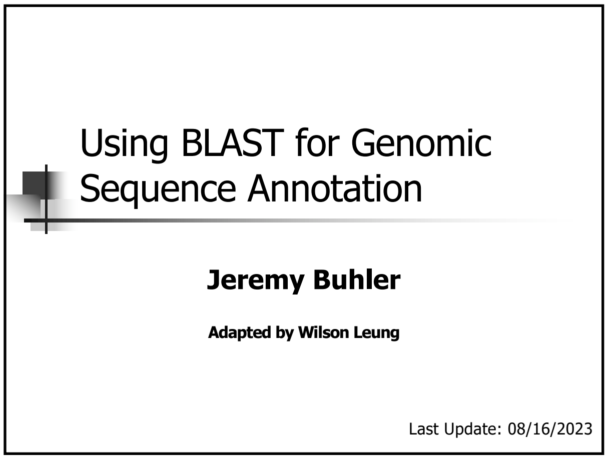


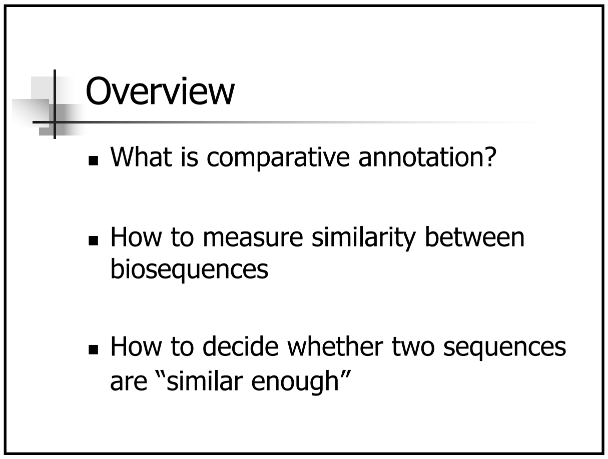
Using BLAST for Genomic Sequence Annotation

Jeremy Buhler

Adapted by Wilson Leung

Last Update: 08/16/2023

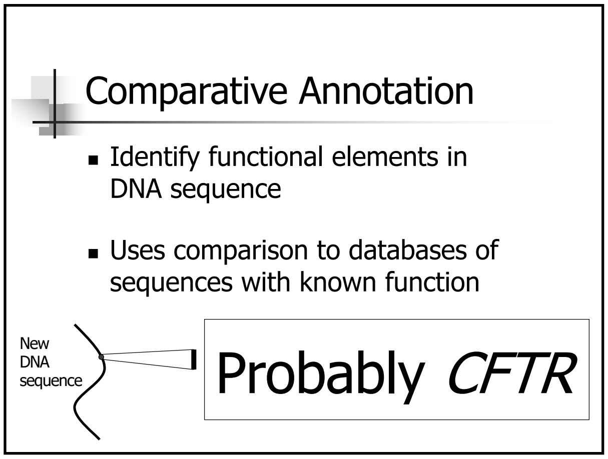
1



Overview

- What is comparative annotation?
- How to measure similarity between biosequences
- How to decide whether two sequences are "similar enough"

2



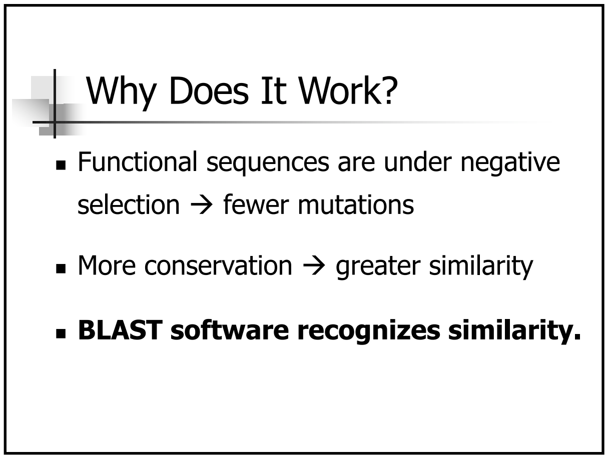
Comparative Annotation

- Identify functional elements in DNA sequence
- Uses comparison to databases of sequences with known function

