Dynamic Programming for Single Nucleotide Polymorphism ID Identification in Systematic Association Studies

Cheng-Hong Yang,[1] *Li-Yeh Chuang,*[2] *Yu-Huei Cheng,*[1] *Cheng-Hao Wen,*[3] *and Hsueh-Wei Chang*[3,4,5]
[1]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences; [2]Department of Chemical Engineering, I-Shou University; [3]Faculty of Biomedical Science and Environmental Biology, [4]Graduate Institute of Natural Products, College of Pharmacy, and [5]Center of Excellence for Environmental Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan.

Single nucleotide polymorphisms (SNPs) play an important role in personalized medicine. However, the SNP data reported in many association studies provide only the SNP nucleotide/amino acid position, without providing the SNP ID recorded in National Center for Biotechnology Information databases. A tool with the ability to provide SNP ID identification, with a user-friendly interface, is needed. In this paper, a dynamic programming algorithm was used to compare homologs when the processed input sequence is aligned with the SNP FASTA database. Our novel system provides a web-based tool that uses the National Center for Biotechnology Information dbSNP database, which provides SNP sequence identification and SNP FASTA formats. Freely selectable sequence formats for alignment can be used, including general sequence formats (ACGT, [dNTP1/dNTP2] or IUPAC formats) and orientation with bidirectional sequence matching. In contrast to the National Center for Biotechnology Information SNP-BLAST, the proposed system always provides the correct targeted SNP ID (SNP hit), as well as nearby SNPs (flanking hits), arranged in their chromosomal order and contig positions. The system also solves problems inherent in SNP-BLAST, which cannot always provide the correct SNP ID for a given input sequence. Therefore, this system constitutes a novel application which uses dynamic programming to identify SNP IDs from the literature and keyed-in sequences for systematic association studies. It is freely available at http://bio.kuas.edu.tw/SNPosition/.

Key Words: association studies, BLAST, BLAT, dynamic programming, single nucleotide polymorphisms
(*Kaohsiung J Med Sci* 2009;25:165–76)

Single nucleotide polymorphisms (SNPs) are the most common genetic variations in heredity. An SNP is a variation of the DNA sequence caused by the replacement of one nucleotide with another, insertion of one or more nucleotides, or nucleotide deletion. Knowledge of SNPs constitutes useful information for personalized medicine [1,2]. Using SNPs, a variety of genetic diseases can be diagnosed, and side effects of medication can be prevented, thus increasing therapeutic efficiency [3].

Comparative SNP searches are very useful; therefore, software packages have been developed to facilitate comparisons in the literature. SNPper [4] and FESD [5], for example, rely on the chromosome position, the cytogenetic band position and the name to query information. SNPHunter [6] provides SNP screening, selection and acquisition. Genewindow provides an interactive tool for visualization of genomic variations [7].

Received: Dec 26, 2008 Accepted: Mar 27, 2009

Address correspondence and reprint requests to: Dr Hsueh-Wei Chang, Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, 100 Shih-Chuan 1st Road, Kaohsiung 807, Taiwan.
E-mail: changhw@kmu.edu.tw

However, these systems are incapable of screening a manually keyed-in sequence for SNP ID identification.

Recently, many servers have begun to provide screening functions for SNP sequences; both SNPServer [8] and GeneSNPs [9] provide a Basic Local Alignment Search Tool (BLAST) function for a sequence; however, the BLAST function, by its very nature, produces some problematic results, which will be described later. Similarly, the National Center for Biotechnology Information (NCBI) BLAST [10] showed different results when different sequences were used, such as those from blastn and megablast. The NCBI dbSNP [11] contains a BLAST program for SNPs called SNP-BLAST [12]. SNP-BLAST contains numerous databanks that provide the BLAST function, but there are still problems when performing an SNP ID query using manually keyed-in sequence. For instance, when using the blastn function of SNP-BLAST *with* megablast and blastn *without* megablast to perform the BLAST function for a partial sequence, the results did not always show the original keyed-in rs#. The BLAST results did not provide high scores or high expected values (which means that the specificity is low), or a corresponding SNP ID could not be found (please see details in the Results and Discussion section). Therefore, in order to correctly query the SNP ID, the system described here uses a dynamic programming algorithm to compare homologs of the input sequences with the FASTA sequence database to avoid such problems.

We used a JAVA-based server to solve all of the above problems. This server includes data for human, mouse and rat SNPs. The system uses dynamic programming [13–15], which allows users to key-in the sequence in different formats, and provides SNP IDs within the input range, as well as allowing users to identify NCBI SNP IDs and user-defined SNPs in an unknown sequence. In addition, a search function for SNP FASTA formats is provided, which overcomes the limitations of the NCBI system, in which SNP-BLAST does not provide only the correct SNP ID.

Methods

System design

A dynamic programming algorithm is one of the most widely used paradigms in computational molecular biology; it can evaluate a search space in polynomial time. Dynamic programming algorithms have been successfully applied to sequence alignment, gene recognition, RNA structure prediction, and to tag SNP selection. In this paper, we describe the development of a web-based SNP information platform. The system provides users with an input function for unknown sequences and performs sequence alignments with the rs_FASTA sequence in the databanks. The alignment results show the SNP ID, contig position, and FASTA sequence information within the input sequence (a sequence in FASTA format begins with a single-line description, followed by lines of sequence data). The system also searches for 5′ and 3′ flanking sequences (not including alleles) of SNP FASTA sequences to provide related information about nearby SNPs. Finally, the SNP information is presented graphically (see Figures 1D and 1E, and Figure 2).

