SNP 資訊平台之建構

莊麗月	楊正宏	鄭煜輝	張學偉
義守大學	高雄應用科技	高雄應用科技	高雄醫學大學
化學工程系	大學電子工程系	大學電子工程系	生物醫學
chuang	chyang	yuhuei.cheng	暨環境生物系
@isu.edu.tw	@cc.kuas.edu.tw	@gmail.com	changhw
			@kmu.edu.tw

摘要

單一核苷酸基因多型性(SNPs; single nucleotide polymorphism) 是遺傳上最常見的基因變異，生物學家可由檢驗出的 SNPs 來研究與基因有關的疾病。目前存在許多 SNP 相關軟體並無法滿足使用者需求。因此，本研究以 NCBI dbSNP 資料庫為基礎，建構在 web 上的 SNP 序列鑑定及 SNP fasta format 搜尋工具，提供自由格式的序列鑑定，包含一般序列格式(ACGT)、[dNTP1/dNTP2] 序列格式或 IUPAC 格式，及已知 SNP ID 的搜尋，包括 reference cluster ID “rs#”、 NCBI assay ID “ss#”、gene name 和 gene ID，可以從序列中找出 SNP ID 及未知的序列鑑定，並解決 NCBI SNP Blast 無法提供精確的 SNP ID 的問題。

關鍵詞：SNP、dbSNP、Blast、fasta、Web-based

一、前言

SNPs (single nucleotide polymorphism) 是目前遺傳上最常見的基因變異(Gene Variation)，主要是指 DNA 序列上的一個核苷酸被另一個核苷酸所取代或是被一個或幾個核苷酸的插入或缺失所造成。生物學家可由檢驗出的 SNP 來研究與基因有關的疾病或與現有的遺傳標記進行各種基因診斷，並避免藥物副作用以提高療效。

隨著 SNP 的相關研究逐漸被重視，其相關的軟體陸續被研發。例如，SNPper[1]可以藉由 position, cytogenetic band 和 name 搜尋 SNP ， SNPHunter[2] 提供 SNP screening 、 selelction 和 acquisition，然而這些系統卻無法藉由使用者輸入的序列而篩選出 SNP。最近，許多伺服器開始提供序列篩選 SNP 的功能，例如，SNPServer[3]提供序列 Blast，但是當測試部份 rs#所提供的 SNP ID 卻是失敗的。同樣的， BLAST 針對各種型態的序列查詢提供不同的功能，主要有 blastn、blastp、megablast、blastx、tblastn、tblastx…等[4]。在 NCBI SNP 的使用上[5]，如選擇 blast 的功能，有專門的 SNP BLAST 程式，即 SNP-BLAST[6]。在 SNP-BLAST 中有許多物種的資料庫可做 blast，然而透過輸入序列在尋找 SNP IDs 上仍有一些問題，例如以 SNP-Blast 使用程式 blastn with megablast 和 blastn without megablast 去 Blast 部分序列，結果即使有對應出原先輸入的 rs#，但卻沒有獲得高的分數(即專一性不足)或是找不到任何 SNP ID。因此，本研究實現 JAVA-based server 去解決上述問題包含 human、 mouse 和 rat SNPs，讓使用者能藉由輸入各種格式的序列而提供 SNP ID，並判別未知序列是否含有 SNP 序列，同時也提供 SNP fasta format 的搜尋功能。

二、研究方法

本研究提出一個 web-based SNP 資訊平台，可提供使用者輸入未知的序列與資料庫中的 rs_fasta 序列資料做比對，比對的結果可以呈現出此序列的 SNP ID 或辨別該序列是否為 SNP 序列，亦即輸入序列包含的鹼基中有

alleles，並且也搜尋此序列旁邊的 SNP 序列，以 fasta format 呈現，最後將搜尋到的序列以視覺化的相對位置方式呈現，以供生物學者做研究，並解決 NCBI Blast 無法提供精確的 SNP ID 的問題。

本資訊平台以 JAVA 程式語言為基礎，採用 MVC (Model-View-Controller) 架構來設計，將資料、程式邏輯和呈現外觀分開，如圖 1 所示，使用 JSP (Java Server Page)作為 View 來呈現資訊，Java Servlet 作為 controller 來控制導向流程以及 Java class 當作底層程式邏輯和與資料庫溝通。以下將分成四個部分:系統架構、資料庫結構、演算法及視覺化等來說明研究方法，詳細說明如下:

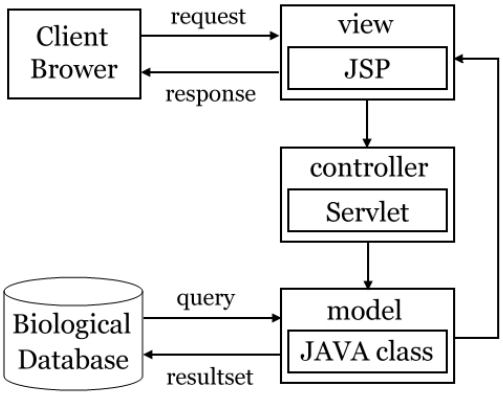


圖 1. MVC (Model-View-Controller)架構設計圖

2.1 系統架構

本資訊平台主要分為五個模組，如圖 2 所示，即(1)Input Module (2)Sequence Process Module (3)Sequence Alignment Module (4)Database (5)Output Module，簡述如下:

