02-251: *Great Ideas in Computational Biology*

Metagenomics

Phillip Compeau

Introduction to Metagenomics

Fill in the Blank: Over half the cells in your body are: ___________.

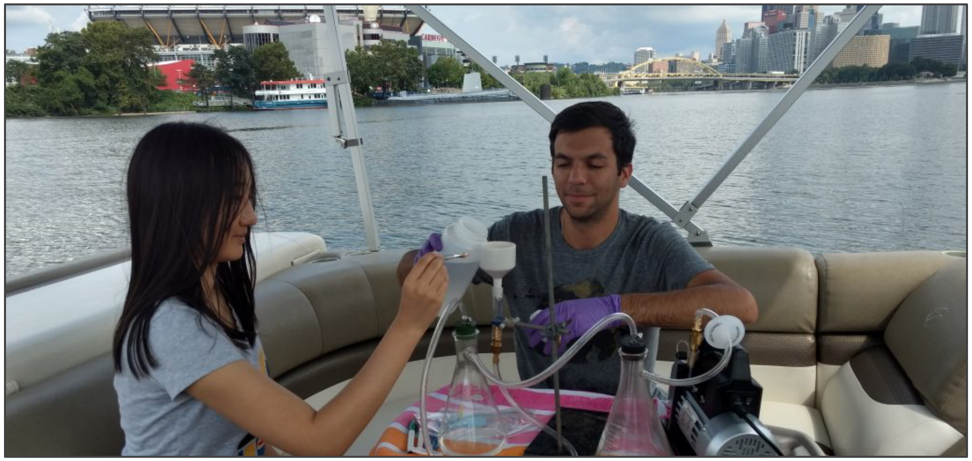
Introduction to Metagenomics

Fill in the Blank: Over half the cells in your body are: *bacteria*_____.

Introduction to Metagenomics

Fill in the Blank: Over half the cells in your body are: *bacteria*_____.

Metagenomics: The study of DNA (or RNA) recovered from an environmental sample.



An Example of Metagenomics

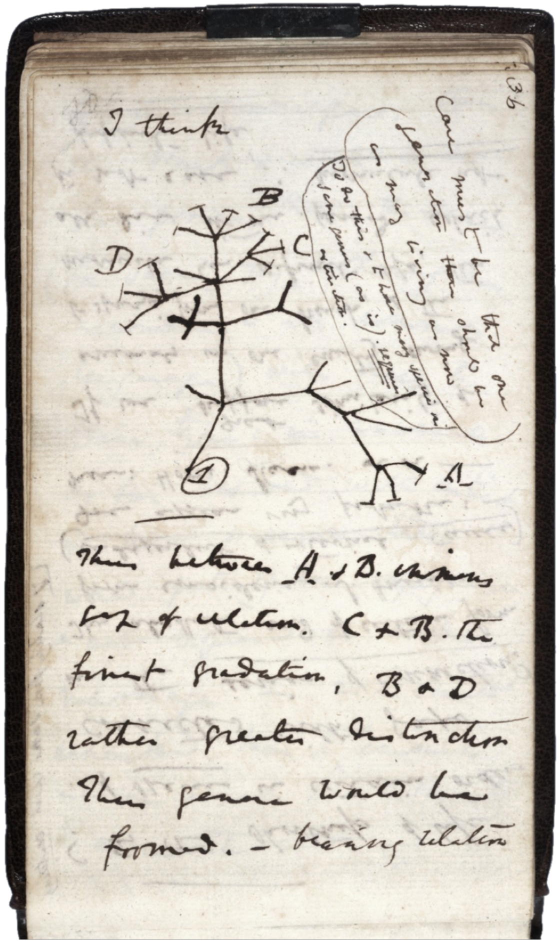
BODY + BRAIN

Nearly Half of the Microbes In New York City's Subways Are Undiscovered Species

BY ALLISON ECK TUESDAY, JUNE 23, 2015 NOVA NEXT

https://www.pbs.org/wgbh/nova/article/nearly-half-of-the-microbes-in-new-york-citys-subways-are-undiscovered-species/

"Why Do We Care About Bacteria?"



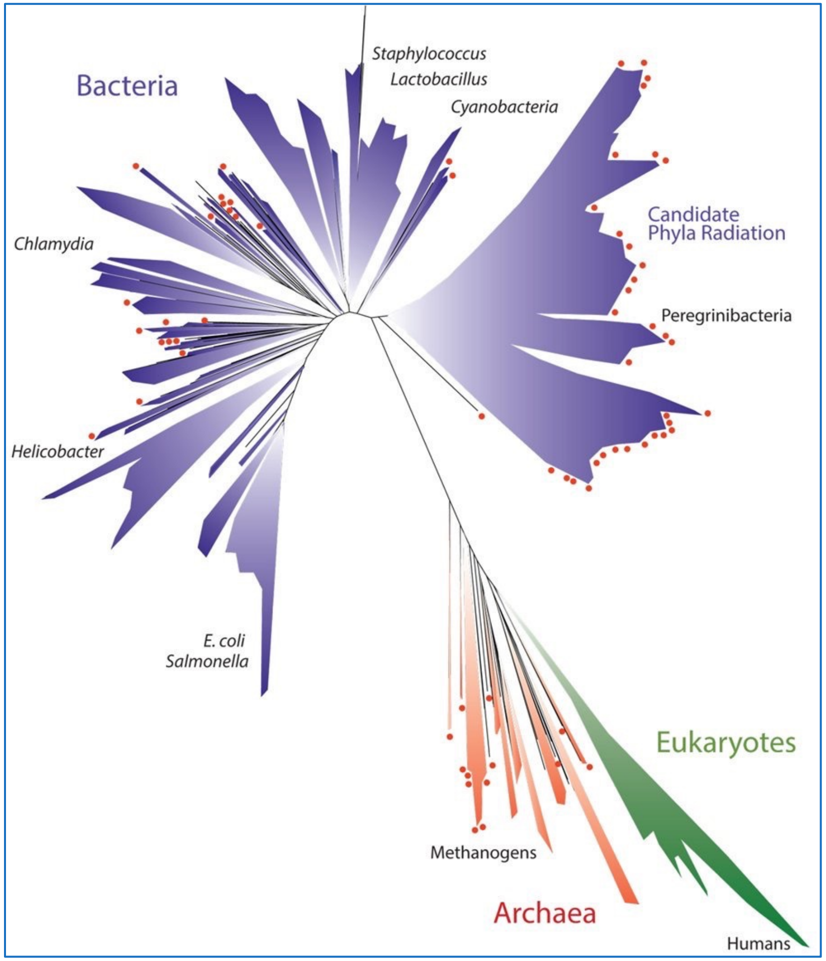
36

I think

case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.

Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation

Darwin's notebook c. 1837



Hug et al., 2016
Courtesy: *Nature Biotechnology, Discovery Magazine*

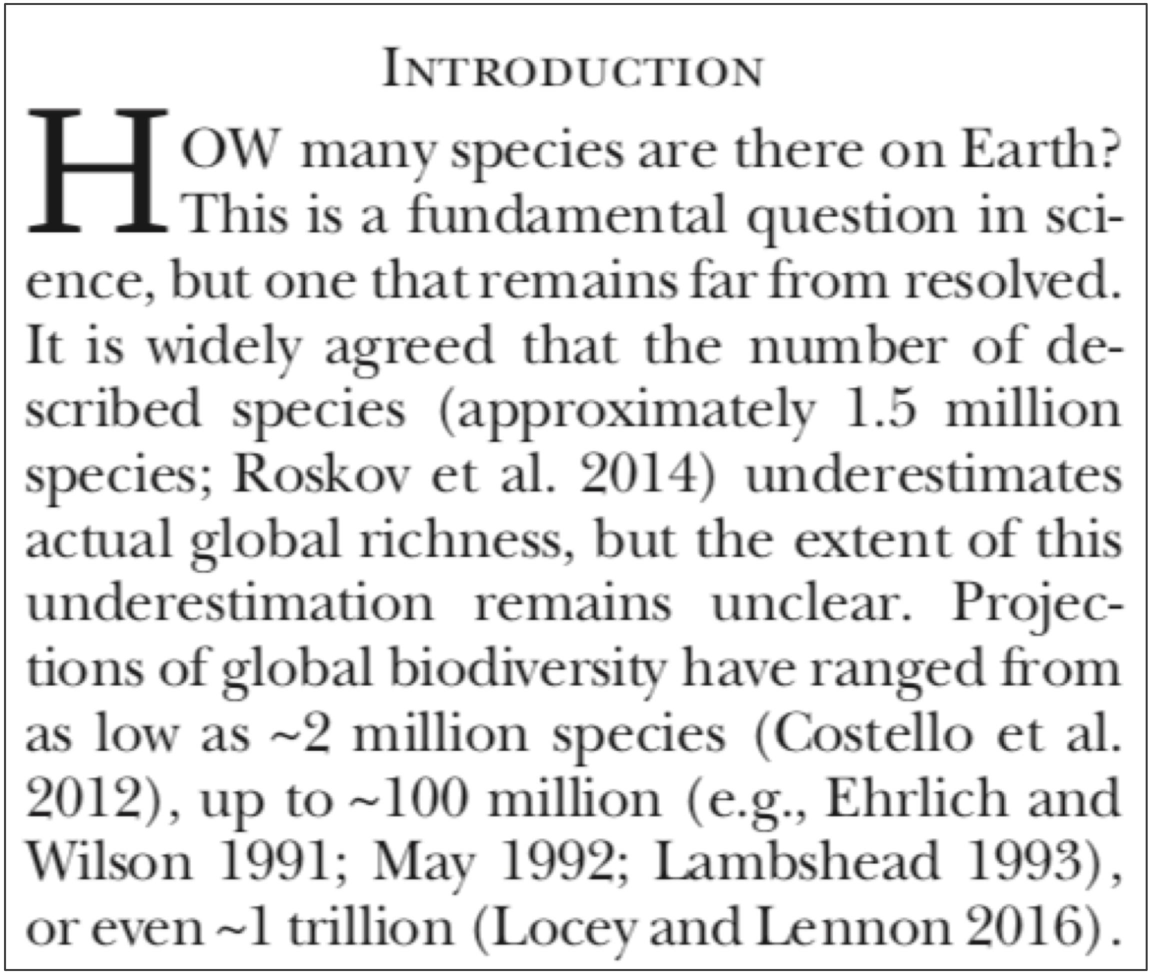
Wait … How Many Species Are There?

Checkpoint: … I'm open to guesses on what you think!

Wait … How Many Species Are There?

Checkpoint: … I'm open to guesses on what you think!

Another can of worms: what is a species?



INTRODUCTION

HOW many species are there on Earth? This is a fundamental question in science, but one that remains far from resolved. It is widely agreed that the number of described species (approximately 1.5 million species; Roskov et al. 2014) underestimates actual global richness, but the extent of this underestimation remains unclear. Projections of global biodiversity have ranged from as low as ~2 million species (Costello et al. 2012), up to ~100 million (e.g., Ehrlich and Wilson 1991; May 1992; Lambshead 1993), or even ~1 trillion (Locey and Lennon 2016).

Larsen et al., 2017

Single-cell Sequencing Offers One Way to Analyze an Environmental Sample

Note: Many species cannot be cultured (i.e., separated and grown from a sample).

There do exist **single-cell sequencing** approaches for isolating a cell, amplifying its DNA, and sequencing its genome. But a small sample may contain thousands of species!