The flowchart for our system, shown in Figure 3, is divided into four modules: an input module, a sequence data preprocessing module, a sequence alignment module, and a visualization module. In the input module, two types of input formats, sequence input and ID input, are allowed. For the sequence input formats, the users may input non-preprocessed DNA sequences or SNP sequences of alleles. The input sequence is then transferred into the sequence data preprocessing module which deletes unnecessary blanks and filters for non-base symbols. The processed sequence is then aligned with the FASTA sequence database in the sequence alignment module. Finally, the alignment results are graphically depicted in the visualization module. The database used by this system contains data for human (ftp://ftp.ncbi.nih. gov/snp/human/), mouse (ftp://ftp.ncbi.nih.gov/snp/mouse/) and rat (ftp://ftp.ncbi.nih.gov/snp/rat/) SNPs in NCBI dbSNP version b126 [11]. Each module in our system is described in more detail below.

Input module

In the input model, a free sequence format is allowed; DNA sequence formats (A, C, G, or T), dNTP1/dNTP2 sequence formats, and the IUPAC format can be used. The input sequence is processed for SNP ID retrieval in the input orientation. If no SNP is found, the input sequence is automatically transferred to its complementary sequence for re-retrieval.

Sequence data preprocessing module

The purpose of this module is to delete unnecessary blanks and line-feeds, and to filter non-base symbols

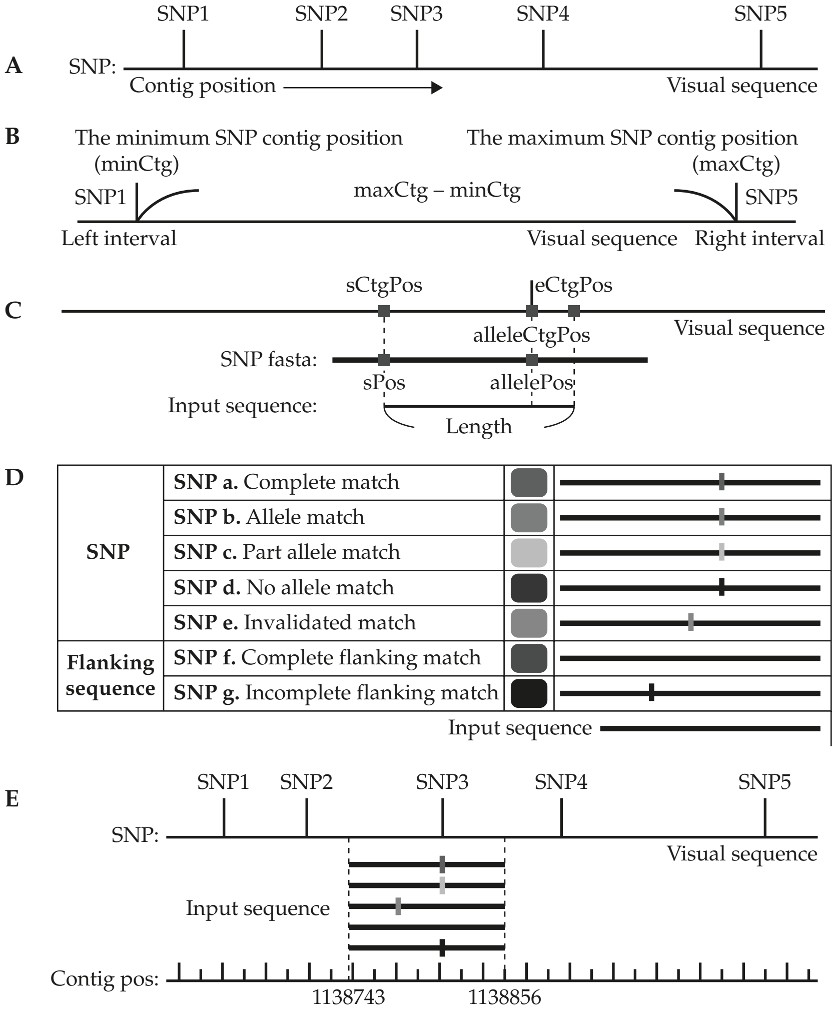
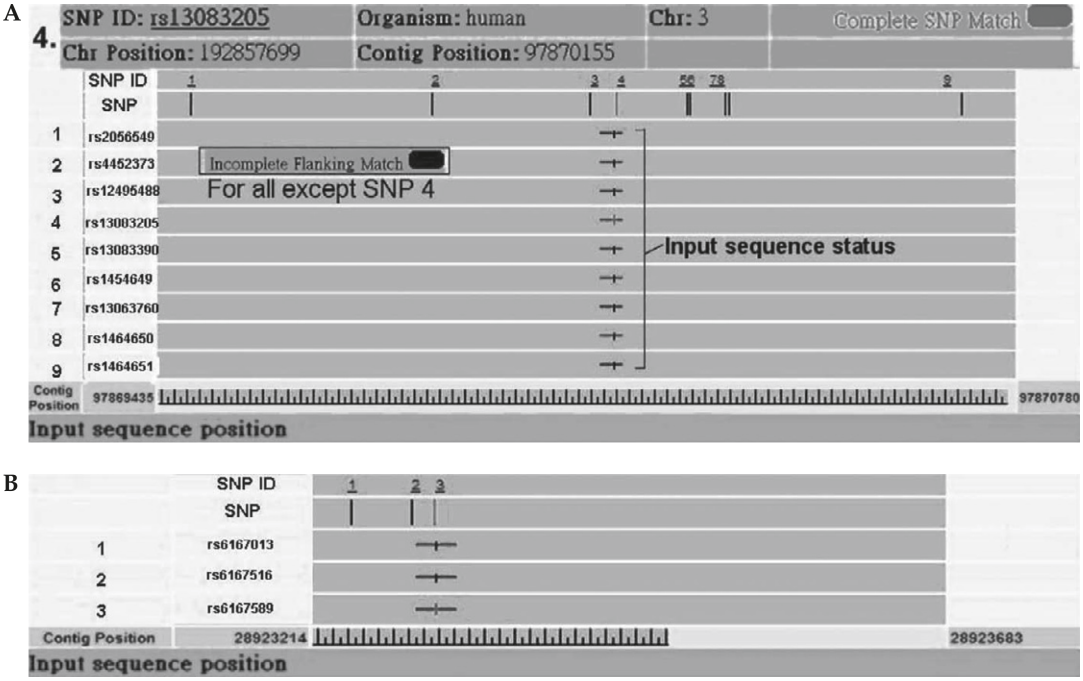


