Jonathan Light
MB&B 452a

A Brief History of Sequence Alignment Methods

Sequence alignment is a key technique in genomics, and its importance has grown over the years as more robust and informative techniques have been developed. Sequence alignments are now very complex, but the concept came from simple origins. In the beginning, performing a sequence alignment involved an inspection of two sequences and consideration of how adding gaps could lead to a better alignment. Needless to say, this method was very time consuming and very imprecise, as there were no standardized criteria for what constitutes a better alignment. The first sequence alignment method that was developed that could be automated was the Needleman-Wunsch method (Needleman & Wunsch 1970). In this method, the originators realized that if you have two sequences, it is possible to align any member of one sequence to any member of the other sequence, thus the process lends itself to a matrix format. In the NW method the sequences are aligned as the rows and columns of a matrix, and first a similarity number is assigned to each possible match. In the earliest NW alignments, similarity was determined by whether or not the two elements being compared were identical or not. After the similarity matrix is constructed, a second matrix is constructed where the scores indicate the best possible alignment that terminates with those two elements aligning. Once this second matrix is complete, it is a simple matter to search for the highest score and draw a path that corresponds to the best alignment. This basic NW was later further refined by adding in penalties for creating and extending gaps, and including a more complex similarity scoring method in which conservative mismatches are given a higher similarity score than very non-conservative mismatches. There are various similarity-scoring methods that have been developed such as BLOSUM and PAM matrices that factor in the probability of a particular mutation over different evolutionary time scales. NW was revolutionary because it actually made alignments quick enough to be useful for reasonably sized sequences. Later, the basics of the NW alignment method were adapted to create a local alignment method known as the Smith-Waterman method (Smith & Waterman 1981). This alignment uses the same algorithm as NW but instead of just comparing the whole sequences it looks at alignments of each possible sub segment of the two sequences to try and find better scores. This is useful because it allows the identification of conserved motifs in two sequences that may be otherwise quite dissimilar.

a

	G	E	S	T	R	P	A	S
G	1	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0
E	0	1	0	0	0	0	0	0
T	0	0	0	1	0	0	0	0
R	0	0	0	0	1	0	0	0
P	0	0	0	0	0	1	0	0
P	0	0	0	0	0	1	0	0
S	0	0	1	0	0	0	0	1

b

	G	E	S	T	R	P	A	S
G								
W								
E								
T				**4**	2	1	1	0
R	2	2	2	2	3	1	1	0
P	2	2	2	2	2	2	1	0
P	1	1	1	1	1	2	1	0
S	0	0	1	0	0	0	0	1

c

	G	E	S	T	R	P	A	S
G	6	4	4	3	2	1	1	0
W	5	4	4	3	2	1	1	0
E	4	5	4	3	2	1	1	0
T	3	3	3	4	2	1	1	0
R	2	2	2	2	3	1	1	0
P	2	2	2	2	2	2	1	0
P	1	1	1	1	1	2	1	0
S	0	0	1	0	0	0	0	1

d

```
G W E - T R P P - S
G - E S T R - P A S
```

Figure 1: Example of NW Sequence Alignment.
a) Similarity matrix
b) Partially completed alignment
c) Finished alignment with best alignment path
d) Final Alignment

The next big innovation in sequence alignment came with the development of various methods of performing multiple sequence alignments. A vast number of different methods were developed for this, including

programs such as CLUSTAL, which performs alignments based on clustering the sequences, and various programs that use Hidden Markov Models in order to create sequence profiles. Today, the most commonly used sequence alignment program is Blast and various other programs derived from Blast. Blast is most noted for its impressive speed. The Blast algorithm works based on a principle of hashing small matching sequences and then extending the hash matches to create High Scoring Segment Pairs until you attain the highest possible score (Altschul, *et al* 1990).

There are many uses for sequence alignments in biology. One of the most common uses is identifying homology between two organisms as a means of demonstrating that they are evolutionarily related. Another common use is determining the function of a new protein by finding homologous proteins in other organisms.

References:

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J., *J. Mol. Biol.* (1990)

215: 403-410

Needleman, S. B., Wunsch, C. D., *J. Mol. Biol.* (1970) 48:443-453

Smith, T. F., Waterman, M. S., *J. Mol. Biol.* (1981) 147:195-197

Jonathan Light
MB&B 452a

Functional Genomics Analysis of the Genome of *M. gersteinius*

Having recently sequenced and identified all of the newly discovered archaea *M. gersteinius*, there are a number of different functional genomics analyses that will aid in the characterization of this unique and fascinating organism. Perhaps the first logical step would be to search for homologous genes in other organisms, especially archaea. A good way to go about this would be to run all of the ORFs of *M. gersteinius* through a BLAST protein search. Any proteins, especially in other archaea, that are found to have a high degree of sequence homology to one of the ORFs in *M. gersteinius* are likely to have a similar function. Thus if we get a homologous hit to a protein of known function, it is likely our ORF serves the same or another similar function in *M. gersteinius.*

Once we have identified likely functions for many of the proteins in *M. gersteinius*, I would say that the next likely step would be to perform a yeast two-hybrid screen of the proteins in *M. gersteinius*. To do this, you would need to create two chimeric constructs for each ORF, one in which the protein is fused to the DNA-binding domain of a transcription factor of a gene for ampicillin resistance, and one in which the protein is fused to the transcription activation domain of the ampicillin resistance gene. Because *M. gersteinius* is a small organism with only 700 ORFs, it is feasible to screen every pair. If the two proteins being screened interact, then the DNA binding domain and the transcriptional activation domain of the transcription factor for the ampicillin resistance will be in close enough proximity to allow transcription of the ampicillin resistance. Thus, any pair that grows on ampicillin is a possible protein-protein interaction. For those that you already have a likely function, this could possibly give useful information about the interacting protein. Although you get a lot of false positives with two-hybrid screening, it can still give useful information on interactions.

