

# Introduction to Basic Local Alignment Search Tool (BLAST)

Philip Ashton  
OUCRU

# Why is sequence similarity important?

- DNA -> RNA -> Protein primary structure -> protein secondary structure -> protein tertiary structure
- Protein tertiary structure is of fundamental importance to the function of proteins
- If two proteins have similar sequences, they are likely to have similar structures
- Therefore, we can make inferences about protein function purely from sequence similarity measures

# Where does biological variation come from?

- Biological sequences show complex patterns of similarity to each other
  - These patterns are often due to homology
  - Homologous sequences are those which share a common ancestor
- Sequences evolve due to **natural selection** acting on **random variation**
- Not all sequence changes we observe are due to natural selection, some are just due to **genetic drift**

# Smith-Waterman/Needleman-Wunsch

- Needleman-Wunsch - global alignment, 1970
- Smith-Waterman – local alignment, 1981
- Guaranteed to find the best alignment (according to scoring criteria)
- Scales quadratically (requires as many calculations as the query length multiplied by the subject length)
- Too slow for many applications, but valuable if we want to be absolutely sure of the answer

**Match Score**   **Mismatch Score**   **Gap Score**

1

-1

-2

	A	C	T	A	T	G	G	G	
	0	-2	-4	-6	-8	-10	-12	-14	-16
A	-2	1 ↑	-1 ↓	-3 ↓	-5 ↓	-7 ↓	-9 ↓	-11 ↓	-13 ↓
C	-4	-1 ↑	2 ↓	0 ↓	-2 ↓	-4 ↓	-6 ↓	-8 ↓	-10 ↓
A	-6	-3 ↑	0 ↑	1 ↑	1 ↓	-1 ↓	-3 ↓	-5 ↓	-7 ↓
T	-8	-5 ↑	-2 ↑	1 ↑	0 ↑	2 ↓	0 ↓	-2 ↓	-4 ↓
G	-10	-7 ↑	-4 ↑	-1 ↑	0 ↑	0 ↑	3 ↓	1 ↓	-1 ↓
A	-12	-9 ↑	-6 ↑	-3 ↑	0 ↑	-1 ↑	1 ↑	2 ↓	0 ↓
G	-14	-11 ↑	-8 ↑	-5 ↑	-2 ↑	-1 ↑	0 ↑	2 ↓	3 ↓

A C T A T G G G  
A C - A T G A G

**Score = 3**

Needleman-Wunsch algorithm  
for global alignment:

1. Choose a scoring system
2. Fill in the table
3. Traceback from the cell on the bottom

# The better option – BLAST, Altschul et al., 1990

“Seed and extend”

1. Break the query into “words” – 3 AA, 11 nt
2. Look for exact matches between the words in the query and in each subject in the database.
3. For each query-subject match extend the alignment, calculating a score as you go.
4. Stop calculating for alignments where score goes below a certain threshold

# How do we score?

Nucleotide

Match +1  
Mismatch -1  
Gap -1

Amino acid

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	

BLOSUM amino acid substitution matrix

# Putting sequence similarity on a firm statistical footing

- BLAST provides an ‘E-score’ or ‘E-value’, E stands for Expectation
- It is the number of times you would expect to see an alignment with a similar score by chance
  - Lower is better;  $10^{-30}$  is a frequently used threshold
- E-value calculation depends on the size of the search space (the query and database size), and the score of the alignment