Figure 1. *Examples and principles for visualization of the output of target single nucleotide polymorphism (SNP) alignment. (A) The related SNPs 1–5 are linearly arranged according to the contig position of the visualized sequence. (B) A graph of the whole sequence based on the interval distance and flanking region of both sides. This range is then reflected to the ruler with a scale, as shown in (E). (C) A related position of the input sequence to the SNP FASTA sequence and the visualized sequence. Only one SNP FASTA sequence is shown although several overlapping SNP FASTA sequences are projected into the visualized sequence. (D) SNP color conditions for SNPs a–g (left side) and the results of the alignment (right side) are displayed individually. The conditions for SNPs a–e and SNPs f and g are discussed with the SNP and flanking region, respectively. Refer to the example with sequences in the text. (E) The alignment results of the input sequence with respect to the visualized sequence. The SNP linear arrangement is based on the contig position.*

other than A, T, C, and G. The IUPAC symbols M, R, W, S, Y, K, V, H, D, B, and N are reserved for the transformation of [dNTP1/dNTP2] to the IUPAC format.

Sequence alignment module

The processed input sequence is aligned with the FASTA sequence database, and a dynamic programming method [13–15] is used to compare their homologs. To determine whether the user's input sequence contains variations or not, the sequence is aligned with the SNP FASTA sequence data. The system compares the bases of an input data sequence one with the SNP rs_FASTA sequences to identify the sequence with the greatest similarity. First, the SNP FASTA sequence and the input sequence of the suffix edit distance E are calculated. Where P(i) represents the base of the user's input sequence in index i, i=1, 2, …, m, and m is the user's input sequence length; T(j) is the base of SNP FASTA sequence in index j, j=1, 2, …, n, and n is the SNP FASTA sequence's length. The procedure for the suffix edit distance is given below. The computational performance when carrying out the proposed analysis is O(Nmn), where



A

4. SNP ID: rs13083205 | Organism: human | Chr: 3 | Complete SNP Match

Chr Position: 192857699 | Contig Position: 97870155

	SNP ID	
1	rs2056549	
2	rs4452373	Incomplete Flanking Match
3	rs12495488	For all except SNP 4
4	rs13083205	
5	rs13083390	Input sequence status
6	rs1454649	
7	rs13063760	
8	rs1464650	
9	rs1464651	

Contig Position 97869435 … 97870780

Input sequence position

B

	SNP ID
1	rs6167013
2	rs6167516
3	rs6167589

Contig Position 28923214 … 28923683

Input sequence position

Figure 2. *The visual output of our system can correctly match single nucleotide polymorphisms (SNP), for example: (A) rs13083205; and (B) rs6167569. The SNP ID rs#, organism, chromosome location, SNP and flanking hits, as well as their chromosome contig positions are displayed. In this system, the relationship between these SNPs is provided in the order of chromosome contig. The SNP hit and flanking hits for the SNP are marked by red and gray colors, respectively.*

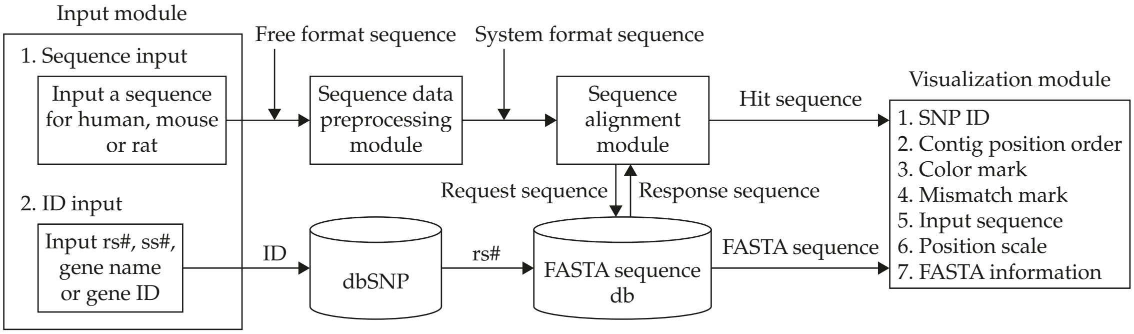


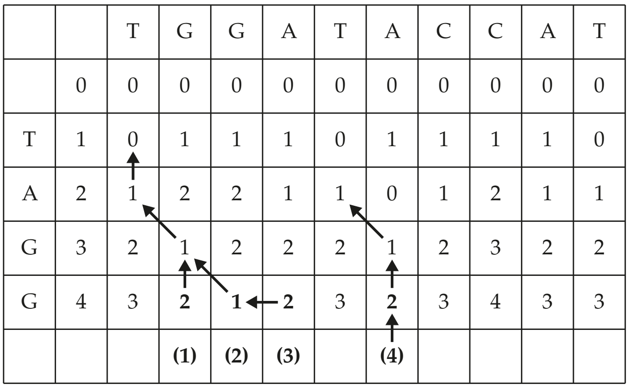
Figure 3. *Block diagram of the proposed system.*

N is the number of SNPs in NCBI dbSNP, and m and n are as described above.

The procedure for determining the suffix edit distance E:

```
for i ← 0 to m do
    E(i, 0) ← i
end
for j ← 0 to n do
    E(0, j) ← 0
end
for i ← 1 to m do
    for j ← 0 to n do
        if(T(j) = P(i)) then
            E(i, j) ← (i − 1, j − 1)
        else
            min ← MIN[E(i − 1, j), E(i, j − 1)]
            E(i, j) ← min + 1
    end
end
return E
```

To find the partially homologous sequence for users, the maximum error tolerance for the input sequence is accepted (Equation 1). Once the error count for aligning P and T is equal to or less than the maximum error tolerance, the input sequence is successfully aligned with the SNP FASTA sequence.



