SWAT: A New Spliced Alignment Tool Tailored for Handling More Sequencing Errors

Yifeng Li[1] and Hesham H. Ali[2]

[1] Department of Pathology and Microbiology,
University of Nebraska Medical Center, Omaha, NE 68198-6805
yl1@unmc.edu
[2] Department of Computer Science, College of Information Science and Technology,
University of Nebraska at Omaha, Omaha, NE 68182-0116
hali@mail.unomaha.edu

Abstract. There are several computer programs that align mRNA with its genomic counterpart to determine exon boundaries. Though most of these programs perform such alignment efficiently and accurately, they can only tolerate a relatively small number of sequencing errors. These programs also highly depend on the GT/AG rule in finding splice sites. Both properties make them less desirable in the case of aligning EST reconstructed transcript with genomic DNA to identify splicing variants, where a lot of sequencing errors and non-canonical splice sites are expected. Using a novel heuristic algorithm, we developed a tool that can handle much more sequencing errors. Test dataset results indicated that SWAT (Sequencing-error Well-handled Alignment Tool) has a much stronger error-handling ability than Sim4 and Spidey, two other popular spliced alignment tools. In the presence of up to 10 percent randomly introduced sequencing errors, it can still give the precise number of exons and exon boundaries in most cases. The robustness of SWAT makes it a desirable tool in cases where sequencing error is a concern. A web service is freely available at http://app1.unmc.edu/swat/swat.html.

1 Introduction

This work is motivated by recent studies on splicing variants using EST data [1, 2]. Alternative splicing is known as the most important mechanism to increase protein diversity. It is estimated that 40–60% of human genes undergo such event and altered mRNA splicing can have a dramatic effect on the structure of the encoded protein [3-5]. Due to its prevalence and biological importance, extensive studies on alternative splicing have been carried out. To identify different splice forms, it is essential to perform an alignment of the transcribed mRNA sequences with the corresponding genomic DNA. ESTs (expressed sequenced tags), which are derived from processed mRNA, in public database provide an ideal resource for transcripts that can be used in such alignment [6]. The major advantage of using ESTs is their abundance, which makes the database likely contain all differently spliced transcripts of the same gene. A disadvantage is their low quality. ESTs often contain errors at a rate much higher

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 927 – 935, 2005.

than those of the finished or even draft genomic sequence [6, 7]. Though most existing tools perform the alignment efficiently and accurately, they usually only allow a relatively small number of sequencing errors. Therefore, to efficiently use ESTs, there is a call for more robust alignment tool. SWAT, which uses a novel heuristic algorithm, achieves fast alignment and efficient error-handling at the same time.

2 Previous Work

There are several programs that perform mRNA-to-genomic alignment. Among them, est_genome [8], Sim4 [9] and Spidey [10] are the three most widely known, whereas MGAlign [11] is the most recently developed one. The following is a brief review of the above-mentioned programs.

Est_genome uses a liner-space dynamic programming recurrence to align spliced sequences to their genomic counterparts. A modified Smith-Waterman scan is performed to locate the maximum-scoring segments. The genome sequence is then searched against forward and reverse strands of the spliced sequence, assuming the splice consensus GT/AG. Est_genome has a strong error-handling ability. However, the long running time limits its use.

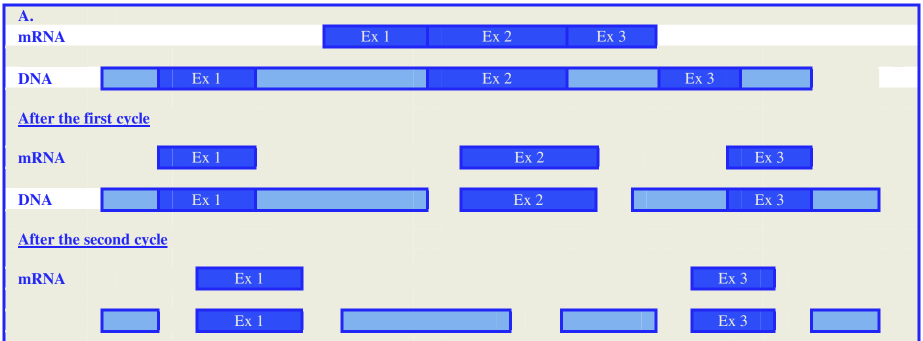
Sim4 uses a heuristic algorithm. It first determines all the high-scoring segment pairs (HSPs) such as those computed by the BLAST program [12]. It then selects a best chain of the HSPs subject to certain constraints and finally applies the GT/AG rule to find exon boundaries. The program only expects a small number of sequencing errors. BLAST-based software suffers from the high granularity of the BLAST program and the problem is worsened when there are sequencing errors.