(1) Input Module

主要分為兩種輸入方式，第一種為 Sequence Input，即輸入未知的序列去鑑定 SNP，第二種為 ID input，即是輸入 SNP ID，如 rs#、ss#、Gene Name 或者是 Gene ID，以搜尋 fasta format 資訊。

(2) Sequence Process Module

把 Sequence Input 輸入方式所輸入的序列，經由去除空白、跳格，及濾除非鹼基核苷酸碼，把雜亂無章的序列轉換成系統可處理的序列。

(3) Sequence Alignment Module

把經由 Sequence Process Module 處理過的序列與 Fasta sequence database 中的序列做比對。

(4) Database

採用 MySQL database 存放 SNP 的 rs_fasta 的資料，建置 Fasta sequence database，並建置 dbSNP 資料庫，此 dbSNP 資料庫被架設在 MS SQL server 上面，以提供大量資料的快速存取，依照 dbSNP 所提供的參考手冊去 create local copy。

(5) Output Module

主要用來呈現比對及搜尋結果，包括 SNP hit、SNP Flanking hit 及 Fasta sequence，並視覺化其比對到的 SNP ID。

本 SNP 資訊平台處理流程說明如下，如圖 2 所示，當使用者將一個未知的序列，經由 Input Module 中的 Sequence Input 模式輸入，首先會將此序列送入 Sequence Process Module 去做處理，例如把多餘的 space 去除，以及將非鹼基的符號過濾，接著將處理過後的序列送 Sequence Alignment Module 與 Fasta sequence database 中的 SNP fasta sequence 做比對的動作，然後將比對之後的 result data 送至 Output Module 去呈現出結果。如果使用者輸入為 SNP

ID，如 rs#、ss#、Gene Name 或 Gene ID 則到 dbSNP database 搜尋所對應的 rs#，然後到 Fasta sequence database 把對應到的 fasta sequence 資料輸出。

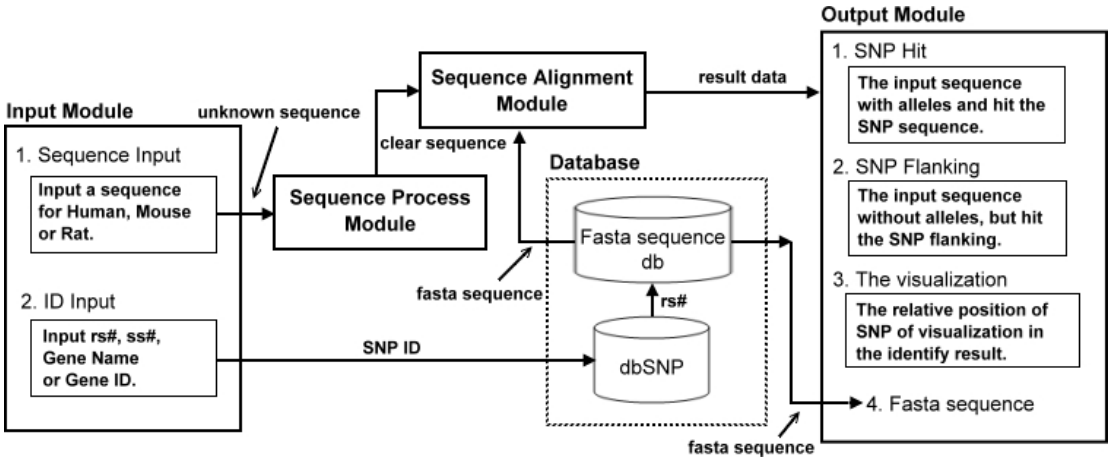


圖 2. 系統架構圖

2.2 資料庫結構

本系統資料庫採用 MySQL database，以 SNP rs_fasta 資料庫為主，含 Human(ftp://ftp.ncbi.nih.gov/snp/human/)、Mouse(ftp://ftp.ncbi.nih.gov/snp/mouse/)、及 Rat(ftp://ftp.ncbi.nih.gov/snp/rat/)的 fasta sequence 資料庫。

2.3 演算法

未知序列與資料庫中的 rs_fasta 序列資料做比對，為了判別輸入的序列是否包含 variation 的鹼基，因此必須搜尋 fasta 的序列資料，然而 fasta sequence database 很大，因此本系統採用平均效率較快的 Boyer-Moore algorithm[7]以提高搜尋比對的效率。Boyer-Moore algorithm 乃利用自右而左的處理比對，與一般習慣性從左往右的處理方式不太相同，但其平均搜尋效率比 Knuth-Morris-Pratt algorithm[7]演算法好，顯然比 Brute Force algorithm[7]的效率高。以下簡述三種方法之差異:

(1) Brute Force algorithm 主要從左到右逐一比對，如果比對錯誤則欲比對的字串則往右邊移動一個位置，不包含預先處理的階段，時間複雜度為 O(mn)。

(2) Knuth-Morris-Pratt algorithm 從左到右開始比對，預先處理的階段需花費 O(m)的空間和時間複雜度，搜尋階段時間複雜度為 O(m+n)。

(3) Boyer-Moore algorithms 從右到左開始比對，預先處理的階段需花費 O(m+σ)的空間和時間複雜度，σ 表示 bad-character shift function 被儲存在 table 的大小，最好的執行效率為 O(n/m)時間複雜度。

本系統所建置的 SNP fasta sequence 資料庫中，fasta 資料有許多欄位，在序列的處理上，只需用到 alleles、sequence 3、variation 及 sequence 5 等四個欄位。輸入序列與 SNP fasta 序列的比對方法如下:

首先，在第一次處理時，先判別欲比對的輸入序列是否包含有 variation，即 M、R、W、S、Y、K、V、H、D、B 和 N，若有則做以下兩個步驟處理:

步驟 1. 比對 fasta 資料庫中的 variation 與輸入序列中所找到的 variation，如果比對正確則做步驟 2，否則代表此輸入序列中的 variation 不在此 fasta 上，繼續往下一筆 fasta 序列比對，直到比對至資料庫最後一筆才完全結束，如圖 3 所示。

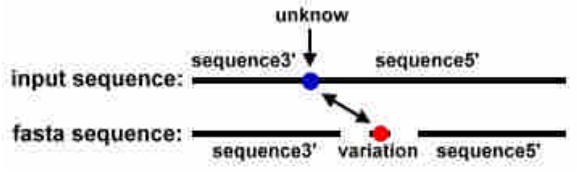


圖 3. variation 比對

步驟 2. 比對輸入序列的 sequence 3’與 fasta 序列的 sequence 3’和輸入序列的 sequence 5’ 與 fasta 序列的 sequence 5’，如果比對成功代表此輸入序列為 hit 到 fasta 的 SNP 序列，如圖 4 所示。

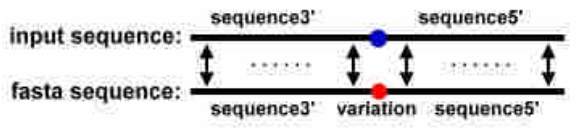


圖 4. variation flanking 比對

如果欲比對的輸入序列中不包含 variation，則做下列步驟:

步驟 1. 未知輸入序列是否為 SNP 序列，因此把輸入序列的每一個鹼基都當成 SNP 的點，接著與 fasta 資料庫中的 alleles 比對，如果比對失敗則往下一個鹼基移動，否則做步驟 3，如圖 5 所示。

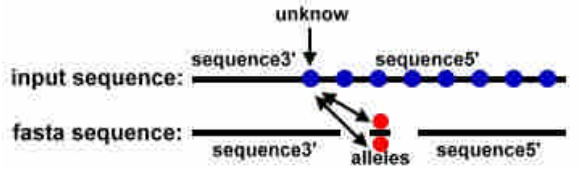


圖 5. alleles 比對

步驟 2. 當輸入序列的鹼基都比對完，若都沒有比對到，則回到步驟 1.往下一筆 fasta 序列繼續比對。

步驟 3. 如果輸入序列的鹼基比對到 fasta 序列的 alleles 時，則比對輸入序列的 sequence 3’與 fasta 序列的 sequence 3’和輸入序列的 sequence 5’ 與 fasta 序列的 sequence 5’，如果比對成功代表此輸入序列為 hit 到 fasta 的 SNP 序列，如圖 6 所示。

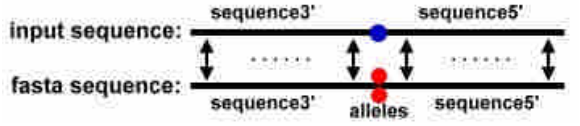


圖 6. alleles flanking 比對

以上為 SNP 序列的比對方法，接著說明 flanking 序列的比對方法，分成兩個步驟，如下所述:

步驟 1. 把輸入序列當成 sequence 3’，然後與 fasta 資料庫中的 sequence 3 做比對，如果比對成功代表此輸入序列為 SNP 序列的 flanking 序列，否則做步驟 2.，如圖 7 所示。

