CONSENSUS SELECTION ALGORITHM FOR AUTOMATIC PRIMER DESIGN SYSTEM

Siriporn Attanoraks [*a], Arinthip Thamchaipenet [b]
Punpiti Piamsa-Nga [c] and Nuanwan Soonthornphisaj [a]

[a] Department of Computer Science, [b] Department of Genetics,
Faculty of Science
[c] Department of Computer Engineering, Faculty of Engineering
Kasetsart University, Bangkok 10900
Thailand.

ABSTRACT

The consensus selection is an important process in the primer design system analysing from sets of protein or nucleotide sequences. An inappropriate consensus can cause poor primers, which can not properly bind to the DNA template. It is time-consuming for experts to select the most appropriate consensus from input alignments. Currently, there are many automatic tools to deal with the primer design problem. But all of them are not compatible and the tool for the complete process has not yet been available.

To improve the primer design system, this research proposes three consensus selection algorithms. The first algorithm creates the consensus using DNA backtracking from conserved region. This method gives the most appropriate consensus with the least or non-degeneracy, but can only be used with the protein sequences that had been recorded in GenBank database. The second algorithm groups the consensus by conserved regions to calculating consensus score from the number of degeneracy and consensus length. This algorithm gives high degeneracy consensus with the longest length. The third algorithm groups the consensus islands from conserved regions by the connection of low degeneracy characters and calculating consensus score. This algorithm gives high specific consensus and can be used with unknown protein sequences.

We integrate these algorithms with the alignments and primer design process in order to create the automatic primer design system that can deal with large protein sequences for designing both PCR primers and probes. Our system includes decision support and primer analysis tools. The pre-version of this automatic primer design system (KUPDS) is available at: http://www.kupds.com

KEYWORDS: Primer Design, Consensus, Alignment, PCR, Bioinformatics.

1. INTRODUCTION

Primer design is an important process for gene and genome researches. Biologists would design primers manually or employ some tools in bioinformatics to help on the creation. However, current primer design systems still have limitations, which could be pointed out as follows:

Each process of the existing systems is separated from each other that cause discontinuous workflow. Each tool for each process is incompatible with the others. The experts, thus, have to learn how to use each tool separately by each condition. The results from difference tools that utilized the same process could be different. Therefore, the experts must have skill on working at an appropriate tool in order to be compatible with that particular tool.

For the multiple sequence alignment process, the experts have to select most appropriate consensus from the results and input them through the translation process in order to translate the protein code into genetic code and DNA. Currently, the experts have to go through the whole steps by themselves which is a time-consuming process, especially for selection of the consensus from the large set of protein sequences. If the experts select an inappropriate consensus, it will cause

* Corresponding author. Tel.: +66-0-29428026
E-mail: fscinws@ku.ac.th

deficient primers, which can not properly bind to the DNA template. Furthermore, the number of the consensus's degeneracy directly effects to the number of the potential primers.

In reality, the experts probably operate each process discontinuously and may take extended time to finish some process before starting the next process. Thus, this may cause non-persistency results.

There are several tools had been developed for the primer design system. They could be categorized into 3 groups, which are the tools for alignment, backtranslation and primer design.

1.1 The multiple sequence alignment tools

The multiple sequence alignment method appeared in 1973 by Sankoff *et al.* [1]. They have been continuously developing into several approaches, such as ClustalW [2], ClustalX [3], PROBCONS [4], Block Maker [5], and MACAW [6]. Some of the multiple sequence alignment programs can compare the consensus results, but none of them can automatically select the most appropriate consensus.

1.2 The backtranslation process

This process is to translate the consensus from protein code to RNA and DNA codes, respectively. There are some tools available at http://arbl.cvmbs.colostate.edu/molkit/
and http://www.bioinformatics.Vg/sms/ rev_trans.html

1.3 The primer design tools

For the primer design process, there are several tools providing automatic primer design system. Examples of these tools are CODEHOP [7], Gene Fisher [8], DePiCt [9] and Primer Premier [10]. These algorithms do not accept the multiple alignments input. Most of primer design tools do not have the consensus selection process. They begin the process with the single target nucleotide sequence such as Primer3 [11], Web primer [12], MethPrimer [13] and PDA [14]. But some of them begin the process with the set of nucleotide sequences such as HYDEN [15].