Spidey is a heuristic also. It first uses the BLAST to align mRNA and the genomic sequence. The BLAST alignments are then sorted by score and assigned into windows by a recursive function which takes the first alignment and then goes down the list to find all alignments that are consistent with the first. Once the genomic windows are constructed, another BLAST search is performed to align entire mRNA with each window at a lower stringency. Spidey then uses a greedy algorithm to generate a high scoring, non-overlapping subset of the alignments from the second BLAST search. Finally, alignments are truncated or extended as necessary so that they terminate at the splice donor site and do not overlap. Due to its BLAST-based manner, Spidey is not expected to handle sequencing errors efficiently either.

MGAlign is a newly developed tool which uses a rapid heuristic method. Its authors claim that it is more accurate and faster than Sim4 and Spidey [13]. However, the details of the algorithm have never been published. We found that this tool does have an improved ability in handling small exons but it is even more sensitive to sequencing errors than Sim4 and Spidey.

3 Proposed Approach

The original problem to be solved is to determine the exon-intron structure for a given pair of mRNA and its parent genomic DNA. SWAT uses a divided and conquer algorithm: suppose an mRNA segment is found matching a DNA segment and such a match can not be extended further in either direction, then by removing the matched segments in both sequences, the original problem can be turned down into two identical but smaller ones (Figure 1). The segment removed is an exon whose boundary has been determined. The smaller cases resulted from the splitting can be treated exactly in the same way: finding the match and then splitting. This strategy can be used repeatedly until there is no mRNA segment left, at which point the original problem is solved. Therefore, the problem of determining exon boundaries can be solved recursively. After each cycle, it always results in an exon with solved boundaries and one or two identical but miniature version of the original problem. The base case for the recursive function will be as the following: given an mRNA and its parent genomic DNA, from a randomly selected point in the mRNA, finding the corresponding exon in both sequences that covers that point.



B.

```
DIVIDE (string RS, string DS){
    Select substring S (S ∊ RS and length (S) = 15)
    Find substring S' ∊ DS that satisfies S' ≡ S
    Extend S and S' at both directions to reach the maximum score
    Retain the boundaries for S and S'
    $RS_L$ = portion of RS left to S
    $DS_L$ = portion of DS left to S'
    $RS_R$ = portion of RS right to S
    $DS_R$ = portion of DS right to S'
    If (length ($RS_L$) > 0)
        DIVIDE (string $RS_L$, string $DS_L$)
    If (length ($RS_R$) > 0)
        DIVIDE (string $RS_R$, string $DS_R$)
```

Fig. 1. Essence of the proposed divided and conquer algorithm. A. Schematic diagram of the key concept. B. Pseudo code for the recursive function call

In each cycle, the starting point in the mRNA is randomly chosen. Given the same pair of sequences the program may go a different path each time it runs, yet reaches the same result. Such randomness is essential in getting the best result when there are errors in the sequence. Current settings guarantee that wrong match will not happen when there are no or very little sequencing errors. However, the chance of having a

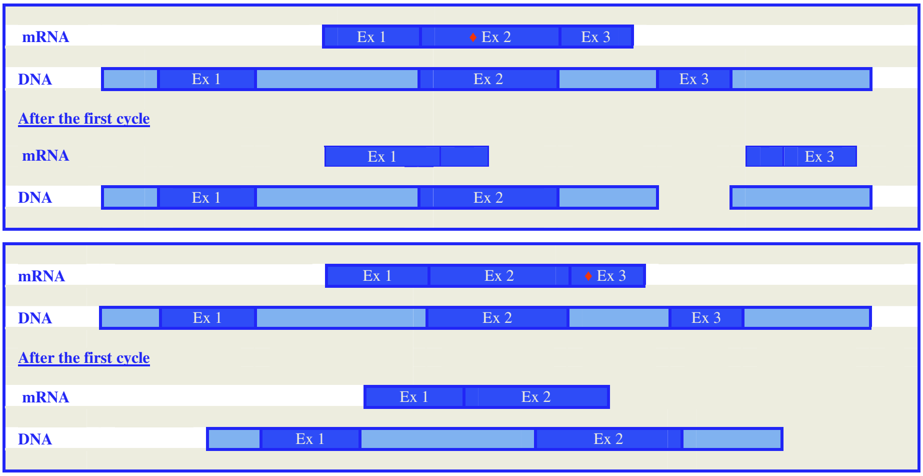


Fig. 2. Graphic illustration of the importance of random start. Top: Wrong match happens between part of Ex2 in the mRNA and a DNA region containing Ex3. This wrong match and subsequent splitting cause problem for finding matches in the next cycle. Bottom: Wrong match is remedied by selecting a different starting point in the mRNA. ♦: the starting point