		T	G	G	A	T	A	C	C	A	T
	0	0	0	0	0	0	0	0	0	0	0
T	1	0	1	1	1	0	1	1	1	1	0
A	2	1	2	2	1	1	0	1	2	1	1
G	3	2	1	2	2	2	1	2	3	2	2
G	4	3	**2**	**1**	**2**	3	**2**	3	4	3	3
			(1)	**(2)**	**(3)**		**(4)**				

Figure 4. *Homologous alignment and possible homologous sequences.*

$$\text{The maximum error tolerance number} = (\text{input sequence length}) \times (\text{tolerant error rate}) \quad \text{(Equation 1)}$$

The homologous sequence can be obtained using the previously determined suffix edit distance E and the maximum error tolerance number based on backwards dynamic programming. Once the suffix edit distance E is less than or equal to the maximum error tolerance number, it is processed. The backward sequence is the homologous sequence that fits with the analog. For example, if an input sequence contains the bases (nucleotides) TAGC, the maximum tolerance error rate is 20%. When the input sequence is aligned with the 10-bases-long SNP FASTA sequence, TGGATACCAT, the maximum error tolerance number is $10 \times 0.2 = 2$. In other words, only two or fewer error alignments are allowed in this case (Figure 4). The boldface arrows in Figure 4 indicate the output of an agreeable homologous alignment, for which the homologous sequences are: (1) TG; (2) TGG; (3) TGGA; and (4) TA.

Visualization module

In the visualization module, the SNPs to be compared and the variations in the SNP flanking regions are presented graphically. The SNPs compared (e.g. SNP1–SNP5) are linearly arranged according to their contig positions, as shown in Figure 1A. The visualization graph of the whole sequence is output based on the interval distance and the flanking regions on both sides, as shown in Figure 1B. Because the input sequence only shows the sequence length, we obtained a similar initial position and allele position for the SNP FASTA sequence using a similar alignment (Figure 1C). In addition, we determined the initial contig position (sCtgPos) and the final contig position (eCtgPos) of the input sequence by using the allele contig position (allelCtgPos) of the SNP within the whole sequence as shown in Figure 1C. The calculation is shown below (Equations 2 and 3).

$$\text{sCtgPos} = \text{alleleCtgPos} - (\text{allelePos} - \text{sPos}) \quad \text{(Equation 2)}$$
$$\text{eCtgPos} = \text{sCtgPos} + \text{length} \quad \text{(Equation 3)}$$

Because the comparison reveals many different results, the proposed system shows the different results coded in different colors, as illustrated in Figure 1D. For alignment with targeting to the NCBI SNP, SNPs a–e are displayed with different colors under different circumstances in Figure 1D. For a flanking sequence alignment without targeting to the NCBI SNP, SNPs f and g are also displayed with different colors. Moreover, the matched SNPs or variations in the flanking sequence are displayed individually in the input sequence, as shown on the right hand side of Figure 1D. Finally, the results of the input sequence are homologous to the SNP linear alignment according to the contig position (Figure 1E). When the visualization is processed, the minimum and maximum contig positions, e.g. 1138743 and 1138856 respectively, must be found within all of the matched SNPs.

Results and Discussion

Ninety percent of polymorphisms in the human genome are SNPs. Knowledge of these SNPs is regarded as an important resource for personalized medicine studies. A table containing 172 different sequences from 23 different chromosomes is provided online at http://bio.kuas.edu.tw/dynamicSNP/ncbi-compare.jsp, and the results obtained by SNP-BLAST in NCBI and our proposed system are compared. Here, the input sequence containing *Homo sapiens* rs13083205 [A/C] for SNP-BLAST of NCBI is described below. As an example, we choose an input sequence copied from an SNP ID—namely for rs13083205 SNP ID [*Homo sapiens*]: AAAATTAATATTCTTCAAAATT**M**TTTCTT CTAAAT (the length of the input sequence is shown in blue as indicated by arrows in Figure 2A; **M** indicates the SNP site).

When the blastn *with* megablast (default in the NCBI) SNP-BLAST programs are used, these sequences gave different results for rs13083205. When

Input 1: AAAATTAATATTCTTCAAAATT**M**TTTCTTCTAAAT

	Blastn with Megablast	**Score**(Bits)	**E** Value	**Blastn without Megablast**	**Score**(Bits)	**E** Value	
A	No significant similarity found			gnl|dbSNP|rs13083390	60.8	4e-08	E
				gnl|dbSNP|rs13083205	**60.8**	**4e-08**	
				gnl|dbSNP|rs13063760	60.8	4e-08	
				gnl|dbSNP|rs12495488	60.8	4e-08	
				gnl|dbSNP|rs4452373	60.8	4e-08	
				gnl|dbSNP|rs2056549	60.8	4e-08	
				gnl|dbSNP|rs1464651	60.8	4e-08	
				gnl|dbSNP|rs1464650	60.8	4e-08	
				gnl|dbSNP|rs1464649	60.8	4e-08	

Input 2: AAAATTAATATTCTTCAAAATT**[A/C]**TTTCTTCTAAAT

				Blastn without Megablast	**Score**(Bits)	**E** Value	
B	No significant similarity found			gnl|dbSNP|rs13083390	57.2	6e-07	F
				gnl|dbSNP|rs13063760	57.2	6e-07	
				gnl|dbSNP|rs12495488	57.2	6e-07	
				gnl|dbSNP|rs4452373	57.2	6e-07	
				gnl|dbSNP|rs2056549	57.2	6e-07	
				gnl|dbSNP|rs1464651	57.2	6e-07	
				gnl|dbSNP|rs1464650	57.2	6e-07	
				gnl|dbSNP|rs1464649	57.2	6e-07	
				gnl|dbSNP|rs13083205	**55.4**	**2e-06**	