To improve the primer design system, we propose automatic consensus selection algorithms that integrate the multiple sequence alignments and primer design process. And then create the automatic primer design system that can deal with large protein sequences for both PCR primers and probes. Furthermore, we propose the appropriate scoring function that satisfied all conditions of the high specific consensus selection algorithm. Our system includes decision support and primer analysis tools. The pre-version of this automatic primer design system is available at: http://www.kupds.com

2. METHODOLOGY

In general, the experts select the most appropriate consensus by comparing all consensus results and select the longest length from the most conserved region with the least and high degeneracy. To develop the algorithm for the consensus selection, we propose three consensus selection algorithms which are DNA backtracking, Conserve Block and Island algorithm.

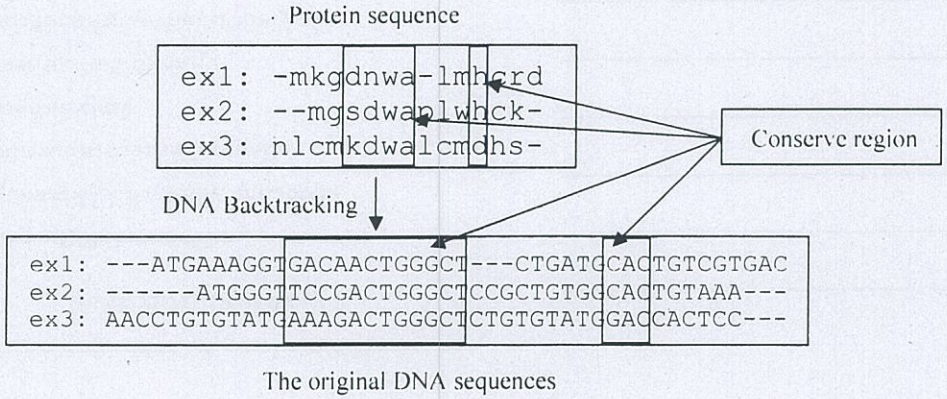


Figure 1. DNA backtracking algorithm.

The first algorithm called the DNA backtracking (Figure 1) that aims to select the most appropriate consensus with the least or non-degenerate by tracing back to see the original DNA sequence of each protein sequence from the GenBank database (http://www.ncbi.nlm.nih.gov). Then the algorithm builds the consensus by selecting the most matched characters from the original DNA sequences and the aligned protein sequences. In case that the character of the original DNA sequence cannot be matched with each other. Codon Usage database (http://www.kazusa.or.jp/codon/) will be applied. However DNA backtracking can only be used when the DNA sequences of the protein sequence had been recorded in the GenBank database.

In case that the DNA sequence of the protein can not be found in the database, we propose to use either the conserve block or the Island algorithm. These two algorithms can produce consensus with different number of degeneracy.

The Conserve Block algorithm groups the consensus by conserved region. Each consensus group is separated by gap "-". Whereas, the Island algorithm groups the consensus from the high conserved region by connecting the low degeneracy protein. After the consensus grouping process has finished, we calculate the length of each consensus group and the pre-ranking score follow by the final score S(i) (see Figure 2).

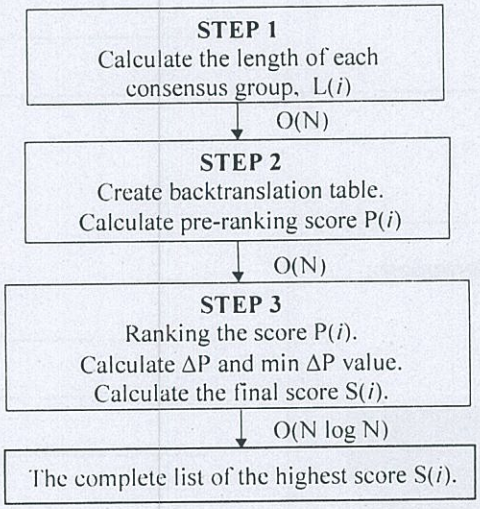


Figure 2.The consensus selection algorithm step by scoring and ranking mechanism.