New DNA sequence

Probably *CFTR*

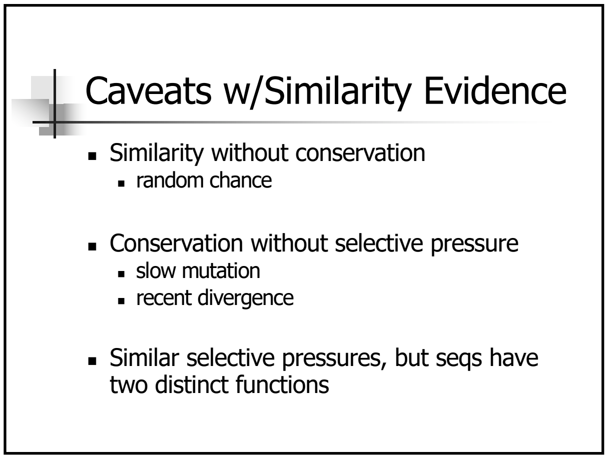
3



Why Does It Work?

- Functional sequences are under negative selection → fewer mutations
- More conservation → greater similarity
- **BLAST software recognizes similarity.**

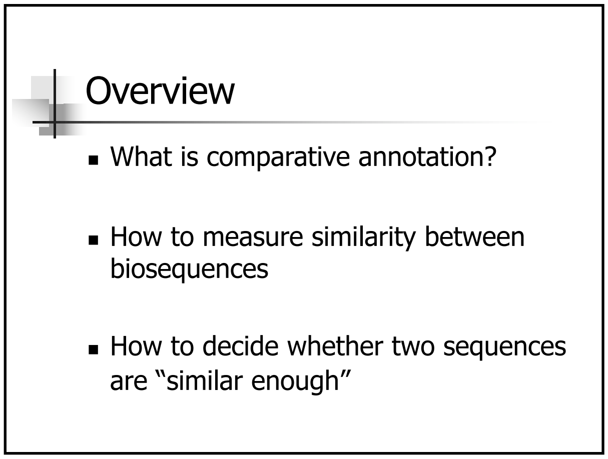
4



Caveats w/Similarity Evidence

- Similarity without conservation
  - random chance
- Conservation without selective pressure
  - slow mutation
  - recent divergence
- Similar selective pressures, but seqs have two distinct functions

5



Overview

- What is comparative annotation?
- How to measure similarity between biosequences
- How to decide whether two sequences are "similar enough"

6

What is **Similarity**?

- How to measure similarity of two DNA seqs?
- Mutations happen…
- Measure should reflect desired evolutionary inference

7

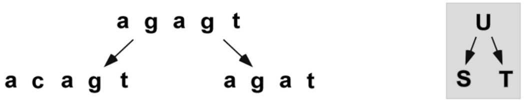
Mutational Model

- Sequences change by series of events of (only) three types:
  - **Substitution** of one base ACG ⟹ ATG
  - **Insertion** of one base ACG ⟹ ACAG
  - **Deletion** of one base ACC ⟹ AG

8

Sequence History (1/2)

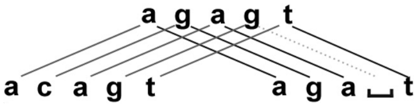
- Suppose seqs S, T diverged from a common ancestral sequence U…



9

Sequence History (2/2)

- Draw lines between bases of S and T that come from same base of U.

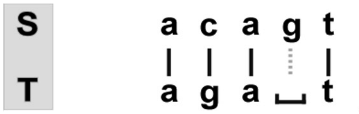


- This is a "tree alignment" of S,T,U.

10

Sequence Alignment

- Now elide the ancestor…

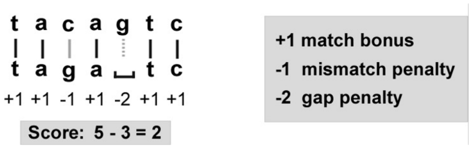


- Result is correspondence between bases of S, T – a sequence alignment

11

Similarity Score of Alignment

- Fewer mutations → more conservation



- Give bonus for matches, penalties for substitutions and gaps

12

One Small Problem…

- Do you own a time machine?
- If not, how do you know
  - ancestral sequence U?
  - history of mutation?
- Hence, how to get correct alignment?