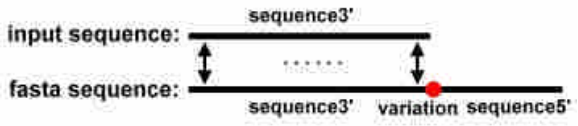


圖 7. sequence 3’ flanking 比對

步驟 2. 把輸入的序列當成 sequence 5’，然後與 fasta 資料庫中的 sequence 5 做比對，如果比對成功代表此輸入序列為 SNP 序列的 flanking 序列，反之則否，如圖 8 所示。

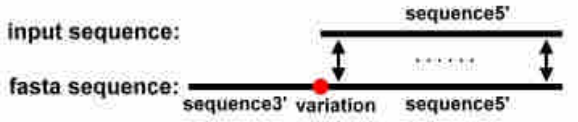


圖 8. sequence 5’ flanking 比對

2.5 視覺化方法

在視覺化部分，將生物學家輸入的序列及比對到的 SNP 及 flanking，做成視覺化的圖形輸出，如圖 9 所示。

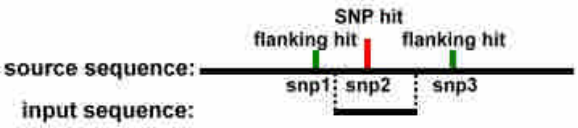


圖 9. 視覺化搜尋到的 SNP 及輸入序列

在圖 9 中所呈現的位置都是相對的，snp1、snp2 及 snp3 都是經由輸入序列所找到的 SNP，只不過其所代表的意義不同，輸入序列所代表的意義是 snp1 fasta 序列中的 sequence 5’的子序列，snp2 fasta 序列所包含到 allele 的子序列以及 snp3 fasta 序列中的 sequence 3’的子序列。

進行視覺化時，必須找出 source sequence，source sequence 可以從此物種的全部 DNA 序列中搜尋到，但要載入全部 DNA 序列去做搜尋的動作，系統肯定無法負荷，因此本系統利用 SNP fasta 序列重疊的特性找出 source sequence，如圖 10 所示。

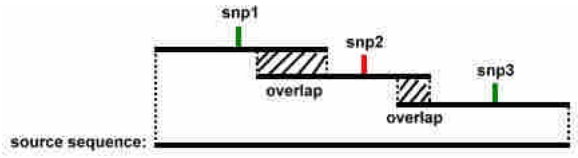


圖 10. source sequence 為 SNP fasta 序列重疊的部分

三、結果與討論

測試下列 SNP ID rs17806770 [Homo sapiens] 的四個序列:

序列 1.

GTGGACCGAAATCCCGCGACAGCAA

[A/G]AGGCCCGTAGCGACCCGCGGTGCTA

序列 2.

GTGGACCGAAATCCCGCGACAGCAA**R**AGGCC CGTAGCGACCCGCGGTGCTA

序列 3.

GTGGACCGAAATCCCGCGACAGCAA**A**AGGCC CGTAGCGACCCGCGGTGCTA

序列 4.

GTGGACCGAAATCCCGCGACAGCAA**G**AGGC CCGTAGCGACCCGCGGTGCTA

加底線的部分代表原本輸入的 SNP ID 和核苷酸，使用程式 blastn 沒有用 megablast，這些序列被 Blast 的結果不同，“score”= 101~50 和“expect”= 9e-20~7e-07。11 個 SNP IDs 被 Blast 如下: rs17883670, rs17883184, rs17882910, rs17880578, rs1421314, rs17883398, rs17551157, rs17551150, rs17806770, rs17880282, 和 rs17881686，然而 rs17806770 卻不是得到高的分數，因此無法得知輸入的序列正確 SNP ID，如圖 11 所示。

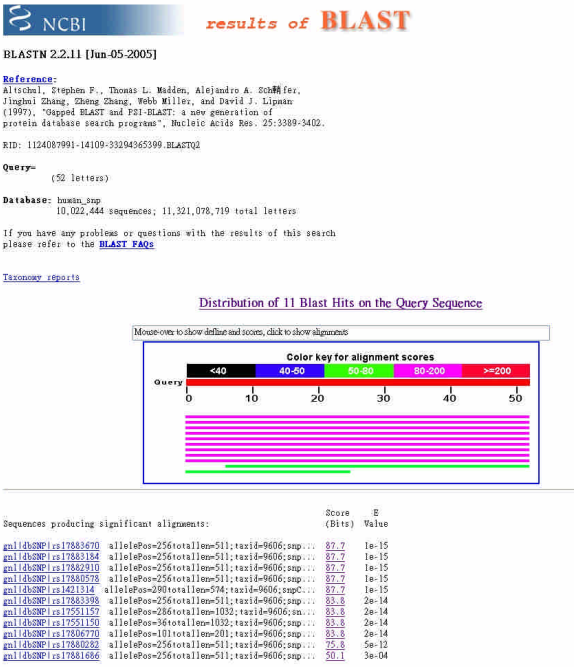


圖 11. 眾多的 SNP IDs，rs17806770 不是最高的分數，無法得到正確的 SNP ID

使用程式 blastn 有用 megablast，這些序列對於 rs17806770 呈現不同的結果，當輸入序列為序列 1 ，SNP-BLAST 的結果是 rs17882910，score = 93.0 和 expect = 3e-17，如圖 12。