Step 1: Grouping the consensus result obtained from the alignment process by the conserved region.

Given:

- N is the number of consensus groups.
- i is the index of consensus group.
- L(i) is the length of consensus sequence in the i^{th} group; determined by the number of characters of the first consensus sequence from each group.
- Lmax is the maximum length of consensus sequence over all groups.

Step 2: Calculate pre-ranking score P(i) of each consensus sequence.

The algorithm creates the table called "Backtranslation table" to define backtranslation degeneracy value and the potential DNA codes (see Table 1). The backtranslation data is the Codon Usage table which obtain from the database. The Codon codes are filtered based on the percentage of the matching (Equation 1). The codon that passed the criteria is called backtranslation. The expert defines the threshold value. The system calculates the percentage value from the Freq value in the Codon Usage table, by considering the maximum Freq (Max) = 100% and the value 0 = 0%.

$$\text{percent} = \frac{\text{Freq}}{\text{Max}} \times 100\,\% \qquad \text{----- (1)}$$

After the filtering process, the backtranslation table contains only DNA codes that has the percentage value greater than or equal to the threshold value.

For example, if the expert defines the threshold value to be 50%, and the potential DNA code from the Codon Usage table of the protein code is "P". Given the maximum Freq value = 15.9 to be 100% and given the value 0 to be 0%. Then we calculate the percentage value for the rest of Freq that will give the result as follow:

fields: [triplet] [frequency: per thousand] ([number]) = [percent from frequency]

CCT	15.9 (34) = 100%	CCA	14.0 (30) = 88.05%
CCC	7.5 (16) = 47.17%	CCG	2.1 (2) = 13.2%

The backtranslation table is as follow:

Table 1. The backtranslation table.

Protein code	Number of degenerate	The potential DNA code
P	1	CCT, CCA

Next, the algorithm fills the backtranslation table by calculating the entire of 20-protein codes.

Given:

MinPrimerLength is the user-defined minimum primer length value. (Default value is 18)
deg(i) is the backtranslation degeneracy of each consensus group.
D(i) is deg(i)+1.
P(i) is the score for pre-ranking.
P is P(i) of any consensus.

Select only the protein consensus group that has L(i) ≥ MinPrimerLength / 3 to calculate the value D(i) of each consensus group. And calculate the value P(i) that is the pre-ranking score of each consensus group by Eqs. (2)

$$P(i) = \frac{L(i)}{1 + \log_6 D(i)} \qquad \text{----- (2)}$$

Step 3: Calculate the final score S(i) of each consensus sequence.

Given:

min ΔP is non-zero smallest difference between P of any two consensus, which is positive.
S(i) is the final score for ranking.

This step will rank the score P(i) and calculate ΔP value, the difference between score P of any two following ranked consensus. Calculate min ΔP using the ΔP values of each consensus group.

Next, the algorithm calculates the score S(i), that is the final score from the score P(i) of each consensus group, appending with the consensus group length. In order to select the consensus group that has least degeneracy with the longest length, we use Eqs. (3)

$$S(i) = \frac{P(i)}{\min \Delta P} + \frac{L(i)}{L_{\max}} \qquad \text{----- (3)}$$

Table 2.The example of the score calculation and complete list of the highest score.

Rank	L(i)	deg(i)	D(i)	P(i)	delta P	S(i)
1	30	2	3	18.60		207.29
2	40	15	16	15.70	2.89	175.38
3	41	22	23	14.91	0.79	166.60
4	40	20	21	14.82	0.09	165.57
5	30	11	12	12.57	2.25	140.33
Lmax = 41				min ΔP = 0.09		

After finishing step 3, the complete list of the highest score S(i) is obtained. At this point, the expert can manually select the preferred consensus or let the system run automatically to give the most appropriate consensus output. Next, the system translates the protein code to the DNA code using Backtranslation table and Codon Usage table.

Building the consensus scoring

The purpose of the consensus scoring function is to rank the consensus according to the degeneracy and length. To define P(i) for pre-ranking, we require the consensus that have the least degeneracy and the longest length. Therefore, the greater value of P(i) means that the consensus has less degeneracy and longer length, or simply satisfied the condition "If D(i) < D(j) and L(i) > L(j), then P(i) > P(j)".