wrong match enhances with increased sequencing errors. Mismatch dues to sequencing errors in a particular run cannot be avoided but the failure it causes can be remedied by running the program several times more. Suppose a wrong match happens in the middle of a run and results in a bad split which in turn causes problems for finding matches in the subsequent cycles, the overall similarity score would be much lower than what is expected. Since the starting point is randomly chosen, another run may go a different path and therefore avoid the bad match (Figure 2). Best result can always be reached by choosing the highest similarity score from 2 to 3 runs. The strategy of random start makes a contribution to the stronger error-handling ability of this tool.

4 Algorithm Description

The proposed algorithm is implemented using C++ programming language and the major steps are described below.

1. *Find the matched segments*. Randomly select an mRNA segment with 10-15 residues in length and find its exact match in the DNA. Extend the matches at both directions as long as exact match can still be found. When exact match is no longer available at either end, extend the current matches to 8 and 12 residues for mRNA and DNA, respectively. For all combinations of length in both extended segments, use the Needleman-Wunsch algorithm [14] to find the segment pair that gives the highest similarity score. This high-score pair will be used as the actual extension. Repeat this if each time it lets the matches grow more than one residue. Otherwise, stop there and no further attempt in extending the match is made. This extension strategy allows

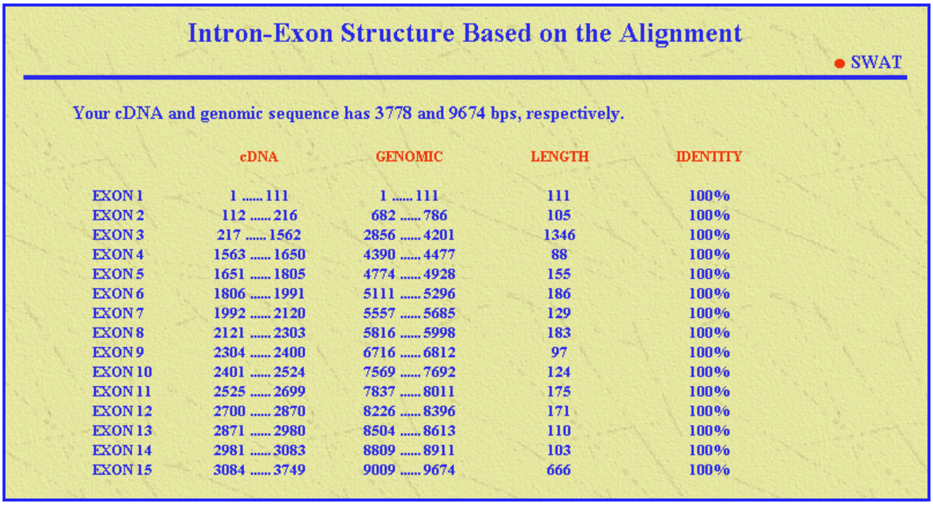
certain number of mismatches, insertion or deletion to be tolerated in determining the final matched segments.

2. *Split both sequences and call the recursive function*. The matched segments detected in the previous step are removed from both sequences and the corresponding position of the cleavage sites are recorded for future exon-intron structure construction. By doing so, both sequences are split into a left portion and a right portion. Each portion of the mRNA and DNA pair forms an identical but miniature version of the original problem, which is solved through a recursive function call. The program steps out of the recursive function when there is no mRNA segment left.

3. *Trim the exon boundaries*. After determining all the candidate exons, GT/AG rule is applied to trim the exon boundaries.