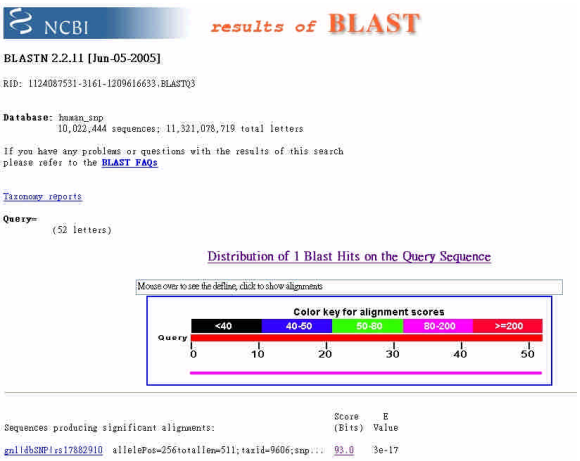


圖 12. rs17806770 序列 1 做 blastn with megablast 沒有 match 到 rs17806770

當輸入序列為序列 2 ，SNP-BLAST 的結果是 no significant similarity was found，也就是沒有提供 SNP IDs，如圖 13。

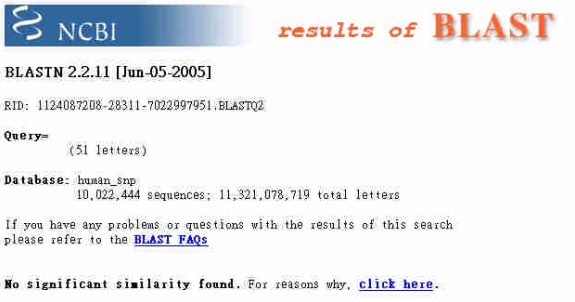


圖 13. rs17806770 序列 2 沒有提供 SNP IDs

當輸入序列為序列 3，SNP-BLAST 的結果是 rs17883398, rs17551157 和 rs17806770 且 score = 97 和 expect = 2e-18，但是三個 SNP IDs 當中，score 都相同，無法判定輸入序列為 rs17806770，如圖 14 所示。

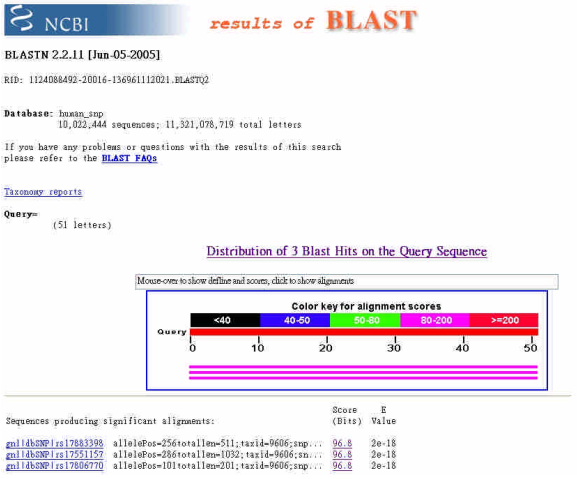


圖 14. 三個 SNP IDs 當中，score 都相同，無法判定輸入序列為 rs17806770

當輸入序列為序列 4，8 個 SNP IDs 被 Blasted：rs17883670, rs17883184, rs17882910, rs17880578, rs1421314, rs17551150, rs17880282 和 rs17881686，但是沒有一個比對到 rs17806770，如圖 15 所示。

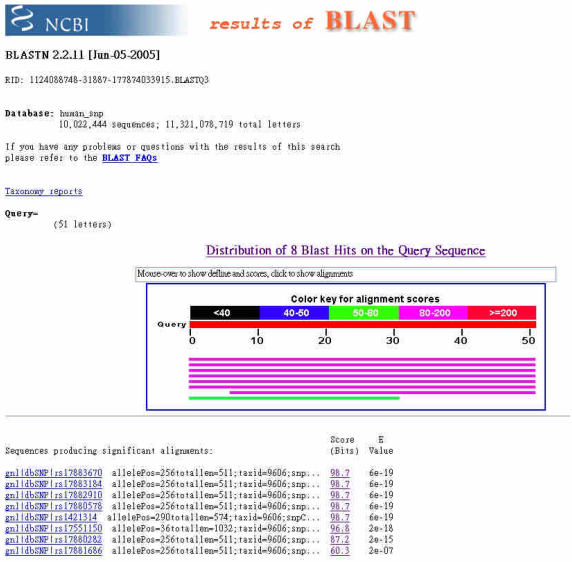
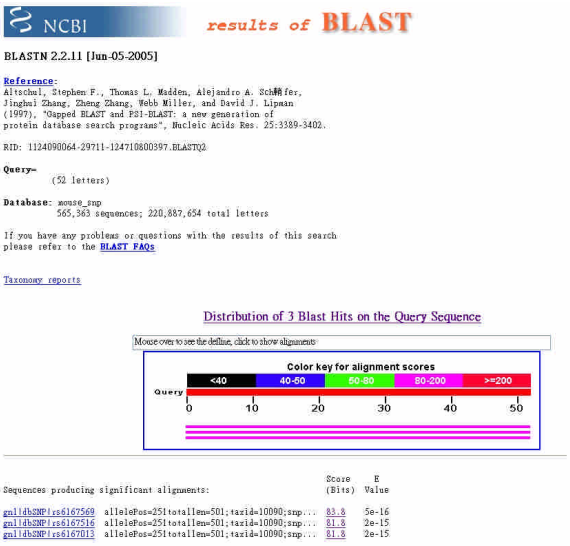