The next thing I would do would be a transposon screen. By transposing the genome of *M. gersteinius* with a transposon that contains a transcription stop codon, you can get insertions in genes that will inactivate them. By sequencing out both ends of the transposon, you can determine the location of the transposon and which gene it is in. This screen lets you determine which genes are required by *M. gersteinius* to live. For instance, if you never see an insertion in the gene that governs flipping through slides really fast, then it is likely that *M. gersteinius* can not survive without that gene.

So, with the above tests we have found likely functions for many of our genes, we have determined many possible pair wise protein interactions, and we have determined which genes are likely to be absolutely required by *M. gersteinius* in order to survive. I think the next step I would suggest would be to enter the wonderful world of microarrays. You could create an array that has complementary DNA to all the mRNAs of *M. gersteinius*. Then, you could grow samples of *M. gersteinius* under widely varying conditions, isolate their total mRNA, and run them over the microarray. By observing the levels of different mRNAs under different growth conditions, you could determine the effects of different growth conditions on the expression of different genes.

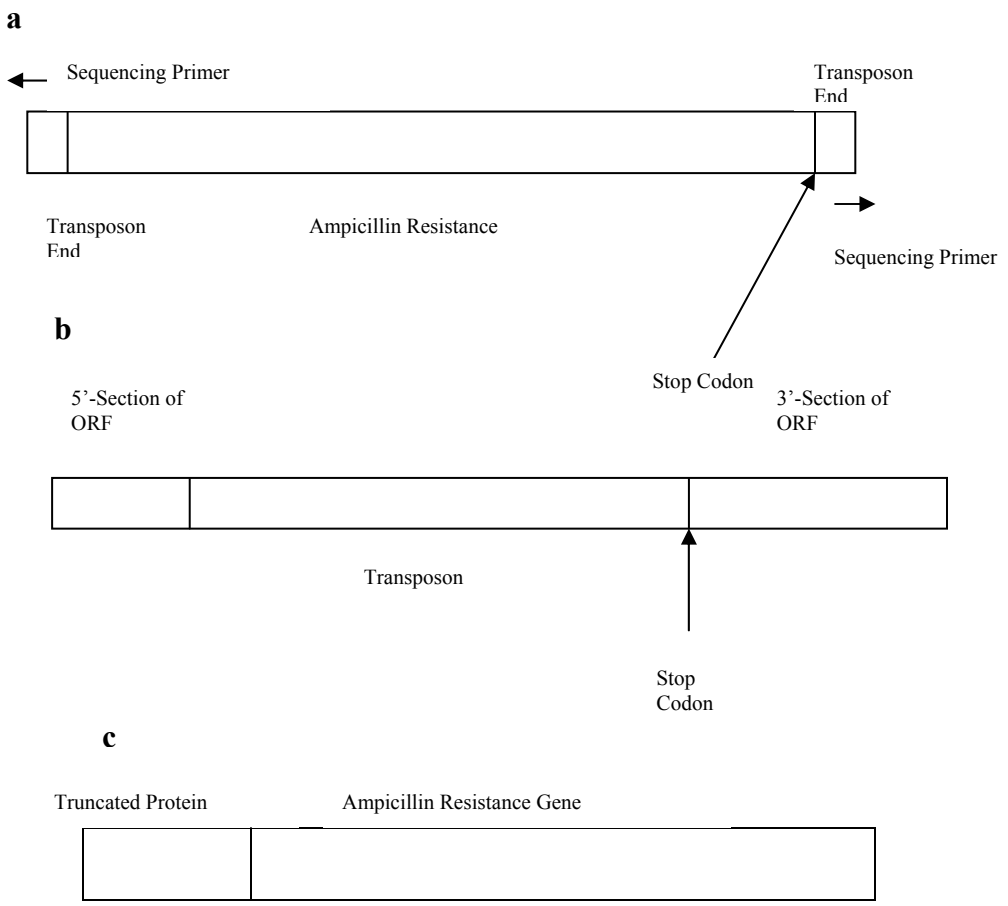

Figure 1: Transposon Screening of *M. gersteinius*: a) The salient features of the transposon are the ends which allow insertion of the transposon into the genome, the ampicillin resistance gene for screening purposes, the primer sequences to locate the transposon, and the stop codon to truncate the protein. b)The transposon inserts itself inside an ORF of *M. gersteinius.* c) The stop codon at the end of the transposon causes the gene product to be truncated. If the gene product is vital, the clone will not survive.