5 Alignment Validation

SWAT was tested with the new multi_exon entries of the GENIE gene finding data set[♠]. There are totally 137 records in this data set. However, matches cannot be found for 5 of them in the GenBank records, so we actually tested 132 cases, of which 64 have 2-5 exons, 38 have 6-10 exons, 19 have 11-15 exons, 5 have 16-20 exons, 2 have 21-25 exons, 2 have 26-30 exons, 1 has 33 exons and 1 has 51 exons. In this study, GenBank records of each mRNA are aligned with the corresponding genomic DNA using our program and the results are compared with the NCBI annotation. For all 132 entries, results given by SWAT are consistent with the NCBI annotation. Like



Intron-Exon Structure Based on the Alignment

● SWAT

Your cDNA and genomic sequence has 3778 and 9674 bps, respectively.

	cDNA	GENOMIC	LENGTH	IDENTITY
EXON 1	1 111	1 111	111	100%
EXON 2	112 216	682 786	105	100%
EXON 3	217 1562	2856 4201	1346	100%
EXON 4	1563 1650	4390 4477	88	100%
EXON 5	1651 1805	4774 4928	155	100%
EXON 6	1806 1991	5111 5296	186	100%
EXON 7	1992 2120	5557 5685	129	100%
EXON 8	2121 2303	5816 5998	183	100%
EXON 9	2304 2400	6716 6812	97	100%
EXON 10	2401 2524	7569 7692	124	100%
EXON 11	2525 2699	7837 8011	175	100%
EXON 12	2700 2870	8226 8396	171	100%
EXON 13	2871 2980	8504 8613	110	100%
EXON 14	2981 3083	8809 8911	103	100%
EXON 15	3084 3749	9009 9674	666	100%

Fig. 3. Sample output of SWAT showing intron-exon structure based on the alignment

[♠] The website for the data set is at http://www.fruitfly.org/seq_tools/datasets/Human.

MGAlign, SWAT also has a strong ability in detecting small exons, which tend to be missed by Sim4 and Spidey. Exon as small as 9 bps can be successfully detected. Two of such cases are included in the 'Alignment Examples' of the service website. In both cases, a small exon was missed by Sim4 and Spidey. A sample output is shown below (Figure 3). In this case, N-deacetylase/N-sulfotransferase 2 (NDST2) transcript is aligned with its corresponding genomic DNA.

6 Error-Handling Ability

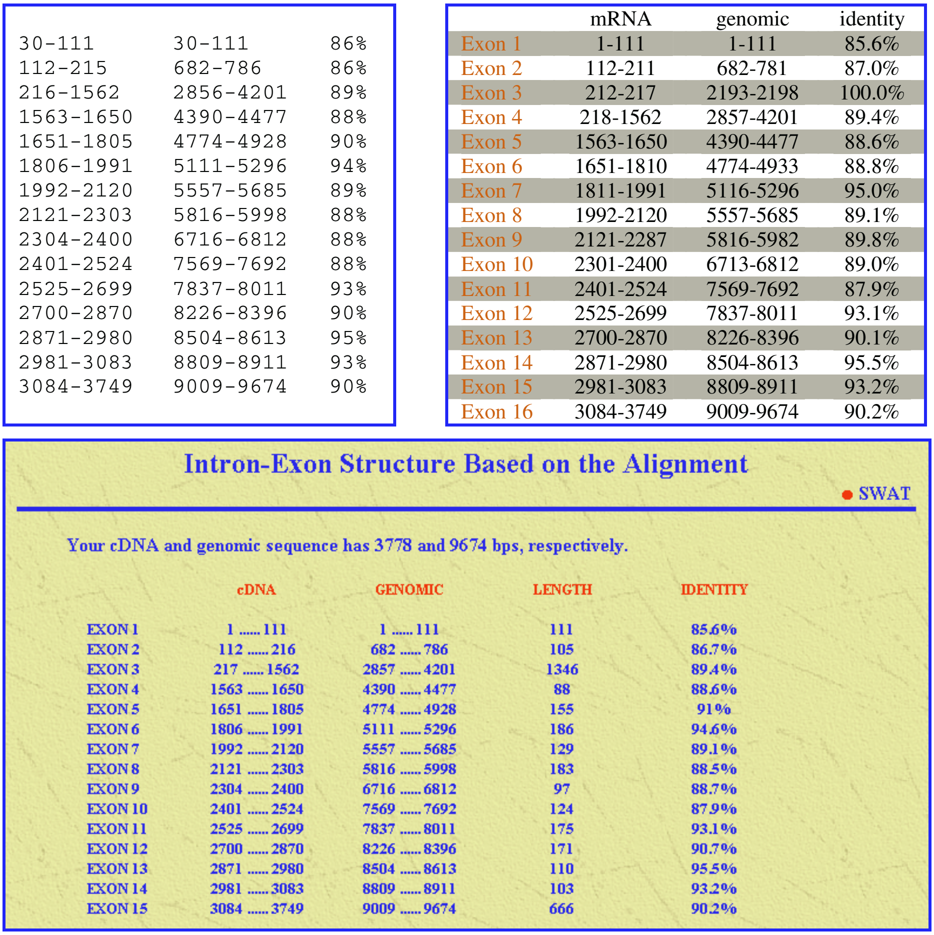
For certain cases in the test dataset, the mRNA segment and the corresponding exon do not perfectly match due to sequencing errors. But for all such cases, the small discrepancy can be properly handled by the program. Two of such cases are included in the 'Alignment Examples' of the service website. Since SWAT is designed to handle more sequencing errors, its actual error-handling ability has been further tested with error-containing mRNA and the results are compared with those given by Sim4 and Spidey . For each of the 132 entries that are used previously for alignment validation, 10 percent sequencing errors were randomly introduced into the mRNA sequence. These error-containing mRNA were aligned with error-free DNA sequence using the three programs. In all cases, SWAT correctly determined the number of exons and in only 9 cases there were small shift in the exon boundaries*. The rate of failure for Sim4 and Spidey are much higher (Table 1). These results clearly show that SWAT has a stronger error handling ability than the other two programs.