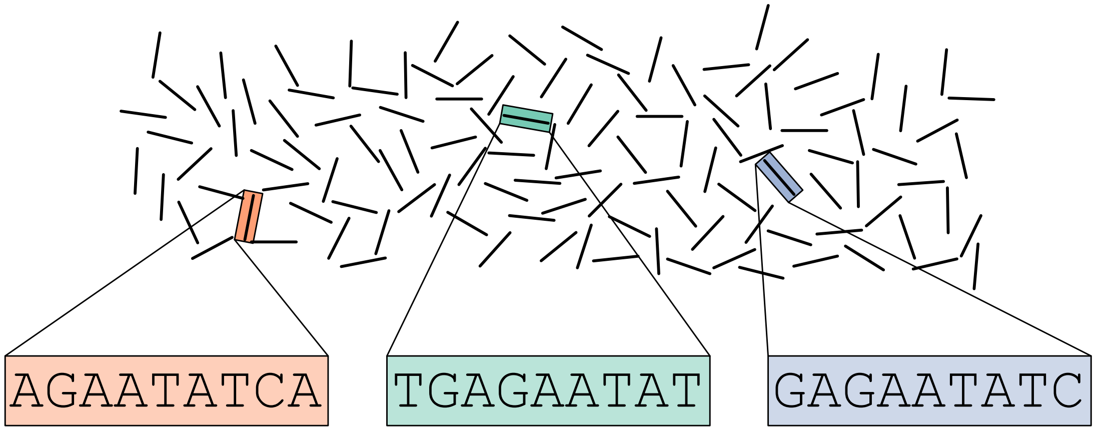
Recall How Sequencing Works

Multiple identical copies of a genome

Shatter the genome into reads

Sequence the reads

(Lab)



AGAATATCA

TGAGAATAT

GAGAATATC

Assemble the genome using overlapping reads

(Computational)

```
     AGAATATCA
    GAGAATATC
   TGAGAATAT
...TGAGAATATCA...
```

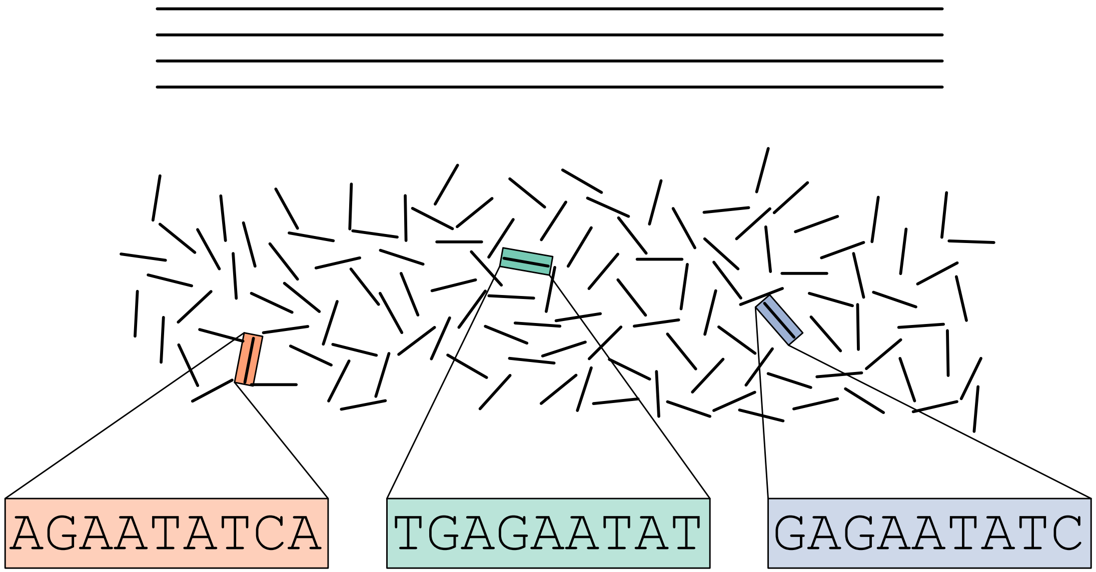
Recall How Sequencing Works

Multiple ~~identical~~ copies of a genome

Shatter the genome into reads

Sequence the reads

(Lab)



Checkpoint: How would you solve this problem computationally?

Two Approaches for Metagenomics

1. Try to reconcile de Bruijn graph approach to construct multiple graphs and do *n* assemblies (where *n* is unknown in advance).

Two Approaches for Metagenomics

1. Try to reconcile de Bruijn graph approach to construct multiple graphs and do *n* assemblies (where *n* is unknown in advance).

2. Use “puzzle on the box” database of all known organisms’ genomes and “align” metagenomics reads to these genomes to find which ones (if any) each read may have come from. This process is called **binning** the reads.