Input 2: AAAATTAATATTCTTCAAAATT**A**TTTCTTCTAAAT

	Blastn with Megablast	**Score**(Bits)	**E** Value	**Blastn without Megablast**	**Score**(Bits)	**E** Value	
C	gnl|dbSNP|rs13083390	65.8	2e-09	gnl|dbSNP|rs13083390	64.4	4e-09	G
	gnl|dbSNP|rs13063760	65.8	2e-09	gnl|dbSNP|rs13063760	64.4	4e-09	
	gnl|dbSNP|rs12495488	65.8	2e-09	gnl|dbSNP|rs12495488	64.4	4e-09	
	gnl|dbSNP|rs4452373	65.8	2e-09	gnl|dbSNP|rs4452373	64.4	4e-09	
	gnl|dbSNP|rs2056549	65.8	2e-09	gnl|dbSNP|rs2056549	64.4	4e-09	
	gnl|dbSNP|rs1464651	65.8	2e-09	gnl|dbSNP|rs1464651	64.4	4e-09	
	gnl|dbSNP|rs1464650	65.8	2e-09	gnl|dbSNP|rs1464650	64.4	4e-09	
	gnl|dbSNP|rs1464649	65.8	2e-09	gnl|dbSNP|rs1464649	64.4	4e-09	
	gnl|dbSNP|rs13083205	**62.1**	**3e-08**	**gnl|dbSNP|rs13083205**	**60.8**	**4e-08**	

Input 3: AAAATTAATATTCTTCAAAATT**C**TTTCTTCTAAAT

	Blastn with Megablast	**Score**(Bits)	**E** Value	**Blastn without Megablast**	**Score**(Bits)	**E** Value	
D	No significant similarity found			**gnl|dbSNP|rs13083205**	**60.8**	**4e-08**	H
				gnl|dbSNP|rs13083390	59.0	1e-07	
				gnl|dbSNP|rs13063760	59.0	1e-07	
				gnl|dbSNP|rs12495488	59.0	1e-07	
				gnl|dbSNP|rs4452373	59.0	1e-07	
				gnl|dbSNP|rs2056549	59.0	1e-07	
				gnl|dbSNP|rs1464651	59.0	1e-07	
				gnl|dbSNP|rs1464650	59.0	1e-07	
				gnl|dbSNP|rs1464649	59.0	1e-07	
				gnl|dbSNP|rs57392751	44.6	0.003	
				gnl|dbSNP|rs12356542	44.6	0.003	

Figure 5. *Single nucleotide polymorphism (SNP) ID identification results obtained using the National Center for Biotechnology Information SNP-tool. The results obtained with SNP-BLAST are separated, showing blastn with megablast on the left side and blastn without megablast on the right side. SNP ID identification result of input sequence with SNP in: (A) International Union of Pure and Applied Chemistry format "M"; (B) [A/C]; (C) A; and (D) C.*

the sequence was input in the IUPAC format, i.e. AAAATTAATATTCTTCAAAATT**M**TTTCTTCTAAAT (Figure 5A) and [dNTP1/dNTP2] sequence format, i.e. AAAATTAATATTCTTCAAAATT**[A/C]**TTTC TTCTAAAT (Figure 5B), SNP-BLAST was unable to locate a significantly similar alignment, i.e. there is no matched SNP. Because the alleles of SNP ID rs13083205 are **A/C**, the input sequence is a general sequence containing the allele base "**A**" (Figure 5C); the results of SNP-BLAST were rs13083390, rs13063760, rs12495488, rs4452373, rs2056549, rs1464651, rs1464650, rs1464649, and **rs13083205**. The "scores" were all 65.8 and the "expected" value was 2e–09, except for rs13083205 (score = 62.1 and expected = 3e–08). Most of these SNP IDs provided in the BLAST results were unable to provide the correct SNP ID from the input sequence, i.e. the input rs13083205 SNP ID-containing sequence. Furthermore, the rs13083205 was hit with low performance (low score bits and high expected values) compared with the other hits. On the other hand, if the input sequence is a general sequence containing allele base "**C**" (Figure 5D), the results of SNP-BLAST show "no significant similarity found". Therefore, the search in SNP-BLAST is sequence format-dependent and differs on a case-by-case basis. In this case, SNP-BLAST does not reveal a single positive result for the C allele of rs13083205, whereas under other circumstances, no significant hit is retrieved by SNP-BLAST. In some cases, SNP IDs are hit without including the target SNP ID. These are common problems for SNP-BLAST in the NCBI when the BLAST function is performed by inputting a sequence to identify SNP ID or IDs within a sequence. These results can mislead researchers and prevent them from obtaining the correctly matched SNP ID in a trial.

Moreover, the input sequence produces different results when using the SNP-BLAST program blastn of SNP-BLAST in the NCBI *without* megablast. When the sequence is input in the IUPAC format "**M**" (Figure 5E), the BLAST results included a total of nine SNP IDs: rs13083390, **rs13083205**, rs13063760, rs12495488, rs4452373, rs2056549, rs1464651, rs1464650 and rs1464649 with a score of 67.2 and an expected value of 2e–09. However, all of the eleven SNPs in rs13083205 share the highest score; therefore, NCBI SNP-BLAST cannot provide the correct SNP ID(s) for this input sequence. Similarly, when the input sequence is entered in the SNP nucleotide format [**A/C**] (Figure 5F) or **A** (Figure 5G), nine matched SNPs are retrieved by SNP BLAST: rs13083390, rs13063760, rs12495488, rs4452373, rs2056549, rs1464651, rs1464650, rs1464649 and **rs13083205** with a score of 55.4–57.2 or 60.8–64.4 and an expected value of 4e–08 to 4e–09 or 2e–10 to 4e–09, respectively. In this case, however, the matched rs13083205 did not have the highest score; thus users are unable to determine whether it is the correct SNP ID. For the sequence AAAATTAATATTCTTCAAAATTCTTT**C**TTCTAAAT (Figure 5H), nine matched SNPs were retrieved by SNP BLAST: **rs13083205**, rs13083390, rs13063760, rs12495488, rs4452373, rs2056549, rs1464651, rs1464650, rs1464649, rs57392751, and rs12357542, with a score of 44.6–60.8 and an expected value of 0.003 to 4e–08. In this case, rs13083205 showed the highest score with SNP-BLAST. However, different kinds of sequence formats may have different performance for SNP ID retrieval in the NCBI SNP-BLAST tool. In general, this will cause confusion, and users cannot precisely identify the target SNP ID for a given input sequence.

