Pairwise sequence alignments & BLAST

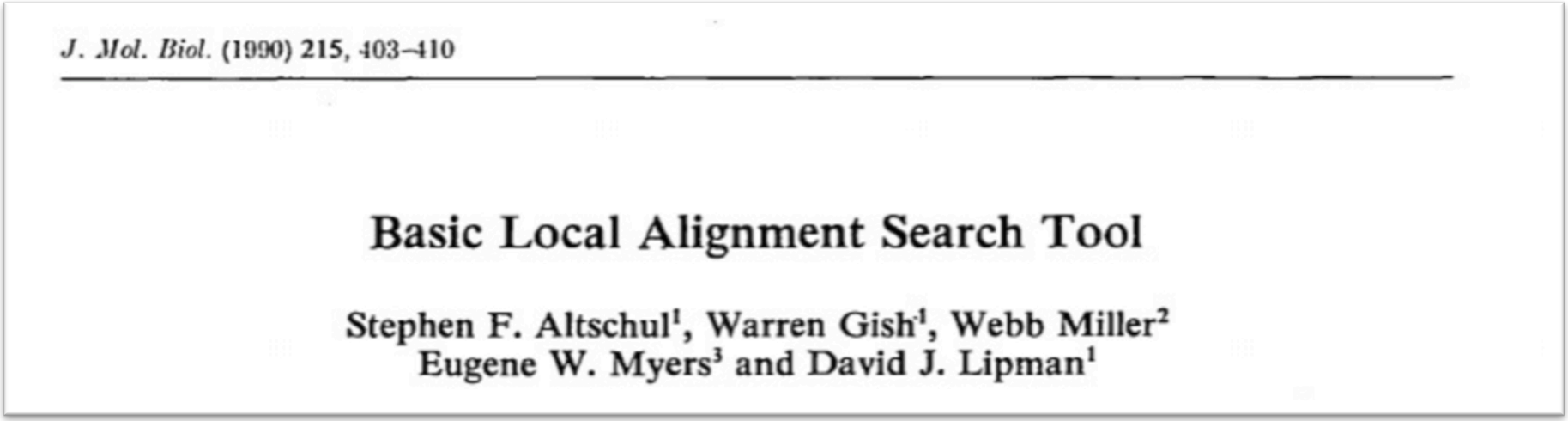
The point of sequence alignment

- If you have two or more sequences, you may want to know
  - How similar are they? (A quantitative measure)
  - Which residues correspond to each other?
  - Is there a pattern to the conservation/variability of the sequences?
  - What are the evolutionary relationships of these sequences?

```
Human: ccatcctcagatccgtcttcagaaccaccttcccctcgatccacggctccattttcatcc
       |||||||||||| |||||||||||| || |||||   ||| ||| |||| ||| | || |||
Mouse: ccatcctcagacccgtcttcagagcccccttc---tcggtccccggccccactgtcttcc
```

BLAST

- Basic Local Alignment Search Tool
- Altschul, et al 1990
- Has been cited over 61,000 times according to Google
- The most highly cited scientific paper in the entire decade of the 1990s



J. Mol. Biol. (1990) 215, 403–410

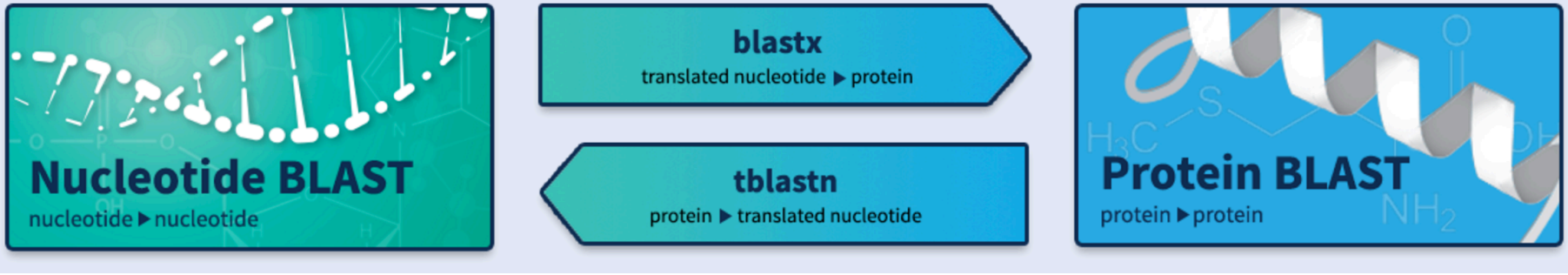
Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

BLAST

- Compares a QUERY sequence to a DATABASE of sequences (also called SUBJECT sequences)
- nucleotide or protein sequences
- Calculates statistical significance
- Available as an online web server , for example, at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi)

Web BLAST



BLAST programs

Program	Query	Database
blastp	protein	protein
blastn	nucleotide	nucleotide
blastx	nucleotide translated to protein	protein
tblastn	protein	nucleotide translated to protein
tblastx	nucleotide translated to protein	nucleotide translated to protein

Why would we want to use translated nucleotides?