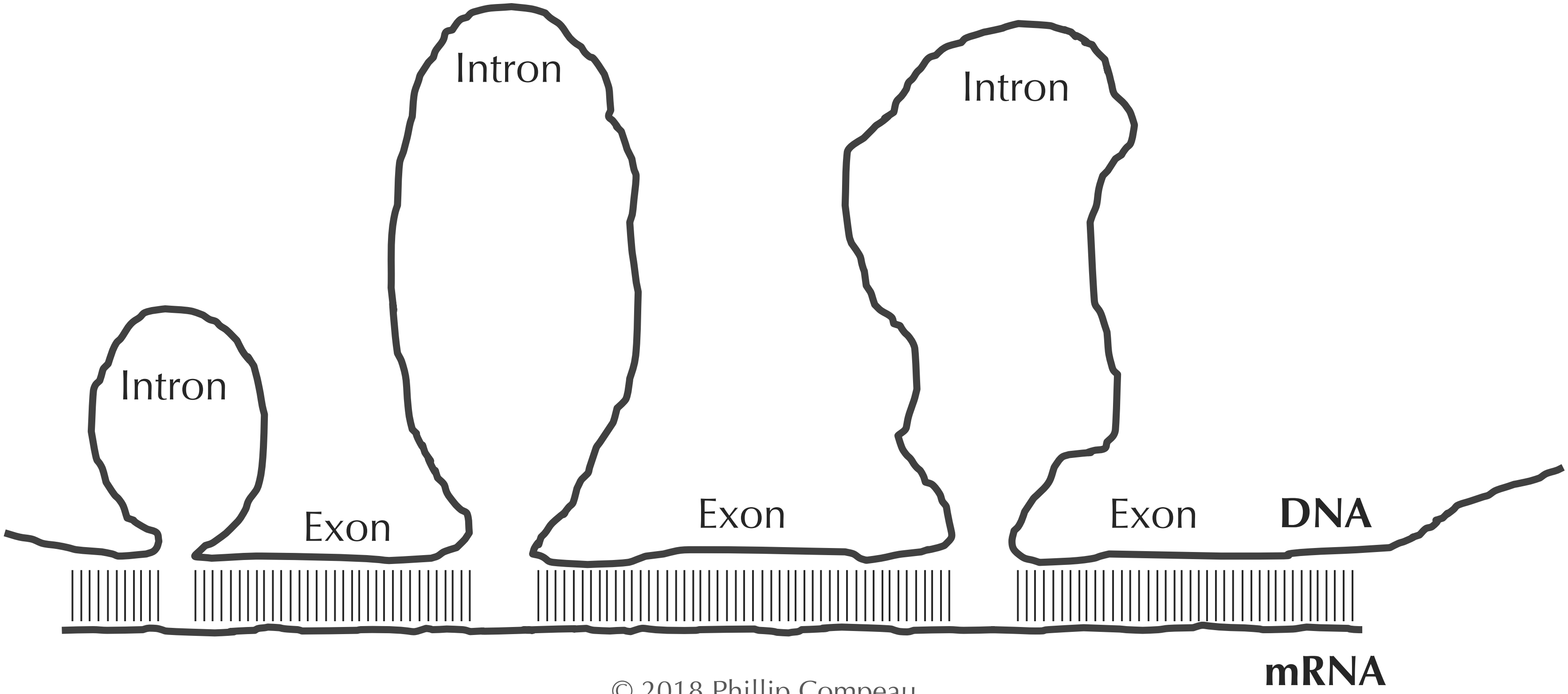
Two Approaches for Metagenomics

1. Try to reconcile de Bruijn graph approach to construct multiple graphs and do *n* assemblies (where *n* is unknown in advance).

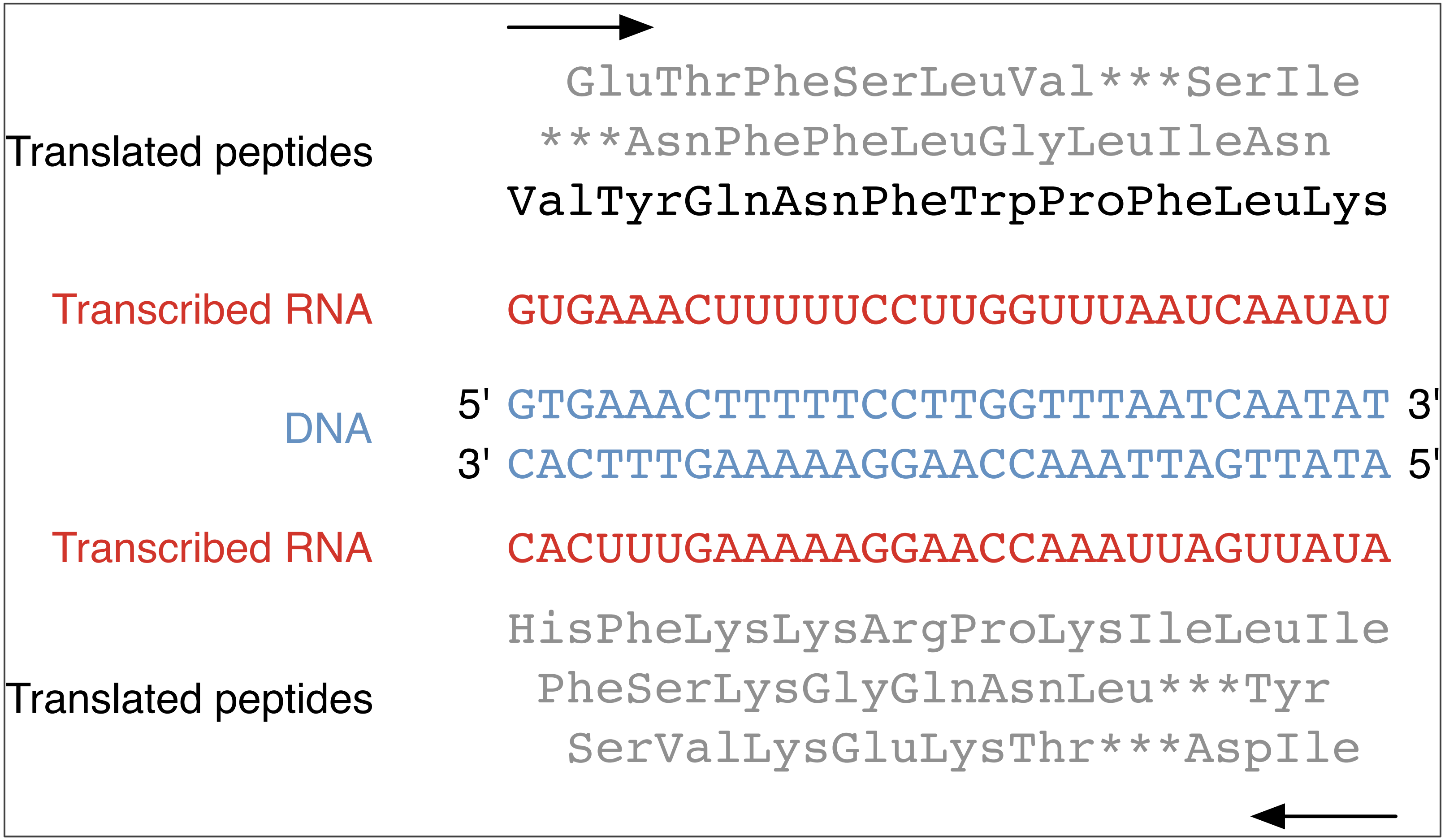
2. Use “puzzle on the box” database of all known organisms’ genomes and “align” metagenomics reads to these genomes to find which ones (if any) each read may have come from. This process is called **binning** the reads.

Prokaryotes (Bacteria and Archaea) Don’t Have Introns

Prokaryotes’ genes generally don’t have introns (i.e., the entire gene is translated into protein).