Let deg(i,j) represents degeneracy of j^{th} character in i^{th} consensus and the character degeneracy can be the integer from 0 to 5. We obtain the following equation:

$$D(i) = (\deg(i,1)+1)\times(\deg(i,2)+1)\times(\deg(i,3)+1)\times\ldots\times(\deg(i,L(i))+1) \quad \text{-----(4)}$$

According to the equation, we have $1 \le D(i) \le 6^{L(i)}$ and $0 \le \log_6 D(i) \le L(i)$.

From the analysis and experiment to find an appropriate P(i), we set

$$P(i) = \frac{L(i)}{1+\log_6 D(i)}$$

To build S(i), a consensus scoring function, we need to consider the following conditions.

1. For any i, j: If P(i) > P(j) then S(i) > S(j).
2. For any i, j: If P(i) = P(j) and L(i) > L(j), then S(i) > S(j).

An appropriate equation that satisfied all above conditions is as follow.

$$S(i) = \frac{P(i)}{\min \Delta P} + \frac{L(i)}{L_{\max}}$$

The proof of the 1st condition : For any i, j; If P(i) > P(j), then S(i) > S(j).

Let $P(i) > P(j)$, we have $P(i) - P(j) > 0$ and $P(i) - P(j) \ge \min \Delta P$

Consequently, $\frac{P(i)-P(j)}{\min \Delta P} \ge 1$ and we have $\frac{L(i)-L(j)}{L\max} \ge \frac{1-L\max}{L\max} > \frac{-L\max}{L\max} = -1$.

Therefore, $S(i) - S(j) = \frac{P(i)-P(j)}{\min \Delta P} + \frac{L(i)-L(j)}{L\max} > 0$. Finally, we get $S(i) > S(j)$.

The proof of the 2nd condition : For any i, j; If P(i) = P(j) and L(i) > L(j), then S(i) > S(j).

Let $P(i) = P(j)$ and $L(i) > L(j)$, we will have $\frac{P(i)-P(j)}{\min \Delta P} = 0$ and $\frac{L(i)-L(j)}{L\max} > 0$.

Therefore, $S(i) - S(j) = \frac{P(i)-P(j)}{\min \Delta P} + \frac{L(i)-L(j)}{L\max} > 0$. Finally, we get $S(i) > S(j)$.

3. RESULTS AND DISCUSSION

We compare our algorithms with the existing tools in the first experiment using the test set of 128 protein sequences of the Influenza A virus H5N1 from GenBank Database separated into 12 groups. Each group consists of 7-14 sequences that have various conserve and various lengths. Table 3 shows the least degeneracy primer obtained from KUPDS (our proposed algorithm) and CODEHOP. The 1st island is the best island that has the highest score and has the least backtranslation degeneracy. The 2nd and 3rd islands have the lower score respectively. We convert the Island to be the IUPAC format in order to show the location of the primer with the direction 5'-3'.

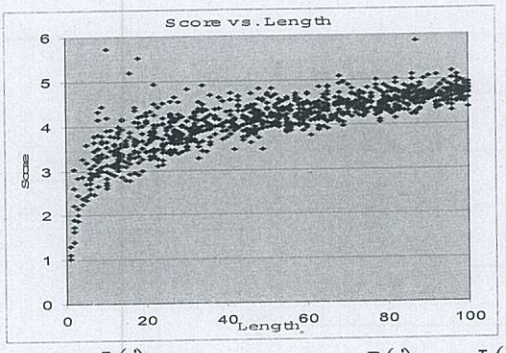
Table 3. The experimental results. The least degeneracy primer of the testing

