In our database, EST quantity are 6020830, mRNA are 24939, number of gene are 26438.

Result

The Differences of the New and Old Version of the EST Sequence Data

In Figure 5, it shows the file architecture of the genome sequences.

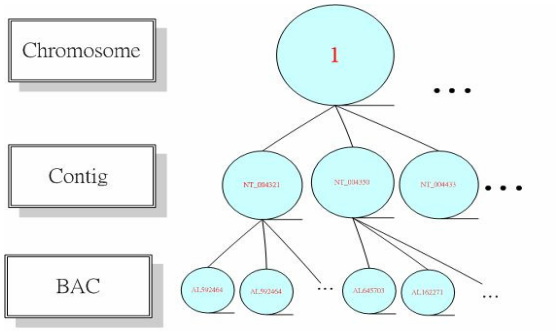

Figure 5: File architecture of the genome sequences.

Contig can be seems to the unit of chromosome when we assemble the genomic. The EST sequence length is usually about 600 nucleotides and the length of a chromosome usually between millions to the hundred millions nucleotides. Hence, we need assemble the EST sequence to the small part, and take these parts to the full chromosome. These small parts of sequences called the contigs.

BAC (Bacterial artificial chromosomes) is the artificial chromosomes that get form bacterial and used to store the cDNA sequence. Contig can see to a set of BACs, and through analyzing the BAC annotation, we can understand the change between new and old chromosome sequence.

We will discuss the differences between NCBI Human Genomic Build 34 (B34) and Human Genomic Build 35 (B35) version of human genome sequences from four points: quantity of contigs, BACs, gaps and genes.

First, we list the quantity of contigs on the human chromosome. The numbers of contigs are 481 in B34 sequence data, in B35 are 377. There are 459 contigs belongs to the B34 version are be deleted in B35 version, 355 contigs are added to the B35 version, 22 contigs are the same between the B34 and B35 data.

Second, we will compare the differences and its quantity between the total lengths of BAC. There are 27014 BACs in B34, 26906 in B35. There are 206 BACs removed from the B34 data, 305 BACs added to the B35. 240 BACs are the difference between the B34 and B35 data, and 26333 BACs are the same between the B34 and B35 data. Table 1 shows the quantity different of BAC between B34 and B35 version of genomic sequence.

Chromosome	no. of BAC (B2, B3)	The same with Build 34(B2)	Difference with Build 34(B2)	Only in Build 35 (B3)
1	2,240	2,159(96.38%)	44(1.96%)	37(1.65%)
2	1,991	1,975(99.20%)	14(0.70%)	2(0.10%)
3	1,714	1,695(98.89%)	17(0.99%)	2(0.12%)
4	1,638	1,625(99.21%)	5(0.31%)	8(0.49%)
5	1,764	1,740(98.64%)	1(0.06%)	23(1.30%)

6	1,780	1,730(97.19%)	2(0.11%)	48(2.70%)
7	1,536	1,481(98.89%)	14(0.91%)	3(0.20%)
8	1,202	1,194(99.33%)	4(0.33%)	4(0.33%)
9	1,009	975(96.63%)	11(1.09%)	23(2.28%)
10	1,139	1,128(99.03%)	5(0.44%)	6(0.53%)
11	1,121	1,098(97.95%)	4(0.36%)	19(1.69%)
12	1,163	1,154(99.23%)	1(0.09%)	8(0.69%)
13	866	865(99.88%)	1(0.12%)	0(0.00%)
14	663	653(98.49%)	0(0.00%)	10(1.51%)
15	678	672(99.12%)	4(0.59%)	2(0.29%)
16	722	633(87.67%)	62(8.59%)	27(3.74%)
17	678	668(98.53%)	8(1.18%)	2(0.29%)
18	597	593(99.33%)	1(0.17%)	3(0.50%)
19	862	861(99.88%)	0(0.00%)	1(0.12%)
20	638	631(98.90%)	1(0.16%)	6(0.94%)
21	494	481(97.37%)	4(0.81%)	9(1.82%)
22	561	535(95.37%)	0(0.00%)	26(4.63%)
X	1,620	1,570(96.91%)	18(1.11%)	32(1.98%)
Y	228	217(95.18%)	6(2.63%)	5(2.19%)
sum	26,904	26,333(97.88%)	227(0.84%)	306(1.14%)

Table 1: The number of BACs of Homo sapiens chromosomes|

The total length of BACs in B34 is $3*10^9$, in B35 is $2.86*10^9$. There are 140 millions of nucleotides removed from B34 genomic sequence. These are $2.81*10^9$ nucleotides that are the same as the two versions, and 13448765 nucleotides changed in chromosome 7. It shows that biologists are still focusing on this chromosome.

Next, we list the quantity of genes between

two versions of genomic data. In B34, there are 26069 genes, and in B35 are 26850. In these genes, 34% of the gene name changed, 23 genes added to the B35 genomic. It shows that biologists have already found more complex genes from these published bio sequences.