# Example of BLAST output

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	<a href="#">botulinum neurotoxin [Clostridium botulinum]</a>	2637	2637	100%	0.0	100.00%	<a href="#">AFV13854.1</a>
<input type="checkbox"/>	<a href="#">botulinum neurotoxin type A [Clostridium botulinum]</a>	2504	2504	100%	0.0	94.37%	<a href="#">WP_014520039.1</a>
<input type="checkbox"/>	<a href="#">botulinum neurotoxin type A [Clostridium botulinum]</a>	2503	2503	100%	0.0	94.37%	<a href="#">WP_078992015.1</a>
<input type="checkbox"/>	<a href="#">botulinum neurotoxin type A [Clostridium botulinum]</a>	2488	2488	100%	0.0	93.75%	<a href="#">WP_011948511.1</a>
<input type="checkbox"/>	<a href="#">botulinum neurotoxin type A [Clostridium botulinum]</a>	2488	2488	100%	0.0	93.75%	<a href="#">WP_061316836.1</a>
<input type="checkbox"/>	<a href="#">neurotoxin A [Clostridium botulinum]</a>	2488	2488	100%	0.0	93.67%	<a href="#">ABM73969.1</a>
<input type="checkbox"/>	<a href="#">RecName: Full=Botulinum neurotoxin type A; Short=BoNT/A; AltName: Full=Bontoxilysin-A</a>	2487	2487	100%	0.0	93.67%	<a href="#">P0DPI0.1</a>

Database is the complete non-redundant NCBI protein database

# Example of hit with e-value and bit score etc.

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	<a href="#">botulinum neurotoxin type A [Clostridium botulinum]</a>	2488	2488	100%	0.0	93.75%	<a href="#">WP_011948511.1</a>
<input type="checkbox"/>	<a href="#">Chain A, Botulinum neurotoxin type A</a>	594	594	25%	0.0	89.85%	<a href="#">6DKK_A</a>
<input type="checkbox"/>	<a href="#">Chain A, Botulinum neurotoxin type A</a>	590	590	25%	0.0	89.54%	<a href="#">6MHJ_A</a>
<input type="checkbox"/>	<a href="#">peptidase M27 [Clostridium botulinum]</a>	85.9	85.9	21%	9e-18	27.72%	<a href="#">WP_011948510.1</a>

Database is a single *C. botulinum* genome

# BLAST is the most important piece of bioinformatics software\*. Why?

1. The problem it solves, sequence similarity search, gives us a really powerful tool for identifying unknowns in the biological world
2. It is fast
3. It is reliable
4. It is flexible, can be used for many sequence analysis scenarios
5. It's so widely used that most biologists understand it as a verb

\* BLAST paper has > 70,000 citations

BLAST; Korf, Bedell, Yandell; O'Reilly Media; 2003



=

# Google!



?

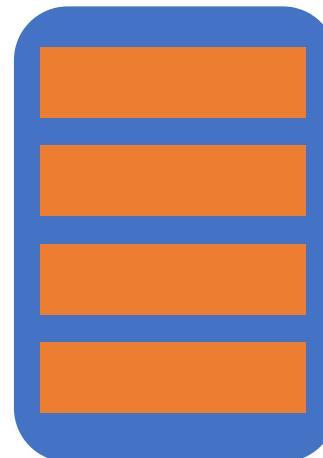
# BLAST Glossary

- Query – the sequence you are interested in
- Subject – the specific sequence you are comparing against
- Database – all the sequences you are comparing against