Table 1. Result of error-handling ability comparison among SWAT, Sim4 and Spidey. The numbers shown are the failed cases

	SWAT	Sim4	Spidey
Incorrect boundaries only	9	34	57
Incorrect numbers & boundaries	0	15	32
Overall rate of failure	6.8%	37.1%	67.4%

A particular case is shown below to give a better idea of how robust the program is (Figure 4). The same example in Figure 3 is used here for easy comparison but now the mRNA sequence contains 10 percent randomly introduced errors. In this case, SWAT gives identical result as there were no errors. Only the identity drops which indicates there are indeed errors in the sequence. Sim4 gives a big shift for the left boundary of exon 1 possibly due to the higher than average local error rate. Spidey detects an extra exon. Result given by MGAlign was even worse where the alignment ends up with a lot of gaps and dramatic boundary shifts (result not shown).

* Shift within 3 residues is not considered as a failure and this standard is applied for all three programs.



```
30-111        30-111        86%
112-215       682-786       86%
216-1562      2856-4201     89%
1563-1650     4390-4477     88%
1651-1805     4774-4928     90%
1806-1991     5111-5296     94%
1992-2120     5557-5685     89%
2121-2303     5816-5998     88%
2304-2400     6716-6812     88%
2401-2524     7569-7692     88%
2525-2699     7837-8011     93%
2700-2870     8226-8396     90%
2871-2980     8504-8613     95%
2981-3083     8809-8911     93%
3084-3749     9009-9674     90%
```

	mRNA	genomic	identity
Exon 1	1-111	1-111	85.6%
Exon 2	112-211	682-781	87.0%
Exon 3	212-217	2193-2198	100.0%
Exon 4	218-1562	2857-4201	89.4%
Exon 5	1563-1650	4390-4477	88.6%
Exon 6	1651-1810	4774-4933	88.8%
Exon 7	1811-1991	5116-5296	95.0%
Exon 8	1992-2120	5557-5685	89.1%
Exon 9	2121-2287	5816-5982	89.8%
Exon 10	2301-2400	6713-6812	89.0%
Exon 11	2401-2524	7569-7692	87.9%
Exon 12	2525-2699	7837-8011	93.1%
Exon 13	2700-2870	8226-8396	90.1%
Exon 14	2871-2980	8504-8613	95.5%
Exon 15	2981-3083	8809-8911	93.2%
Exon 16	3084-3749	9009-9674	90.2%

Intron-Exon Structure Based on the Alignment

● SWAT

Your cDNA and genomic sequence has 3778 and 9674 bps, respectively.