DNA Has Six "Reading Frames" for Translation into Protein



Translated peptides

```
  GluThrPheSerLeuVal***SerIle
 ***AsnPhePheLeuGlyLeuIleAsn
ValTyrGlnAsnPheTrpProPheLeuLys
```

Transcribed RNA

GUGAAACUUUUUCCUUGGUUUAAUCAAUAU

DNA

5' GTGAAACTTTTTCCTTGGTTTAATCAATAT 3'

3' CACTTTGAAAAAGGAACCAAATTAGTTATA 5'

Transcribed RNA

CACUUUGAAAAAGGAACCAAAUUAGUUAUA

Translated peptides

```
HisPheLysLysArgProLysIleLeuIle
 PheSerLysGlyGlnAsnLeu***Tyr
  SerValLysGluLysThr***AspIle
```

Binning Reads is a Protein Comparison Problem

Moreover, most of the prokaryotic genome is made up of genes.

Binning Reads is a Protein Comparison Problem

Moreover, most of the prokaryotic genome is made up of genes.

Because protein-protein comparisons can be more fruitful, we should compare all six protein translations of each DNA read against a (huge) database consisting of all known prokaryotic proteins.

Why Not Use Sequence Alignment?

Checkpoint: We could perform a local alignment of each protein product of each read against the entire database. What would be its runtime?

Why Not Use Sequence Alignment?

Checkpoint: We could perform a local alignment of each protein product of each read against the entire database. What would be its runtime?

Answer: O(|*Database*|*|*Pattern*|) for each pattern, for a total of O(|*Database*|*|*Patterns*|). Not to mention building the array …

Why Not Just Use BWT?

Recall our BWT-based algorithm for finding all pattern matches with up to *d* mismatches.

Why Not Just Use BWT?

Recall our BWT-based algorithm for finding all pattern matches with up to *d* mismatches.

1. Divide *Pattern* into *d*+1 "equal" segments (called **seeds**).

Why Not Just Use BWT?

Recall our BWT-based algorithm for finding all pattern matches with up to *d* mismatches.

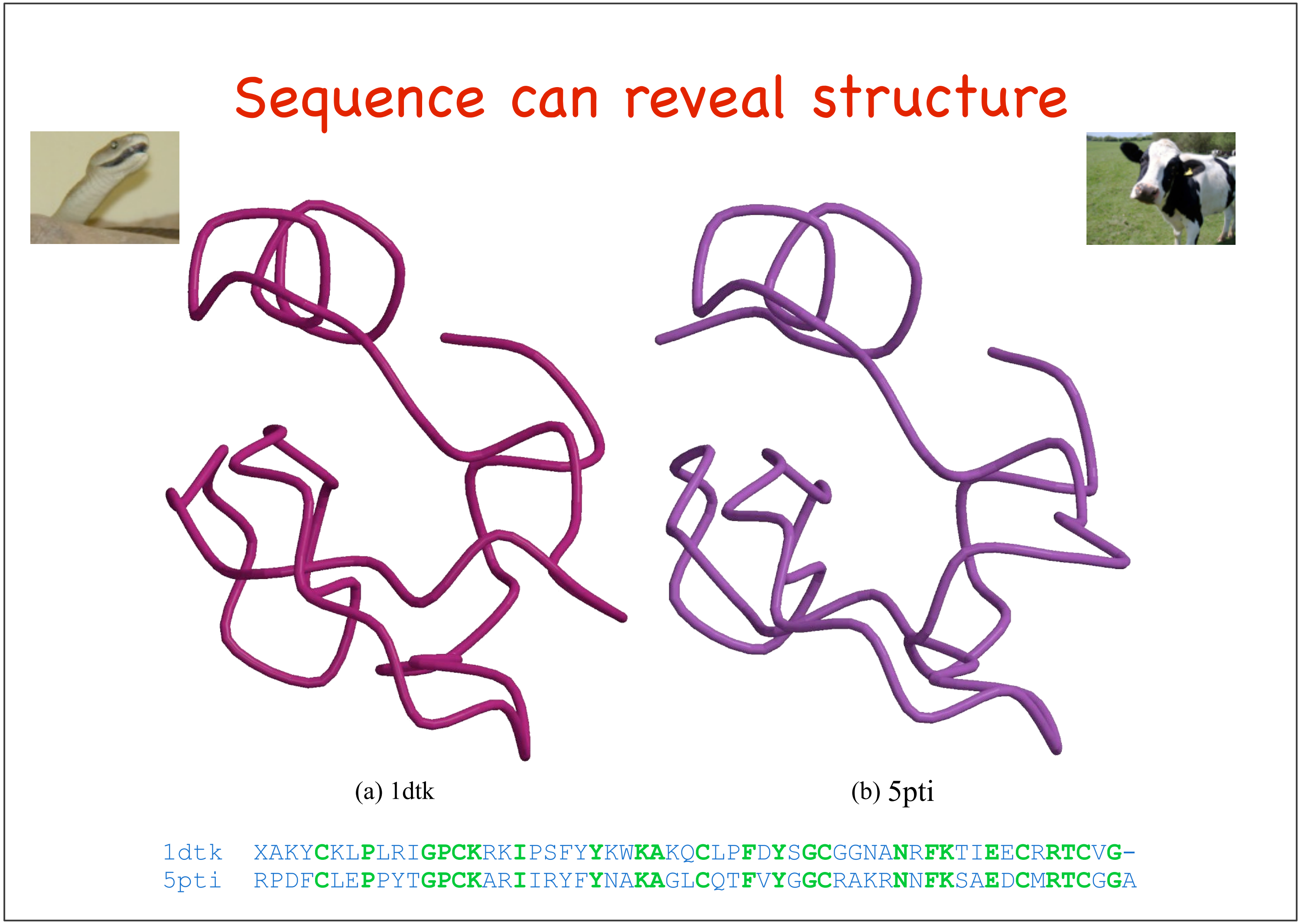
1. Divide *Pattern* into *d*+1 ″equal″ segments (called **seeds**).
2. Find which seeds match *Text* exactly using BWT (**seed detection**).

Why Not Just Use BWT?

Recall our BWT-based algorithm for finding all pattern matches with up to *d* mismatches.

1. Divide *Pattern* into *d*+1 "equal" segments (called **seeds**).
2. Find which seeds match *Text* exactly using BWT (**seed detection**).
3. Attempt to extend all seeds in both directions to verify whether *Pattern* occurs with at most *d* mismatches (**seed extension**).