圖 15. 8 個 SNP IDs 沒有一個 match 到 rs17806770

同樣的，SNP-BLAST 在其他的物種如 Mus musculus 也有相同的問題，例如：rs6167569 的序列為 TCTTGCGTAGATCC GTCACAGCCCT**[C/T]**TTTCACCCGCCAGGGCT CCCGACAA，使用 blastn 沒有 megablast，rs6167569 的序列種類包含**[C/T]**, **Y**, **C** 和 **T**，都 Blast 到三個 SNP IDs: rs6167569, rs6167516, 和 rs6167013，如圖 16，雖然這次 rs6167569 獲得了最高分，但是使用 blastn 有 megablast， rs6167569 包含有 **Y** 的序列沒有 blast 到任何的 SNP ID，如圖 17，rs6167569 包含有**[C/T]**和 **C** 的序列 blast 到 rs6167516，跟原本輸入的 rs6167569 不一樣，score 分別為 89.1 和 94.9 且 expect 分別為 1e-19 和 2e-19，如圖 18、圖 19 所示，而 rs6167569 包含 **T** 卻正確的 match 到 rs6167569，如圖 20 所示。


NCBI results of BLAST

BLASTN 2.2.11 [Jun-05-2005]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1124090064-29711-124710800397.BLASTQ2

Query=
(52 letters)

Database: mouse_snp
565,363 sequences; 220,887,654 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQs

Taxonomy reports

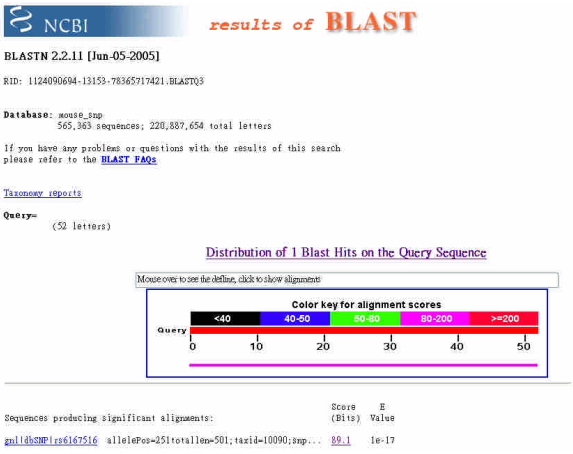
Distribution of 3 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Color key for alignment scores
<40 | 40-50 | 50-80 | 80-200 | >=200
Query
0 10 20 30 40 50

Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs6167569	allelePos=251totallen=501;taxid=10090;snp...	83.8	5e-16
gnl|dbSNP|rs6167516	allelePos=251totallen=501;taxid=10090;snp...	81.8	2e-15
gnl|dbSNP|rs6167013	allelePos=251totallen=501;taxid=10090;snp...	81.8	2e-15

圖 16. rs6167569 序列種類包含**[C/T]**, **Y**, **C** 和 **T**，rs6167569 獲得了最高分


NCBI results of BLAST

BLASTN 2.2.11 [Jun-05-2005]

RID: 1124090694-13153-78365717421.BLASTQ3

Database: mouse_snp
565,363 sequences; 220,887,654 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQs

Taxonomy reports

Query=
(52 letters)

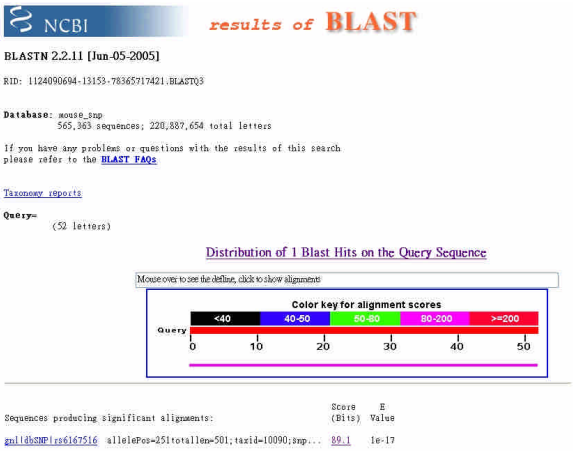
Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Color key for alignment scores
<40 | 40-50 | 50-80 | 80-200 | >=200
Query
0 10 20 30 40 50

Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs6167516	allelePos=251totallen=501;taxid=10090;snp...	89.1	1e-17

圖 17. rs6167569 包含 **Y** 的序列沒有提供 SNP IDs


NCBI results of BLAST

BLASTN 2.2.11 [Jun-05-2005]

RID: 1124090694-13153-78365717421.BLASTQ3

Database: mouse_snp
565,363 sequences; 220,887,654 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQs

Taxonomy reports

Query=
(52 letters)

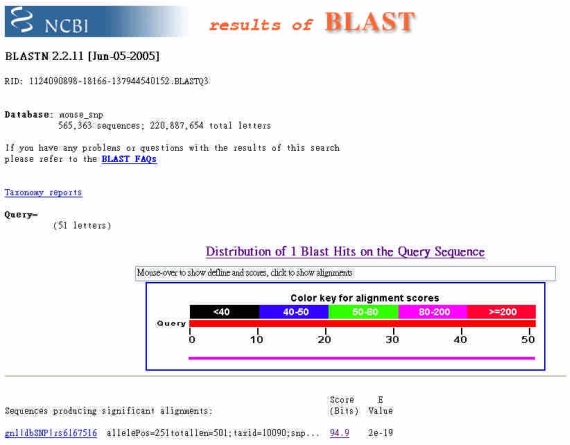
Distribution of 1 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Color key for alignment scores
<40 | 40-50 | 50-80 | 80-200 | >=200
Query
0 10 20 30 40 50

Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs6167516	allelePos=251totallen=501;taxid=10090;snp...	89.1	1e-17

圖 18. rs6167569 包含**[C/T]**的序列沒有 match 到 rs6167569


NCBI results of BLAST

BLASTN 2.2.11 [Jun-05-2005]

RID: 1124090898-18166-137944540152.BLASTQ3

Database: mouse_snp
565,363 sequences; 220,887,654 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQs

Taxonomy reports

Query=
(51 letters)

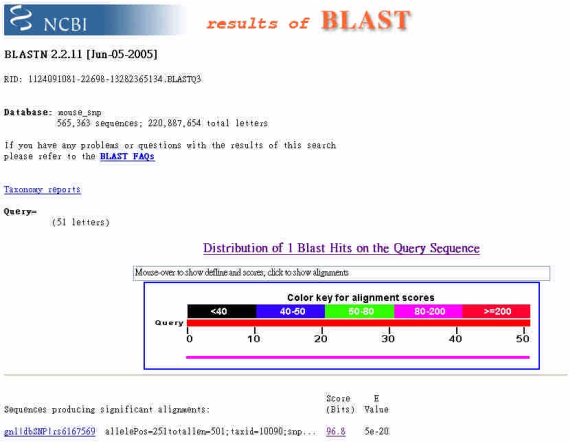
Distribution of 1 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Color key for alignment scores
<40 | 40-50 | 50-80 | 80-200 | >=200
Query
0 10 20 30 40 50

Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs6167516	allelePos=251totallen=501;taxid=10090;snp...	94.9	2e-19

圖 19. rs6167569 包含 **C** 的序列沒有 match 到 rs6167569


NCBI results of BLAST

BLASTN 2.2.11 [Jun-05-2005]

RID: 1124091081-22698-13282365134.BLASTQ3

Database: mouse_snp
565,363 sequences; 220,887,654 total letters

If you have any problems or questions with the results of this search please refer to the BLAST FAQs

Taxonomy reports

Query=
(51 letters)

Distribution of 1 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Color key for alignment scores
<40 | 40-50 | 50-80 | 80-200 | >=200
Query
0 10 20 30 40 50

Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs6167569	allelePos=251totallen=501;taxid=10090;snp...	96.8	5e-20

圖 20. rs6167569 包含 **T** 的序列 match 到 rs6167569

使用本系統鑑定 rs17806770 序列為 GTGGACCGAAATCCCGCGACAGCAA **[A/G]**AGGCCCGTAGCGACCCGCGGTGCTA 包含 **R**, **A**, **G** 序列，四者的序列輸出顯示結果如圖 21 所示，皆正確 match 到此序列有 SNP rs17806770 存在，當然鑑定 rs6167569 的四個序列，也都正確 match 到 rs6167569，如圖 22 所示。