Group	1: 1st Island with 5'-Primer-3'	2: 1st Island with 5'-Primer-3'
KUPDS	GRATGAARTGGATGATGG	GRATGATGATGGGRATG
Island	AGRATGAARTGGATGATGGCHATGAA	GGNATGATGATGGGNATGTTYA AYA
CODEHOP	CGGATGAAGTGGATGATGGC	CCGGCATGATGATGGG
Group	3: 1st Island with 5'-Primer-3'	4: 1st Island with 5'-Primer-3'
KUPDS	GGGRATGGARATGAGRAG	TGGARTTYTTYTGGACWA
Island	AARATYAARATGAARTGGGG NATGGARATGMGDMGDTGT	GGNMGDATGGARTTYTTYTGGA CVA
CODEHOP	GCACCTCCAAGATCAAGATGAARTGGGG	CCGGCCGGATGGAGTTYTTYTG GAC
Group	5: 1st Island with 5'-Primer-3'	6: 1st Island with 5'-Primer-3'
KUPDS	RAGRGCHATGATGGAYCA	GGAYCCHAAYGGRTGGAC
Island	CARMGDGCKATGATGGAYCAR	GGATCCHAAYGGNTGGACVGAR ACVGA
		6: 2nd Island with 5'-Primer-3'
		CAYAARATHTTYAARATGGA RAARG
CODEHOP	GCGGGCCATGATGGAYCA	CCCACAAGATCTTCAAGATGGA RAARG
Group	7: 1st Island with 5'-Primer-3'	8: 1st Island with 5'-Primer-3'
KUPDS	GATGTGGGARATHAAYGG	GGRATGATGATGGGRATG
Island	ATGATGTGGGARATYAAYGGNCCHGAR	CCHGGNATGATGATGGGNATG TTYAA
CODEHOP	CCTCCTCCATGATGTGGGA	CCCGGCATGATGATGGG
Group	9: 1st Island with 5'-Primer-3'	10: 1st Island with 5'-Primer-3'
KUPDS	TYTTYGGRTGGAARGARC	TGYGAYGAYCARTGYATG
Island	TAYGATGCKATYAARTGTATGAARACVT TYTTYGGNTGGAARGARCCHAARATYAT YAARCCHCATGARAARGGNATY	TAYCAYAARTGYGAYGAYCAR TGYATG
		10: 3rd Island with 5'-Primer-3'
		GAYAARATGAAYAARCARTAY GAR
CODEHOP	ATCAAGTGYATGAARAC	ACAACATGGTGGACAAGATGAAYAARCA
Group	11: 1st Island with 5'-Primer-3'	12: 1st Island with 5'-Primer-3'
KUPDS	TGATGGARAAYGCHAARC	GGRAAYGAYATHTGGATG
Island	YATGAGRACWGARATHATHAGRATGATG GARAAYGCHAARCCHGARGAY	AARGGNTGGGCKTTYGATAAYGGNAAYGAT ATYTGGATGGGNMGDACVATYAAR
CODEHOP	CATGCGGACCGAGATCATCMGNATGATG GA	CGACAACGGCAACGACATHTGGATGGG

The experimental results show that the score of the consensus island obtained from our algorithm can be used to predict the location of the least degeneracy primers.

We found that most of the best islands are located in the area of the conserved block that has the highest score and they are coincident with the area of the high specific primers. Our proposed algorithm can predict the location of high specific primers, which makes the calculation much faster. The expert can easily select the consensus based on the ranking score in order to design the high specific primer.

We analyse our consensus scoring function in the second experiment using the test set which are randomly selected from the database (Figure 3). We obtain 1,000 consensus sequences with the random number of degeneracy and consensus length.



$$P(i) = \frac{L(i)}{1 + \log_6 D(i)}, \qquad S(i) = \frac{P(i)}{\min \Delta P} + \frac{L(i)}{L_{max}}$$

Figure 3 The performance analysis of our scoring function.

The experimental results show that our scoring function yields better score distribution. The score obtained from our function loosely depends on the length of consensus, which means that the consensus that have the same length can have the different score. Hence our function is more suitable for the consensus selection criteria.