Query

Subject



Database

# BLAST vs web service

The screenshot shows the NCBI BLAST search interface. At the top, there are links for NIH, U.S. National Library of Medicine, NCBI, and Sign in to NCBI. The main title is "BLAST® > blastn suite". Below it, the "Standard Nucleotide BLAST" section is visible. The "Enter Query Sequence" field is highlighted. Other tabs include "blastn", "blastp", "blastx", "tblastn", and "tblastx". The "Enter accession number(s), gi(s), or FASTA sequence(s)" input field is empty. There are "Clear" and "Query subrange" buttons. Below the input field, there are fields for "From" and "To" dates. A "Choose file" button for uploading a file is present, showing "No file chosen". A "Job Title" input field and a "Enter a descriptive title for your BLAST search" field are also shown. A checkbox for "Align two or more sequences" is unchecked. The "Choose Search Set" section includes fields for "Database" (set to "Others (nr etc.)"), "Organism" (set to "Nucleotide collection (nr/nt)"), "Exclude" (checkboxes for "Models (XM/XP)" and "Uncultured/environmental sample sequences"), "Limit to" (checkbox for "Sequences from type material"), and "Entrez Query" (input field with "Create custom database" link). The "Program Selection" section has an "Optimize for" dropdown set to "Highly similar sequences (megablast)". Other options include "More dissimilar sequences (discontiguous megablast)" and "Somewhat similar sequences (blastn)". A "Choose a BLAST algorithm" link is at the bottom.

```
[flashton@Philips-MacBook-Pro:~/Dropbox/talaromyces_marneffii/phylo/results/2018.10.19$ blastn -h
USAGE
```

```
blastn [-h] [-help] [-import_search_strategy filename]
[-export_search_strategy filename] [-task task_name] [-db database_name]
[-dbsize num_letters] [-gilist filename] [-seqidlist filename]
[-negative_gilist filename] [-entrez_query entrez_query]
[-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
[-subject subject_input_file] [-subject_loc range] [-query input_file]
[-out output_file] [-evalue evalue] [-word_size int_value]
[-gapopen open_penalty] [-gapextend extend_penalty]
[-perc_identity float_value] [-qcov_hsp_perc float_value]
[-max_hsps int_value] [-xdrop_ungap float_value] [-xdrop_gap float_value]
[-xdrop_gap_final float_value] [-searchsp int_value]
[-sum_stats bool_value] [-penalty penalty] [-reward reward] [-no_greedy]
[-min_raw_gapped_score int_value] [-template_type type]
[-template_length int_value] [-dust DUST_options]
[-filtering_db filtering_database]
[-window_masker_taxid window_masker_taxid]
[-window_masker_db window_masker_db] [-soft_masking soft_masking]
[-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
[-best_hit_score_edge float_value] [-window_size int_value]
[-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
[-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
[-outfmt format] [-show_gis] [-num_descriptions int_value]
[-num_alignments int_value] [-line_length line_length] [-html]
[-max_target_seqs num_sequences] [-num_threads int_value] [-remote]
[-version]
```