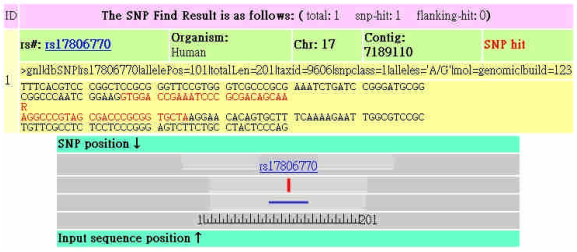


圖 21. 正確 match 到 rs17806770

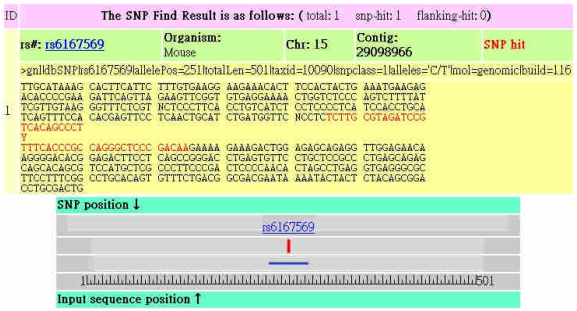


圖 22. 正確 match 到 rs6167569

在結果呈現方面，本系統依照 SNP 的 contig position 做順序性的輸出，並以相對位置視覺化方式呈現，讓生物學家清楚的明白輸入序列的位置及 SNP 的位置，方便做進一步研究，如圖 23。

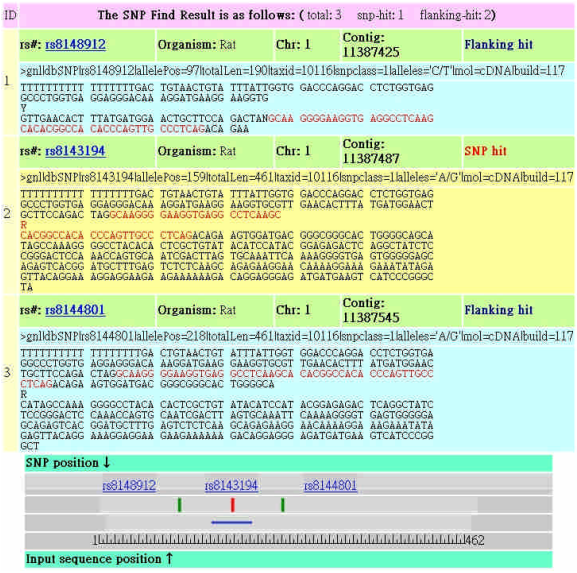


圖 23. contig position 順序性的輸出，並以相對位置視覺化方式呈現

本資訊平台能藉由輸入序列提供精確的 SNP ID，並依照 contig position 做順序性及視覺化呈現，解決 NCBI Blast 無法精確提供 SNP ID 的問題，然而 NCBI Blast 對於大資料量的比對效率仍是本系統所無法望其項背的，表 1 為本 SNP 資訊平台與 NCBI Blast 的比較表。

表 1. 本系統與 NCBI Blast 比較表

系統 / 功能	本系統	NCBI Blast
可比對的 organism	Human, Mouse 和 Rat	NCBI 資料庫有的，均可比對
SNP ID 精確性	包含 SNP hit, 和 SNP flanking, 精確性高	以 score 及 except 為評分標準，精確性不高
視覺化	輸入序列、SNP ID 位置、hit color	僅利用 color 顯示比對分數

四、結論

NCBI Blast 是目前最常用且功能強大的序列比對工具，承續了 Heuristic 演算法的觀念，以較小的時間複雜度找尋序列中相似的片段，然而還是有其缺點存在，如在 blastn program 預設 Use Megablast 無法正確比對到所輸入的序列。本研究提出一個 web-based SNP 資訊平台能從輸入序列中提供 SNP ID，並能夠藉由 reference cluster ID “rs#”、 NCBI assay ID “ss#”、gene name 和 gene ID，提供快速搜尋 SNP fasta 資訊，解決 NCBI SNP Blast 不能提供精確的 SNP ID 的問題，藉此輔助生物學家更方便的進行 SNP 的相關研究。

五、參考文獻

[1] Riva, A. and I.S. Kohane. 2002. SNPper: retrieval and analysis of human SNPs. Bioinformatics 18:1681-1685.

[2] Wang, L., S. Liu, T. Niu and X. Xu. 2005. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. BMC Bioinformatics 6:60.

[3] Savage, D., J. Batley, T. Erwin, E. Logan, C.G. Love, G.A. Lim, E. Mongin, G. Barker, et al. 2005. SNPServer: a real-time SNP discovery tool. Nucleic Acids Res 33:W493-495.

[4] National Center for Biotechnology Information BLAST, http://www.ncbi.nlm.nih.gov/BLAST/.

[5] National Center for Biotechnology Information, Single Nucleotide Polymorphism, dbSNP

build124, http://www.ncbi.nlm.nih.gov/projects/SNP/.

[6] National Center for Biotechnology Information Blast SNP, http://www.ncbi.nlm.nih.gov/projects/SNP/snp_blastByOrg.cgi.

[7] Christian Charras et Thierry Lecroq, Handbook of Exact String Matching Algorithms, King's College London Publications, 2004.