Let's Borrow CK's Slide …



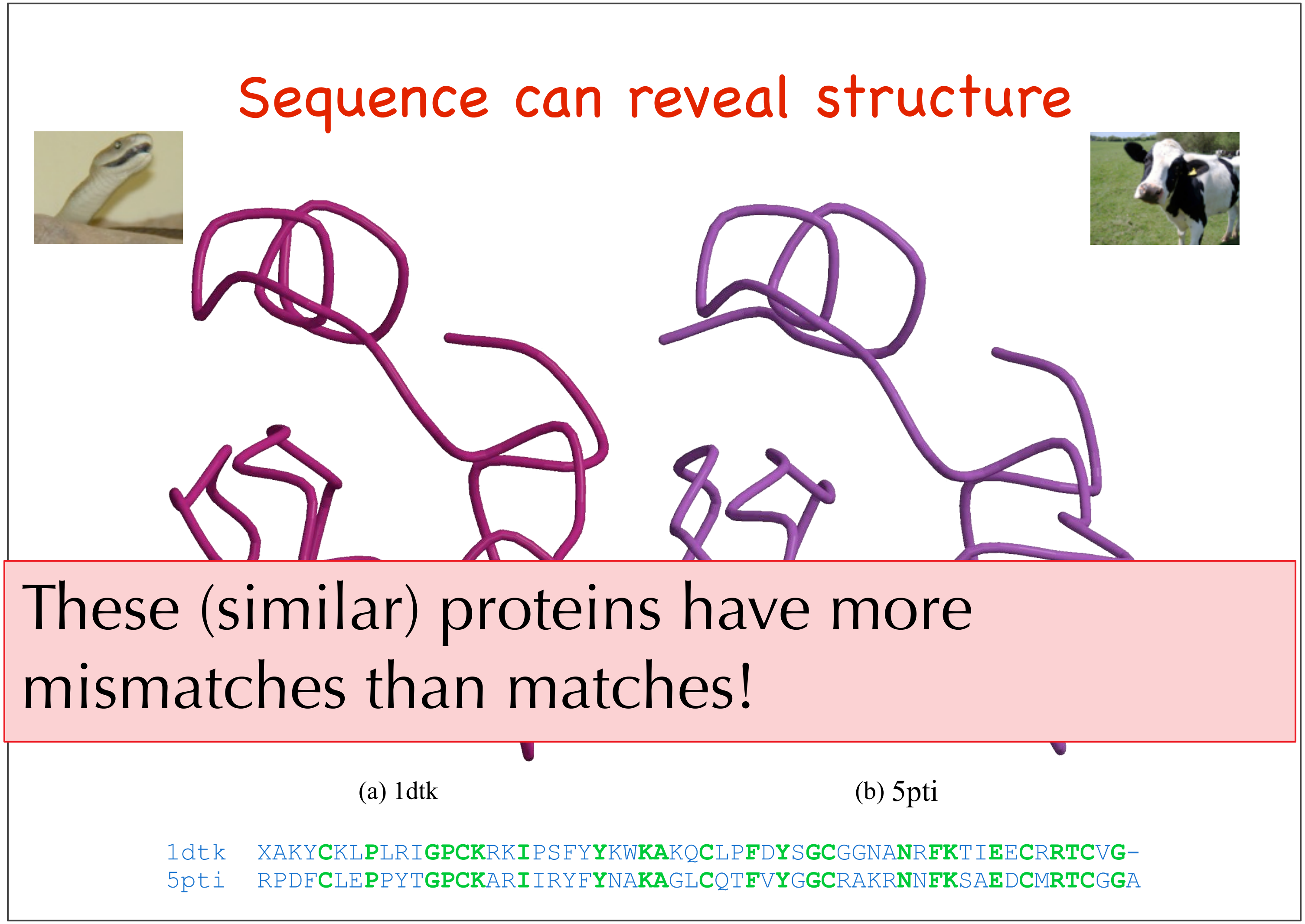
Sequence can reveal structure

(a) 1dtk

(b) 5pti

```
1dtk   XAKYCKLPLRIGPCKRKIPSFYYKWKAKQCLPFDYSGCGGNANRFKTIEECRRTCVG-
5pti   RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
```

Let's Borrow CK's Slide ...



Sequence can reveal structure

These (similar) proteins have more mismatches than matches!

(a) 1dtk

(b) 5pti

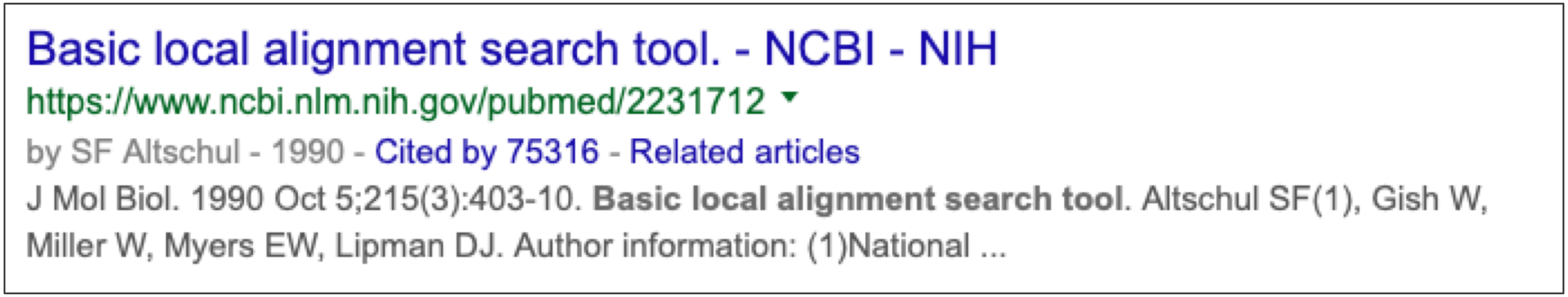
```
1dtk   XAKYCKLPLRIGPCKRKIPSFYYKWKAKQCLPFDYSGCGGNANRFKTIEECRRTCVG-
5pti   RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
```

We Need a “Just Right” Binning Approach

Accuracy: Able to take protein-level comparisons into account and find “correct” alignments.

