Using BLAST for Genomic Sequence Annotation

Jeremy Buhler

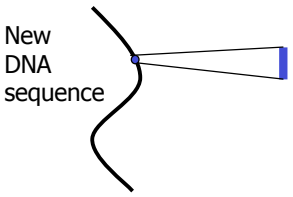
For HHMI / BIO4342
Tutorial Workshop

Overview

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

Comparative Annotation

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



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

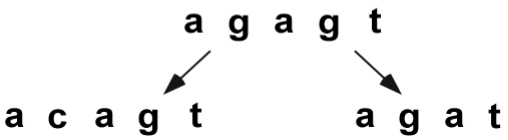
Mutational Model

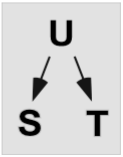
- Sequences change by series of events of (only) three types:
  - **Substitution** of one base `acg` ⇒ `atg`
  - **Insertion** of one base `acg` ⇒ `acag`
  - **Deletion** of one base `acg` ⇒ `ag`

8

Sequence History (1/2)

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

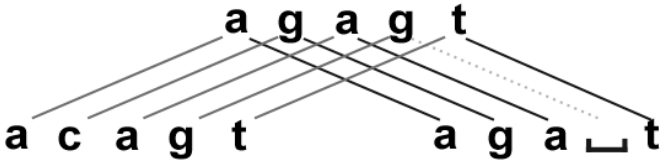




9

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

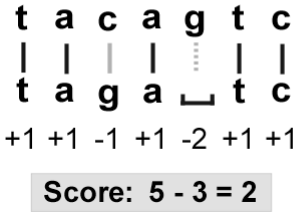
- Now elide the ancestor…

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

11

Similarity Score of Alignment

- Fewer mutations → more conservation



Score: 5 - 3 = 2

+1 match bonus
-1 mismatch penalty
-2 gap penalty

- 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

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

19

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

20

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 $E > 10^{-5}$

21

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.

22