Last, we list the total number of gaps on genomics. The gap is a small vacant region on contig. These are some sequences that we unable clone, and it caused the gaps when we take BACs to assemble to contig. A perfect chromosome sequence should not contain any gaps, and the quantities of gaps on a chromosome is shows the quality of genome sequencing, like chromosome 14. Table 4-6 and 4-7 shows the statistics number of the total gap by reading the genome annotation. In B35 genomic data, we can find the total lengths of gaps decrease clearly, but the length of gap is highly increasing on chromosome 8. It is possible that biologists already cloned some region of sequence that could not be cloned previously, and the positions of each contig on chromosome are adjusted. It also shows that the quality of genome sequence increases gradually. Table 4-8 and 4-9 list the total length of gap on each contig between B34 and B35 version of genomic data.

From our statistics the quantity of BACs, there are 96.33% of BACs not changed between B34 and B35 data. It also means the result of EST alignment and alternative splicing does not change very much. If we just re-compute the variant data, it will save us a lot of time. Hence, we will focus on the rest 4.63% of the variant EST sequence, and align these sequences to the new genome to get the updated result.

The time we saved and the evaluation of the result

According to our statistics on the time cost to construct the alternative splicing database, using 600 millions of EST sequences and NCBI human B35 genomic sequences as the experiment material, the computing with 10 personal computers take 30 days and 6 hours.

Using the method we have proposed in this research, first, we need to translate these static results between new and old version of genomic to the new version data. It took 3 hours to translate these 4635828 EST sequence. Next, we need recomputed other variant EST sequences, and it took 24 hours to align these 118605 EST sequences. Hence, compare this method with the traditional one, and the process time decrease upper to 93%.

The functions of web interface

How to give users an interface can efficiently and easily to query data is also the purpose of this research. Next, we will descript the function of the database web we support.

The Genome and EST Report

For the huge genome data, how to get their differences between the new and old version

genome data in fast way? Hence, we provide the genome report function. Users just input the chromosome number through the web interface, and the genome report will show the position discrepancy of each contig both on new and old version of chromosome. At the same time, we can also query the total number of EST on each contig by this interface.

The Mail Report and the Statistics Report

This database includes the mechanism of registration. Users can give their names and e-mail addresses, when the data in this database are changed, the mail report agent will mail them the last report of data statistics to the users. It will be conveniently for the researchers to do related research. The statistics report agent can also update the last statistics result of database automatically.

Discussion and conclusion

The other alternative splicing database we known have not include the automatic update function. The result of alternative splicing need to compute many types of sequence data to get it, and it also need work force to monitor the data update, flow processes are complex and hard to maintain. After we use the intelligence agent technology, these steps can run automatically to save many work force and time for us, and increase the data accuracy.

Following the increase of the sequence, to build up and update database will becomes more difficult. But in this research according to analyze the position difference of each BAC, gene and contig between the new and old version of genomic to degrade the computation time, and it decrease upper to 93% time we need. The processes of build up a database are still complex, but we try to make it easy.

In this paper, we designed an agent for AVATAR to keep its information up to date. It is interesting to propose a generalized tool for all databases generated from genomic sequences. Database developer could just specify data definition, like the source of data, type of sequences, etc. Then, the tool could generate an agent for updating the database automatically.

Acknowledgements

This work was supported by the National Science Council, Republic of China, under Grant.
NSC 92-3112-B-468-001.

The author is grateful to professor Hsu (Feng Chia University), Yang(Taichung Healthcare and Management University), Chang (National Tsing Hua University) and Hsiao (Taichung Healthcare and Management University) for their comments and careful reading of the draft of this paper.

Reference

C. Lee, L. Atanelov, B. Modrek, and Y. Xing, ASAP: the Alternative Splicing Annotation

Project, Nucleic Acids Res., Jan 2003, vol. 31, pp. 101 - 105.

T.A. Thanaraj, S. Stamm, F. Clark, J.-J. Riethoven, V. Le Texier, and J. Muilu, ASD: the Alternative Splicing Database, Nucleic Acids Res., January 1,2004, vol. 32(90001), pp. D64 - D69.

I. Dralyuk, M. Brudno, M. S. Gelfand, M. Zorn, and I. Dubchak, ASDB: database of alternatively spliced genes, Nucleic Acids Res., Jan. 2000, vol. 28, pp. 296 - 297.

J. Leipzig, P. Pevzner, and S. Heber, The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome, Nucleic Acids Res., Aug. 2004, vol. 32(13), pp. 3977 – 3983.

F.R. Hsu, H.Y. Chang, Y.L. Lin, Y.T. Tsai, H.L. Peng, Y.T. Chen, C.Y. Cheng, M.Y. Shih, C.H. Liu, and C.F. Chen, AVATAR: A database for genome-wide alternative splicing event detection using large scale ESTs and mRNAs, Bioinformation, Apr. 2005, Vol.1 No.1, pp. 16-18.

R.E Breitbart, A. Andreadis and B. Nadal-Ginard, Alternative splicing: a ubiquitous mechanism for generation of multiple protein isoforms from single genes, Annu. Rev. Biochem. 1987, vol. 56, pp. 467-495.

L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller, A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence, Genome Res., Jul. 1998, pp. 967-974.