Speed: But still fast enough to be practical.

BLAST in a Nutshell



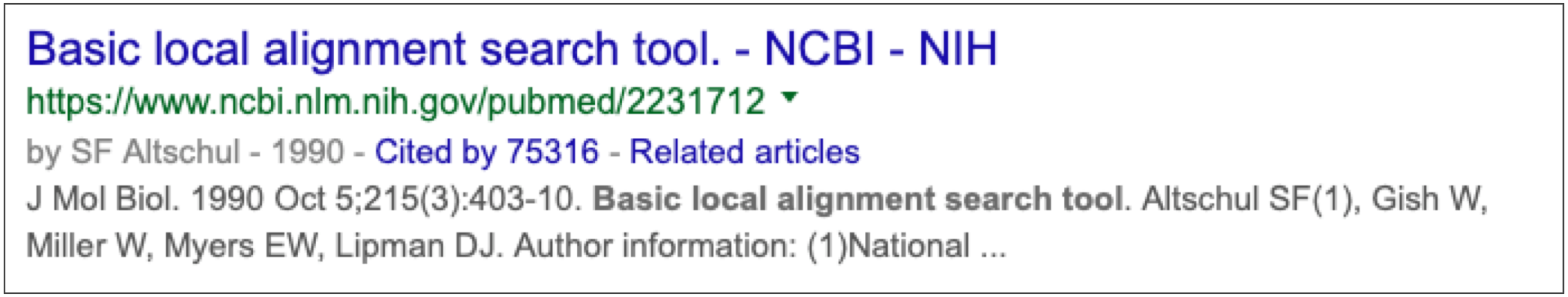
Basic local alignment search tool. - NCBI - NIH

https://www.ncbi.nlm.nih.gov/pubmed/2231712 ▾

by SF Altschul - 1990 - Cited by 75316 - Related articles

J Mol Biol. 1990 Oct 5;215(3):403-10. **Basic local alignment search tool**. Altschul SF(1), Gish W, Miller W, Myers EW, Lipman DJ. Author information: (1)National ...

BLAST in a Nutshell



Basic local alignment search tool. - NCBI - NIH

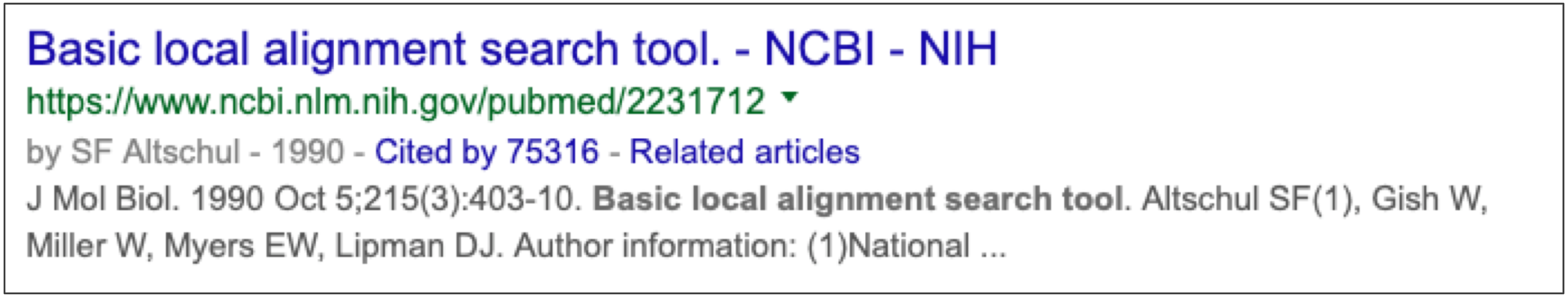
https://www.ncbi.nlm.nih.gov/pubmed/2231712

by SF Altschul - 1990 - Cited by 75316 - Related articles

J Mol Biol. 1990 Oct 5;215(3):403-10. **Basic local alignment search tool.** Altschul SF(1), Gish W, Miller W, Myers EW, Lipman DJ. Author information: (1)National ...

Key Point: BLAST is a **heuristic**, meaning that it does not promise that it will bin all reads correctly; in other words, it is not solving a computational problem exactly.

BLAST in a Nutshell



Basic local alignment search tool. - NCBI - NIH

https://www.ncbi.nlm.nih.gov/pubmed/2231712 ▾

by SF Altschul - 1990 - Cited by 75316 - Related articles

J Mol Biol. 1990 Oct 5;215(3):403-10. **Basic local alignment search tool**. Altschul SF(1), Gish W, Miller W, Myers EW, Lipman DJ. Author information: (1)National ...

- **Input:** a *database* (in this case a protein) and a *query* (in this case a protein corresponding to one of six reading frame translations of sequencing read). Plus, some parameters (later).
- **Output:** A collection of high-scoring local alignments of the query against the database (there may be none, or more than are found).

BLAST in a Nutshell

BLAST uses a modified “seed and extend” approach, in which the seeds are based on a protein scoring matrix (e.g., BLOSUM62) to allow for more robust scoring.

BLAST in a Nutshell

BLAST uses a modified “seed and extend” approach, in which the seeds are based on a protein scoring matrix (e.g., BLOSUM62) to allow for more robust scoring.

Gaps are “expensive” computationally (they cause sequence alignment to go from $O(n)$ to $O(n^2)$), so BLAST waits until the last possible moment to incorporate them.

Step 1: Divide A Given Read into *k*-mers

Note: *k* is one of the additional parameters that we mentioned.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Step 1: Divide A Given Read into *k*-mers

Note: *k* is one of the additional parameters that we mentioned.