The NCBI SNP-BLAST tool shows similar problems for other types of sequences from other species, such as *Mus musculus*. For example, the rs6167569 sequence is TCTTGCGTAGATCCGTCACAGCCCT[**C/T**]TTTCACCCGCCAGGGCTCCCGACAA. When using blastn *with* megablast, the allele of rs6167569 with T provided the correct SNP IDs in SNP-BLAST with a score 93.5 and an expected value of 7e–18, but the allele of rs6167569 with Y, C and [C/T] did not match the correct SNP ID. Using blastn *without* megablast, rs6167569 contains different sequences, including [C/T], Y, C and T. They all provided the same BLAST results with three SNP IDs: rs6167569, rs6167516 and rs6167013; of these, rs6167569 attained the highest scores. Therefore, NCBI SNP-BLAST may provide too much data to determine a correct SNP ID for the input sequence, or it cannot provide the SNP ID at all. In addition, the SNPs and flanking hits are not marked.

In our novel system, we used the same sequences for rs13083205, located in human chromosome 3, as for SNP-BLAST. The sequence contains SNPs in the nucleotide format R, [A/G], A or G, and the choice for mismatched bases is 1, by default. The four sequence inputs show the same results, and all are correctly matched to the SNP sequence rs13083205, as shown in Figure 2A. Similarly, with rs6167569, located in mouse chromosome 15, the different sequences including [C/T], Y, C and T, were all matched correctly to only a single SNP, rs6167569, as shown in Figure 2B. In addition, the nearby SNP IDs rs6167013 and rs6167516 are provided as flanking hits distinguished from the SNP hit itself. This provides a complete topography of the input sequence, and it corrects the inherent problem of the NCBI SNP-BLAST tool. In Figure 2A and 2B, the SNP contig position is presented graphically. The color coding of the visualization helps the biologists to identify all of the relevant data contained in the SNP sequence. In addition, users can test their SNPs for novelty if they are unable to find the corresponding SNP ID in our proposed system, i.e. user-defined SNPs marked with a bar were discovered.

We designed an experiment to test how often the SNP ID within the input sequence is incorrectly provided using the method shown in Figure 5. Using 172 different sequences from 23 different chromosomes (a list is provided online at http://bio.kuas.edu.tw/dynamicSNP/ncbi-compare.jsp), we found that our system achieved a success rate of 100%, while the NCBI SNP-BLAST tool, *with* and *without* the megablast function, provided the complete score matched (similar to Figure 5H; correct SNP ID hit is the single highest score among hits) for 13.34% versus 40.70%, the high score matched (Figure 5E; several hits share the highest score but the correct SNP ID hit is not shown in the first row) for 11.05% versus 34.88%, no high score matched (Figures 5C, 5F and 5G; the correct SNP ID hit is not the highest score) for 6.98% versus 20.35%, and no matched or no significance (similar to Figures 5A, 5B and 5D; no hits provided) for 68.60% versus 4.07%, respectively. These results are provided online at the website address given above. For complete-score-matched, high-score-matched and no-high-score-matched conditions, the system is able to provide the correct SNP ID hits, although it is difficult for most users to identify them. In contrast, no-matched and no-significance situations show false-negative results for the SNP ID hit.

The NCBI BLAST tool uses a heuristic algorithm that seeks local, as opposed to global, alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity [16]. It obtains all statistically significant alignments. However, it does not explore the entire search space between two sequences. This is the reason why BLAST sometimes cannot hit the correct SNP ID. The overlapping nature of several SNP flanking sequences found by the NCBI SNP-BLAST tool is shown in Figure 6 and in our previous results with SNP ID-info software [17]. This may partly explain why false-positive hits (providing many hits rather than the correct SNP ID hit alone) for NCBI SNP-BLAST were common in our 172 tests with different sequences from 23 different chromosomes. Many SNP flanking sequences commonly overlap each other, although the SNPs are located separately.

Recently, the University of California Santa Cruz Genome Browser Database [18] was developed to