BLAST

- Also available as a command line tool (guess which one we'll be using???)

- Need to conquer some basic concepts
  - Alignment
  - Scoring an alignment
  - Substitution matrices

Alignment

```
String A =   a b c d e
String B =   a c d e f

A (good) alignment would be:

String A =   a b c d e –
             |   | | |
String B =   a - c d e f
```

Many alignments are possible, we want to find the best

```
g c t g a a c g
c t a t a a t c
```

Bad:

```
g c t g a a c g - - - - - - -
- - - - - - - c t a t a a t c
```

Many alignments are possible, we want to find the best

```
g c t g a a c g
c t a t a a t c

Better?

g c t g - a a - c g
  | |     | |   |
- c t a t a a t c
```

To decide which alignment is best we need

- A way to examine all possible alignments
- A way to compute a score that gives the quality of the alignment

Scoring sequence similarity

- A simple scheme
  - +1 for a match
  - -1 for a mismatch

```
String A =      a b c d e
                |   | | |
String B =      a c c d e
```

```
+ 4
- 1

Total Score: 3
```

Scoring based on Biology

- Nucleotides are not mutated randomly
- Transition mutations are more common
  - Purine (A/G) to purine (A/G)
  - Pyrimidine (C/T) to pyrimidine (C/T)
- Transversion mutations are less common
- Can build a scoring scheme to reflect this:
  - Residue is the same = +1
  - Residue undergoes transition = 0
  - Residue undergoes transversion = -1

Scoring Based on Biology

- Amino Acids are not mutated at random either
- Those of similar physicochemical types are more likely to replace each other
- Instead of guessing what these rates might be, can measure empirically

Scoring Based on Biology

- Margaret Dayhoff (1978)
  - Collected statistics on protein substitution frequencies
  - Built the first set of protein substitution matrices
  - Point accepted mutation (PAM) matrices

BLOSUM

- BLOSUM (BLOck SUbstitution Matrix) - Henikoff and Henikoff
- A new substitution matrix, preferred today
- Much better for more divergent species (constructed using divergent species alignments)
- BLOSUM62 is the matrix used by default in most recent alignment applications such as BLAST.

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Scoring Gaps

- What about gaps?
- Usually, a gap opening is more of a penalty than a gap extension
- Why? A single mutational even may insert more than one base.

- Commonly used is the affine gap penalty:

  Gap opening penalty of 11

  Gap extension penalty of 1 for each additional residue

Scoring Wrap Up

- Now we have good a way to score a particular alignment
  1. Score substitutions appropriately reflecting biology
  2. Score gaps appropriately reflecting biology
- But how to generate all the possible alignments?

Approximate Methods

- Need more speed!
- Approximate methods have been developed that are
  - Great at detecting close relationships
  - Inferior to exact methods for picking up distant relationships
  - Approximate! (IE no guarantee that the optimal match is found)
- Start with identical "words"
  - Called k-tuples or k-mers
  - Use these words to quickly find perfect matches
  - Then use the more slow methods to grow the matches
- BLAST works this way

Heuristic – any that employs a practical methodology not guaranteed to be optimal or perfect, but sufficient for the immediate goals

Significance of Alignments

- Now we can find the best scoring alignment (or at least approximately if using BLAST)
- But is it significant in the statistical sense?
  - What is the likelihood that you are observing true biological similarity (evolution) vs random chance?
- E (expect) value = the number of hits one can "expect" to see by chance when searching a database of a particular size
- Takes into account the size of the database but not the number of queries (beware of multiple testing!)
- Lower = more biologically meaningful

E values

E Value	How many random alignments just as good?
1	1 in 1
.2	1 in 5
1e-5	1 in 100,000
1e-9	1 in 1,000,000,000
0	0%

BLAST

- high-scoring segment pairs (HSP)
- A query and a match sequence can have more than one HSP

Review

- Compares a sequence query to a set of sequences
- Uses scoring and statistics to find a good alignment
- Heuristic – approximates the best alignment

Want to learn more about how BLAST works?

- Wheeler and Bhagwat. BLAST Quick Start http://www.ncbi.nlm.nih.gov/books/NBK1734/
- Wikipedia https://en.wikipedia.org/wiki/BLAST