#### DESCRIPTION

Nucleotide-Nucleotide BLAST 2.4.0+

Use '-help' to print detailed descriptions of command line arguments

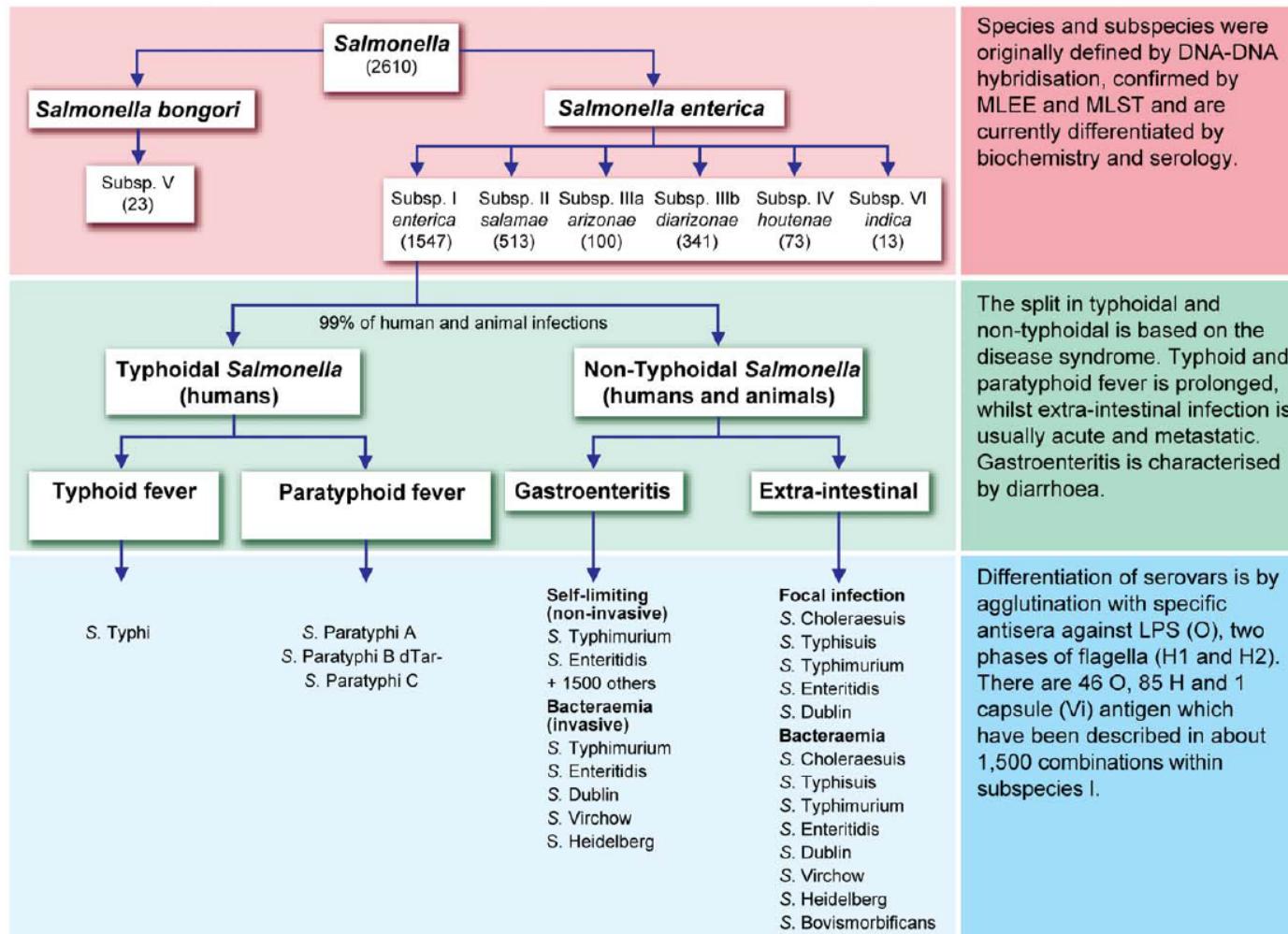
# Repeats

- Simple
  - Repeats of the same nucleotide e.g. TTTTTTTTTTTTT
  - Repeats of the same di-nucleotide e.g. CACACACACACACA
  - Repeats of the same tri-nucleotide e.g. TGCTGCTGCTGC
  - Etc
- Complex
  - Non-coding RNAs like ribosomal RNA – *E. coli* has 7 identical copies of rRNA encoding locus
  - Transposons – ‘jumping genes’, and their ‘cargo’ (e.g. AMR genes)
  - Repeated protein domains

# Different types of BLAST

<b>Program</b>	<b>Query sequence type</b>	<b>Target sequence type</b>	
BLASTP	Protein	Protein	Compares an amino acid query sequence against a protein sequence database
BLASTN	Nucleotide	Nucleotide	Compares a nucleotide query sequence against a nucleotide sequence database
BLASTX	Nucleotide (translated)	Protein	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database
TBLASTN	Protein	Nucleotide (translated)	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
TBLASTX	Nucleotide (translated)	Nucleotide (translated)	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

# Introduction to Salmonella serotyping



- Kauffmann-White scheme
- 46 O antigens (lipopolysaccharide)
- 85 H antigens (flagellar)
- 1500 combinations in subsp *enterica*
- O antigen; phase 1; phase 2
- E.g. *S. Paratyphi A* is 1,4,5,12; b; 1,2

# Further reading

- BLAST; Korf, Bedell, Yandell; O'Reilly Media; 2003
  - <http://shop.oreilly.com/product/9780596002992.do>
- Having a BLAST with bioinformatics, Pertsemlidis & Fondon, 2002
  - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC138974/>