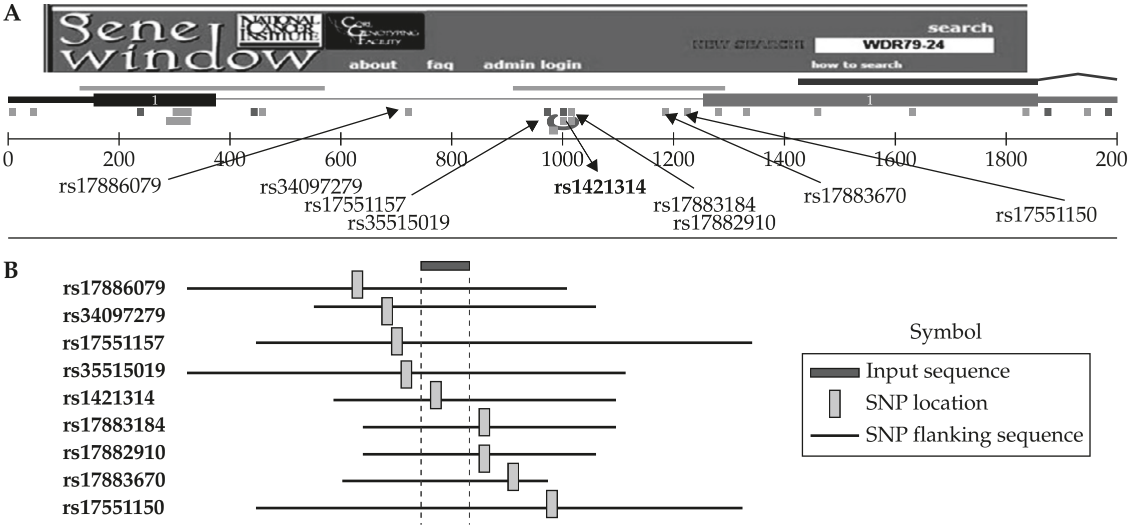


Figure 6. *Example of the overlapping nature of several sequences containing single nucleotide polymorphisms. (A) Neighboring environmental viewing using Genewindow software. Nine single nucleotide polymorphisms located in chromosome 17 were selected as an example. The relative locus of the chromosomal contig position is provided. (B) The overlapping relationship between the nine single nucleotide polymorphisms flanking sequences can be seen in (A).*

Table. Comparison of our novel system and the National Center for Biotechnology Information search tool SNP-BLAST

Information for SNP ID retrieval	NCBI SNP-BLAST	Proposed system
Method	BLAST	Dynamic programming/BLAST*/BLAT* (*is used for the primary search for the chromosome location to improve the subsequent SNP identification).
Sequence source	Only SNP FASTA sequences	SNP FASTA sequences are integrated with the chromosome and contig positions.
Accuracy	*For input sequence containing NCBI SNP:* 1. Indistinguishable to the SNP hit and flanking hits. 2. Sometimes, no significant hit for SNP ID. 3. No environment for SNP ID map, i.e. without nearby SNPs map in sequence. *For input sequence without NCBI SNP within:* 1. Cannot confirm whether any SNP IDs exist within the input sequence.	*For input sequence containing NCBI SNP:* 1. Provides the correct target SNP ID (or IDs) for input sequence (SNP hit). 2. Provides the flanking SNP IDs, i.e. SNP nearby but not within the input sequence (flanking hits). *For input sequence without NCBI SNP within:* 1. Provides SNPs with flanking hits if no SNP hit is present within the input sequence. 2. Presents the position and range of the input sequence under the SNP FASTA sequence and environment by visualization.
Positional information (chromosome/contig positions)	1. Not shown immediately or systemically. 2. SNP IDs are not listed in the order of chromosome and contig positions.	1. Provided immediately and systemically. 2. All retrieved SNP IDs are displayed in the order of chromosome and contig positions.
Analytic parameters	1. Length and SNP density of the input sequence are proportional to SNP number retrieved from input sequence. 2. Megablast: different BLAST results for SNP-BLAST with or without megablast function. 3. Listed according to order of score and E values. a. Highest score and lowest E values are not always representative correct SNP ID. b. Provides too many SNP IDs with same performance (score and E values) to call. c. The score and E values of SNP-BLAST make no sense for the SNP ID identification of an input sequences.	1. Length of input sequence: Retrieval SNP number is increasing with the length of input sequence. 2. No megablast choice problem. Consistent results for SNP ID identification. 3. SNPs listed in the order of chromosome and contig positions.
Display type	1. SNP ID panel listed. 2. Listed in the order of score and E values. 3. SNP and flanking hits not marked.	1. Visualization for the range and position of the input sequence under SNP ID environment. 2. Color-match provides illustration of SNP and flanking hits. 3. Only the correct SNP hit is provided. Nearby SNPs, which are not included in the input sequence, are provided and marked as flanking hits.
Major feature	Searches similar flanking sequences for SNP ID.	1. Identifies SNP ID within an input sequence. 2. Provides SNP hit and flanking hits in the order of chromosome and contig positions.

SNP = single nucleotide polymorphism; ID = SNP reference cluster identification card number; NCBI = National Center for Biotechnology Information; BLAST = Basic Local Alignment Search Tool; BLAT = BLAST-Like Alignment Tool; E = expected value.

provide sequence and annotation data for the genomes of many species and model organisms. With its BLAST-Like Alignment Tool (BLAT) function [19], it provides genome annotations including SNP IDs within the input sequence that are similar to our novel system. However, SNPs and flanking hits are not provided by BLAT. This problem is similar to that of SNP-BLAST in terms of the acceptable sequence formats. The results are sometimes sequence format-dependent and many browser hits are provided for selection. The IUPAC code and [dNTP1/dNTP2] formats for SNP are not suitable for BLAT because this sometimes provides the wrong chromosome location and many browsers with different chromosome locations. Furthermore, SNP-BLAST and BLAT do not indicate mutations other than the SNPs in the NCBI dbSNP. In contrast, our proposed system marks both user-defined SNPs and SNPs in dbSNP. However, because SNP-BLAST and BLAT provide fast SNP retrieval, we introduced these functions into our proposed system for sequences with unknown chromosome locations to improve the screening speed. By using dynamic programming in our system, we overcame the limitations inherent in SNP-BLAST and BLAT, and free formats of SNP input sequences are accepted.

Dynamic programming has been used for sequence alignment [13–15], and it has recently been applied to haplotype block studies [20,21]. However, none of these studies performed SNP ID targeting for input sequences. To our knowledge, we are the first to identify SNP IDs in NCBI rs# using dynamic programming. A general comparison of the performance between NCBI SNP-BLAST and our proposed system is shown in the Table. The results indicate that our system is more precise and more informative for SNP ID identification than the NCBI SNP-BLAST, particularly regarding SNP ID targeting and general information of the SNP. Flanking hits are provided in the order of the chromosome and contig positions. Once users obtain an SNP ID from NCBI dbSNP, software, including our newly developed SNP-RFLPing [22], can provide SNP information for genotyping in association studies.