	cDNA	GENOMIC	LENGTH	IDENTITY
EXON 1	1 111	1 111	111	85.6%
EXON 2	112 216	682 786	105	86.7%
EXON 3	217 1562	2857 4201	1346	89.4%
EXON 4	1563 1650	4390 4477	88	88.6%
EXON 5	1651 1805	4774 4928	155	91%
EXON 6	1806 1991	5111 5296	186	94.6%
EXON 7	1992 2120	5557 5685	129	89.1%
EXON 8	2121 2303	5816 5998	183	88.5%
EXON 9	2304 2400	6716 6812	97	88.7%
EXON 10	2401 2524	7569 7692	124	87.9%
EXON 11	2525 2699	7837 8011	175	93.1%
EXON 12	2700 2870	8226 8396	171	90.7%
EXON 13	2871 2980	8504 8613	110	95.5%
EXON 14	2981 3083	8809 8911	103	93.2%
EXON 15	3084 3749	9009 9674	666	90.2%

Fig. 4. Error-handling ability comparison among Sim4, Spidey and SWAT on a test case. 10% sequencing errors are randomly introduced to the mRNA. Upper left: Result given by Sim4. Upper right: Result given by Spidey. Bottom: Result given by SWAT

7 Discussion

SWAT uses a divided and conquer algorithm. The original problem is kept turning down into identical but smaller ones and solved recursively. As indicated by the test dataset results, in most cases up to 10 percent randomly introduced sequencing errors can be properly handled. The strong error-handling ability first comes from the strategy used to extend local optimal match and then is further strengthened by the fact that best result can always be reached by choosing the highest similarity score from several runs, which is made possible by the divided-and-conquer algorithm itself and the randomness in selecting the starting point. Besides its ability in handling errors, SWAT has several other desirable properties as well. First, the program is extremely fast. The average running time per alignment is less than 0.5 second when performed on a 1.40 GHz Intel Pentium 4 system with 256 Mb of RAM. Second, non-canonical sites are likely to be detected. Though GT/AG rule is applied in the program, similarity is given the first priority and under no circumstances will it be sacrificed to satisfy the rule. Third, due to the manner of how the original sequence is broken down, all subsequent exons found are in consistent order. Therefore, consistency (mRNA and genomic DNA are non-overlapping and linearly consistent) is automatically achieved and no post-alignment allocation is needed. Forth, best result is guaranteed. The random way in selecting the starting point in each cycle makes it always possible to choose the best result from several runs. Both its strong error handling ability and high speed make SWAT an ideal tool for mRNA and genomic sequence alignment especially when sequencing error is a concern.

Acknowledgements

We thank Dr. Mark Pauley, Xutao Deng and Alexandre Tchourbanov for their helpful comments and suggestions. We thank Yan Wu for running the test dataset. We would also like to thank Joe Ziskovsky for helping us set up the web service.

References

1. Kan Z, Rouchka EC, Gish WR and States DJ. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res. (2001) 11, 889–900
2. Modrek B, Resch A, Grasso C and Lee C. Genome-wide analysis of alternative splicing using human expressed sequence data. Nucleic Acids Res. (2001) 29, 2850–2859
3. Graveley BR. Alternative splicing: Increasing diversity in the proteomic world. Trends Genet (2001) 17: 100–107
4. Black DL. Protein diversity from alternative splicing: A challenge for bio-informatics and post-genome biology. Cell (2000) 103: 367–370
5. Modrek B and Lee C. A genomic view of alternative splicing. Nat Genet (2002) 30: 13–19
6. Jongeneel CV. Searching the expressed sequence tag (EST) database: panning for genes. Brief Bioinform (2000) 1: 76-92
7. Boguski MS, Lowe TM and Tolstoshev CM. dbEST — database for "expressed sequence tags". Nat Genet (1993) 4, 332-333
8. Mott R. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. (1997) 13**:** 477–478.
9. Wheelan SJ, Church DM and Ostell JM. Spidey: a tool for mRNA-to-genomic alignments. Genome Res. (2001) 11: 1952-1957
10. Florea L, Hartzell G, Zhang Z, Rubin GM and Miller W. A computer program for aligning a cDNA sequence with a genomic sequence. Genome Res. (1998) 8: 967-74
11. Lee BT, Tan TW and Ranganathan S. MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences. Nucleic Acids Res. (2003) 31: 3533-3536

12. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic Local Alignment Search Tool. J. Mol. Biol. (1990) 215, 403-410
13. Ranganathan S, Lee BT and Tan TW. MGAlign, a reduced search space approach to the alignment of mRNA sequences to genomic sequences. Genome Informatics (2003) 14: 474-475
14. Needleman, SB. and Wunsch, CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. (1970) 48: 443-453