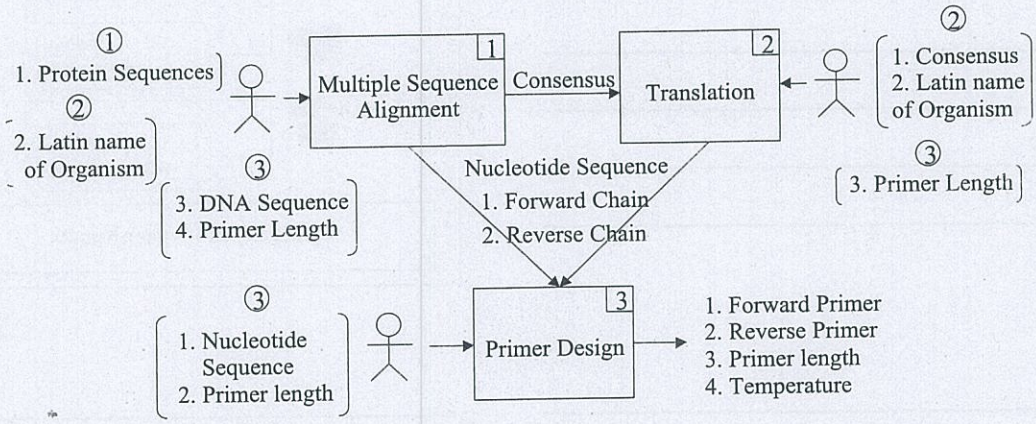


Figure 4 The automatic primer design system.

We further develop the current primer design system to be an automatic tool that can provide the flexible and efficient performance (Figure 4). The experts can start and finish at any process that they desire, and they can manually or automatically operate with our system. With the intention of solving the discontinuous workflow problem, we present three consensus selection algorithms and integrate them with the alignment and the primer design process.

Figure 5-8 demonstrate our primer design system "KUPDS". Which was implemented based on the Web-based application.

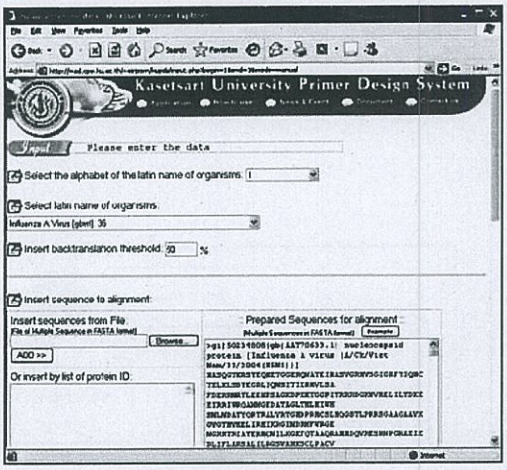


Figure 5. The implementation of the web-based automatic primer design system "KUPDS". This figure shows the entering page to insert the input data.

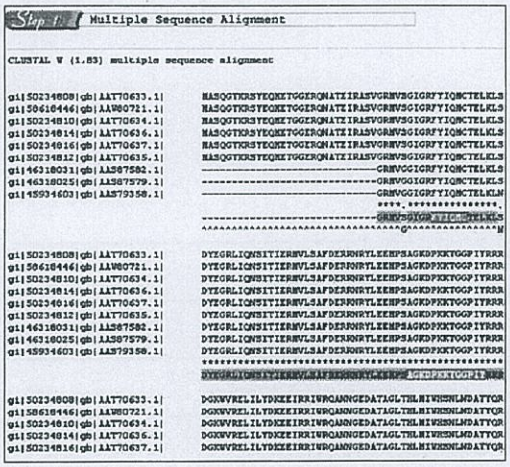


Figure 6. The multiple sequence alignment process. This program will automatically preview the least degenerate region of the consensus.

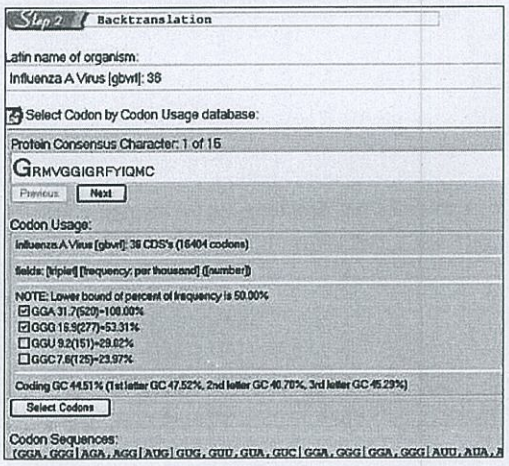


Figure 7 The backtranslation process. This program can automatically translate the protein consensus to the nucleotide sequence using the Codon Usage database and DNA backtracking.