13

What We Do In Practice

- Guess an alignment that minimizes # of hypothesized mutations

```
a a g c c — — — — —     S
— — — — — a a t c c     T

          a a g c c     S
          | |   | |
          a a t c c     T
```

- (more precisely, maximizes score)

14

Overview

- What is comparative annotation?
- How to measure similarity between biosequences
- How to decide whether two sequences are "similar enough"

15

Deciding What to Report

- Any two sequences can be aligned with ***some*** score.
- Higher scores are better…
- When is score high enough to be evidence of conservation?

16

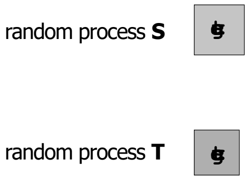
Idea: Test a Null Hypothesis

- Suppose two DNA seqs S, T are **completely unrelated**.
- What is probability that best alignment between S, T has score at least Θ?
- If score(S,T) is unlikely to occur by chance, then report (S,T)

17

Null Model Assumptions

- Bases of seqs S, T generated independently at random

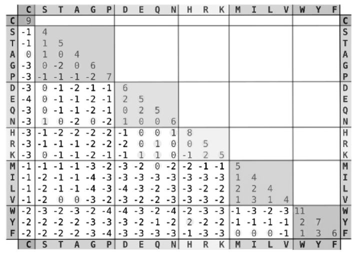


random process **S**

random process **T**

18

Scoring Systems



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

BLOSUM62 Matrix

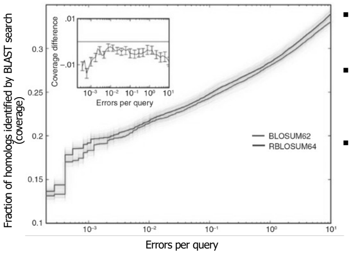
Amino acids grouped by the Dayhoff classification scheme

"BLOSUM". Wikipedia. Ppgardne.

- Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004 Aug;22(8):1035-6.

19

Errors in the BLOSUM Matrices



Styczynski MP, et al. BLOSUM62 miscalculations improve search performance. Nat Biotechnol. 2008 Mar;26(3):274-5.

- Benchmark: detect distant homologs in the ASTRAL database
- The "correct" matrix (**RBLOSUM64**) performs worse than the **BLOSUM62** matrix
- Subsequent "fix" of the BLOSUM matrices:
  - Hess M, et al. Addressing inaccuracies in BLOSUM computation improves homology search performance. BMC Bioinformatics. 2016 Apr 27;17:189.

20

P-values

- Given random seqs S′, T′ with same base distributions as S, T
- Karlin-Altschul theory tells us probability that S′, T′ align with score at least Θ
- If p(Θ) is small, report alignment of S,T

21

E-values

- For computational reasons, BLAST reports not p(Θ) but rather E(Θ)
- E(Θ) = expected # times alignment with score at least Θ happens by chance in current search
- If E(Θ) < 1, then score Θ is interesting

22

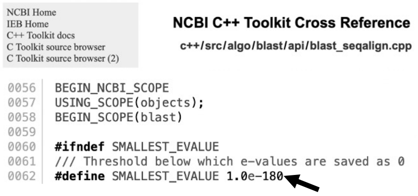
Caveats about E-values

- Model from which E-values are computed is too simple for real bioseqs
- Large margin of safety is wise
- Be very skeptical of "matches" with E > 10^{-5}

23

Explanation for E-value = 0.0

- E-value is less than **1.0e-180**



NCBI C++ Toolkit Cross Reference

c++/src/algo/blast/api/blast_seqalign.cpp

```
BEGIN_NCBI_SCOPE
USING_SCOPE(objects);
BEGIN_SCOPE(blast)

#ifndef SMALLEST_EVALUE
/// Threshold below which e-values are saved as 0
#define SMALLEST_EVALUE 1.0e-180
```

24

Summary

- Comparative annotation with BLAST uses similarity as evidence for conserved function.
- Similarity score based on hypothesized evolutionary relations among sequences.
- E-values indicate whether scores are high enough to be real biological conservation.

25

Additional Resources

- Introduction to Dynamic Programming
  - Overview of the algorithms for calculating global, semiglobal, and local alignments
- From Smith-Waterman to BLAST
  - Discuss the heuristics used by BLAST to reduce the search space and quickly report high-scoring local alignments

26