Conclusion

The NCBI BLAST is currently the most commonly used sequence alignment tool. It continues the heuristic calculation concept and requires a shorter time to search for alignment similarities. Nevertheless, it has limitations, such as its inherent problem in the BLAST program, which uses megablast, and the fact that it cannot always correctly align an input sequence. In this paper, we described the development of a web-based SNP system that uses a dynamic programming algorithm to compare homologs. This system provides the correct SNP ID when entering a sequence, and the reference cluster ID "rs", the NCBI assay ID, the "ss", the gene name and the gene ID, which facilitates faster searching of SNP FASTA information. This not only solves the inherent limitations of NCBI SNP-BLAST, but also assists biologists trying to establish a systematic database while conducting association research.

Acknowledgments

This work was partly supported by the National Science Council in Taiwan under grants 96-2221-E-214-050-MY3 and 97-2622-E-151-008-CC2, and by the grants KMU-EM-97-1.1a and KMU-EM-98-1.4.

References

1. Suh Y, Vijg J. SNP discovery in associating genetic variation with human disease phenotypes. *Mutat Res* 2005; 573:41–53.
2. Erichsen HC, Chanock SJ. SNPs in cancer research and treatment. *Br J Cancer* 2004;90:747–51.
3. Shastry BS. SNPs and haplotypes: genetic markers for disease and drug response (review). *Int J Mol Med* 2003; 11:379–82.
4. Riva A, Kohane IS. SNPper: retrieval and analysis of human SNPs. *Bioinformatics* 2002;18:1681–5.
5. Kang HJ, Choi KO, Kim BD, et al. FESD: a Functional Element SNPs Database in humans. *Nucleic Acids Res* 2005;33:D518–22.
6. Wang L, Liu S, Niu T, et al. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. *BMC Bioinformatics* 2005;6:60.
7. Staats B, Qi L, Beerman M, et al. Genewindow: an interactive tool for visualization of genomic variation. *Nat Genet* 2005;37:109–10.
8. Savage D, Batley J, Erwin T, et al. SNPServer: a real-time SNP discovery tool. *Nucleic Acids Res* 2005;33: W493–5.
9. University of Utah Genome Center. *GeneSNPs-blast function.* Available at: http://www.genome.utah.edu/genesnps/cgi-bin/blast.cgi [Date accessed: August 23, 2006]

10. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 2004;32:W20–5.
11. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–11.
12. National Center for Biotechnology Information. *SNP-BLAST.* Available at: http://www.ncbi.nlm.nih.gov/SNP/snp_blastByOrg.cgi [Date accessed: January 9, 2009]
13. Eddy SR. What is dynamic programming? *Nat Biotechnol* 2004;22:909–10.
14. Giegerich R. A systematic approach to dynamic programming in bioinformatics. *Bioinformatics* 2000;16: 665–77.
15. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;48:443–53.
16. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25: 3389–402.
17. Yang CH, Chuang LY, Cheng YH, et al. SNP ID-info: SNP ID searching and visualization platform. *OMICS* 2008;12:217–26.
18. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006;34:D590–8.
19. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;12:656–64.
20. Zhang K, Qin Z, Chen T, et al. HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* 2005;21:131–4.
21. Su SC, Kuo CC, Chen T. Inference of missing SNPs and information quantity measurements for haplotype blocks. *Bioinformatics* 2005;21:2001–7.
22. Chang HW, Yang CH, Chang PL, et al. SNP-RFLPing: restriction enzyme mining for SNPs in genomes. *BMC Genomics* 2006;7:30.

在系統相關性研究中使用動態規劃法找尋單核苷酸多型性身分證 (SNP ID)

楊正宏[1] 莊麗月[2] 鄭煜輝[1] 溫政浩[3] 張學偉[3,4,5]

[1]高雄應用科技大學 電子工程學系

[2]義守大學 化學工程學系

高雄醫學大學 [3]生物醫學暨環境生物學系 [4]天然藥物研究所 [5]環境醫學頂尖研究中心

單一核苷酸多型性 (single nucleotide polymorphisms、SNP) 在個人化醫學上扮演很重要的角色。然而，很多關聯性研究對於 SNP 資訊的描述僅提供 SNP 核苷酸或胺基酸的位置，而沒有提供 NCBI SNP ID。因此，一個能夠提供 SNP ID 辨識且具有友善使用者介面的工具是必需的。本研究使用動態規劃方法來比對輸入序列與 SNP FASTA 資料庫間的相似性，並且提出一個基於 NCBI dbSNP 資料庫的 web-based 系統，來提供 SNP 序列辨識和 SNP FASTA 格式。此系統能夠輸入各種格式的序列，包括一般化的序列格式 (如 ACGT、[dNTP1/dNTP2] 或 IUPAC 格式) 和不同方向的序列。相較於 NCBI SNP-BLAST，本研究提出的系統能依照染色體和 contig 位置提供正確的目標的 SNP ID (SNP hit)，而且也能提供鄰近的 SNPs (flanking hits)，並且也解決了 SNP-BLAST 對於輸入序列常常不能夠提供正確 SNP ID 的問題。因此，我們將動態規劃方法新穎地應用在 SNP ID 搜尋與辨識，讓使用者可以查詢文獻上的序列及輸入序列中所包含的 SNP ID，而提供系統性的相關性研究。這個程式在網頁 http://bio.kuas.edu.tw/SNPosition/ 可以免費使用。

關鍵詞：關聯性研究，BLAST，BLAT，動態規劃方法，單一核苷酸多型性

(高雄醫誌 2009;25:165–76)

收文日期：97 年 12 月 26 日
接受刊載：98 年 3 月 27 日

通訊作者：張學偉副教授
高雄醫學大學生物醫學暨環境生物學系
高雄市三民區十全一路 100 號