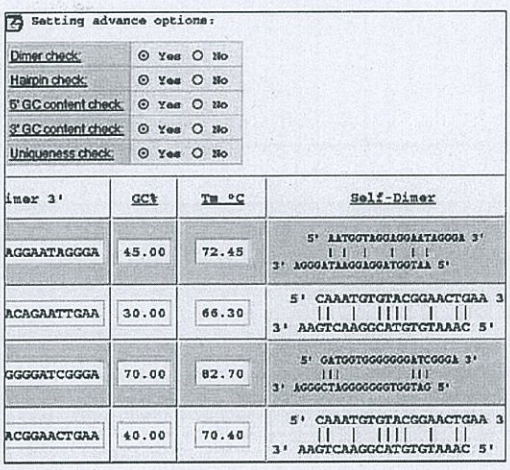


Figure 8 The primer design process. This program can design degenerate primer. It can design both PCR primer and probe. The expert can edit the sequence and specify more features of the designing. Our program includes the primer analysis tool.

4. CONCLUSIONS

In this paper, we propose three new algorihms for the consensus selection algorithms, which are DNA backtracking, Conserve Block and Island algorithm. Our algorithms have an ability to select the most appropriate least, high or non-degenerate consensus with the longest length. We also propose the new consensus scoring function that has an advantage that it can output the consensus with different degeneracy and different lengths that match the user requirements. These algorithms have been implemented as the automatic primer design system. The system can deal with the large set of protein sequences and support for both PCR primers and probes. It can support many multiple sequence alignment tools and can easily apply to select the high degenerate consensus for the undiscovered gene research.

5. ACKNOWLEDGEMENTS

This research is supported by Kasetsart University Research and Development Institute (KURDI) under grant No. 12008582.

REFERENCES

[1] Chen, S. H., Lin, C.Y., Cho, C.S., Lo, C.Z. and Hsiung, C.A. **2003**. Primer Design Assistant (PDA). Nucleic Acids Res. 31; 3751 - 3754.

[2] Thompson, J., Higgins, D., and Gibson, T. **1994**. CLUSTAL W. Nucleic Acids Res. 22; 4673-4690.

[3] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. **1997**. CLUSTAL X. Nucleic Acids Res. 25; 4876-4882.

[4] Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S. **2005**. ProbCons. Genome Res. 15;330-340.

[5] Henikoff, S., Henikoff, J.G., Alford, W.J. and Pietrokovski, S. **1995**. Block Maker. Gene. 163; 17-26.

[6] Schuler, G. D., Altschul, S. F., and Lipman, D. J. **1991**. A workbench for multiple alignment construction and analysis. Proteins. 9;180-190.

[7] Rose, T.M., Henikoff, J.G. and Henikoff, S. **2003**. CODEHOP. Nucleic Acids Res. 31; 3763-3766.

[8] WWW. http://www.bibiserv.techfak.uni-bielefeld.de/genefisher/

[9] Xintao Wei, David N. Kuhn, Giri Narasimhan. **2003**. Degenerate Primer Design via Clustering. IEEE Computer Society Bioinformatics Conference (CSB'03), pp 75.

[10] WWW. http://www.PremierBiosoft.com

[11] Rozen, S. and Skaletsky, H. **2000**. Primer3. Krawetz, S. and Misener, S. (eds.). Bioinformatics Methods and Protocols. Humana Press, Totowa, NJ, pp 365-386.

[12] WWW http://www.seq.yeastgenome.org/cgi-bin/web-primer

[13] Li, L-C. and Dahiya, R. **2002**. MethPrimer. Bioinformatics. 18;1427-1431.

[14] Sankoff D., Morel C. and Cedergren RJ. **1973**. Evolution of 5S RNA and the non-randomness of base replacement. Nature NewBiol. 245; 232-234.

[15] Linhart, C. and Shamir, R. **2002**. The degenerate primer design problem. Bioinformatics. 18 Suppl. 1: S176-S180.