The number of alignments that BLAST produces for a given collection of parameters is called its **sensitivity**.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Step 1: Divide A Given Read into *k*-mers

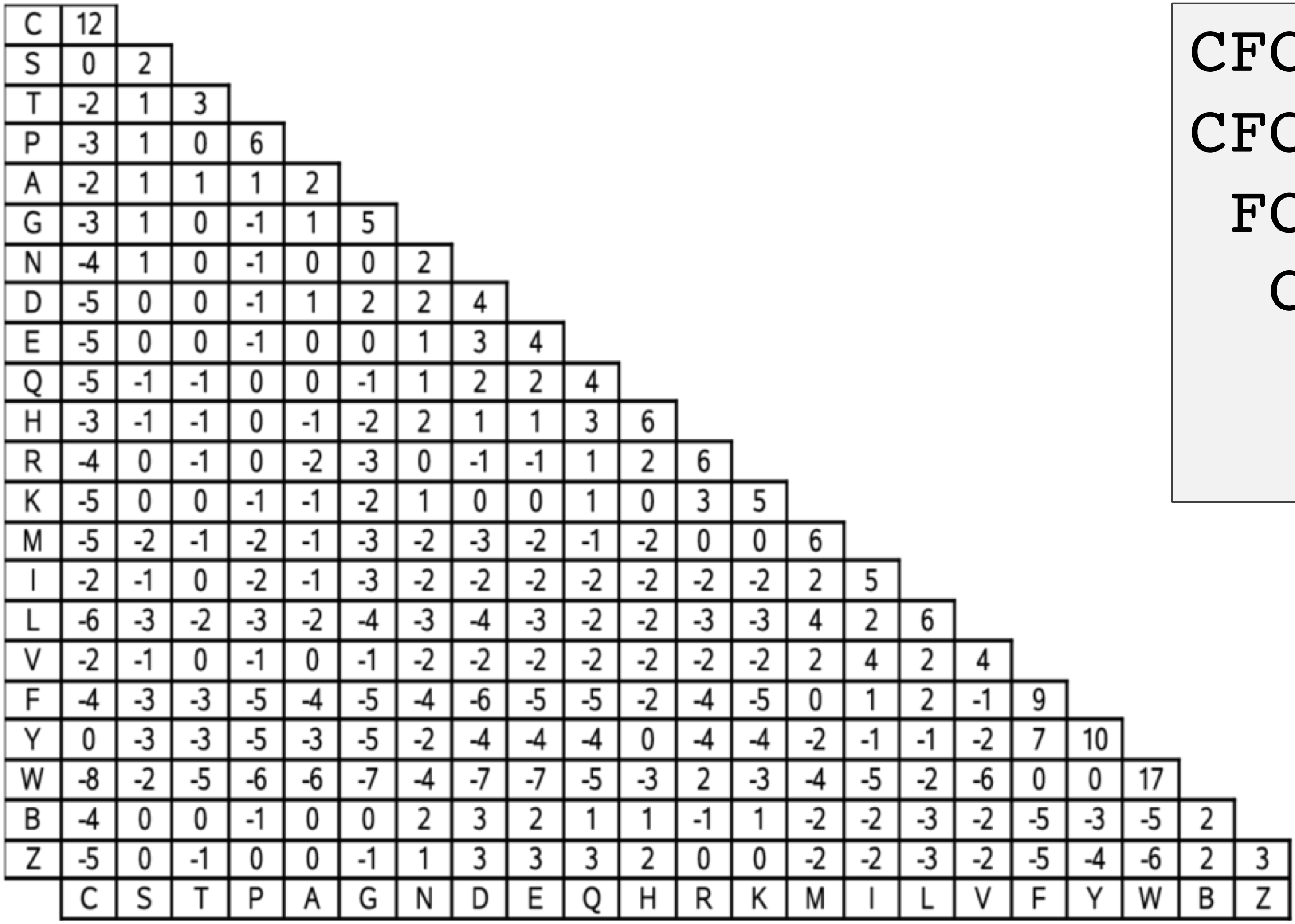
Note: *k* is one of the additional parameters that we mentioned.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

The number of alignments that BLAST produces for a given collection of parameters is called its **sensitivity**.

Checkpoint: What do you think happens to the sensitivity of BLAST as *k* increases/decreases?

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?



C	12																					
S	0	2																				
T	-2	1	3																			
P	-3	1	0	6																		
A	-2	1	1	1	2																	
G	-3	1	0	-1	1	5																
N	-4	1	0	-1	0	0	2															
D	-5	0	0	-1	1	2	2	4														
E	-5	0	0	-1	0	0	1	3	4													
Q	-5	-1	-1	0	0	-1	1	2	2	4												
H	-3	-1	-1	0	-1	-2	2	1	1	3	6											
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6										
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6						
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9				
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
B	-4	0	0	-1	0	0	2	3	2	1	1	-1	1	-2	-2	-3	-2	-5	-3	-5	2	
Z	-5	0	-1	0	0	-1	1	3	3	3	2	0	0	-2	-2	-3	-2	-5	-4	-6	2	3
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z

PAM$_{250}$

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?

C	12																					
S	0	2																				
T	-2	1	3																			
P	-3	1	0	6																		
A	-2	1	1	1	2																	
G	-3	1	0	-1	1	5																
N	-4	1	0	-1	0	0	2															
D	-5	0	0	-1	1	2	2	4														
E	-5	0	0	-1	0	0	1	3	4													
Q	-5	-1	-1	0	0	-1	1	2	2	4												
H	-3	-1	-1	0	-1	-2	2	1	1	3	6											
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6										
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6						
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9				
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
B	-4	0	0	-1	0	0	2	3	2	1	1	-1	1	-2	-2	-3	-2	-5	-3	-5	2	
Z	-5	0	-1	0	0	-1	1	3	3	3	2	0	0	-2	-2	-3	-2	-5	-4	-6	2	3
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	B	Z

PAM$_{250}$

Exercise: Five 3-mers score > 23 against `CFC`. What are they?

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?

High Scoring *k*-mers Problem:

- **Input:** an amino acid *k*-mer *x*, a scoring matrix *Score*, and a threshold value *T*.
- **Output:** All amino acid *k*-mers *y* such that $Score(x_1, y_1) + Score(x_2, y_2) + \ldots + Score(x_k, y_k)$ is $> T$.

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?

High Scoring *k*-mers Problem:

- **Input:** an amino acid *k*-mer *x*, a scoring matrix *Score*, and a threshold value *T*.
- **Output:** All amino acid *k*-mers *y* such that $Score(x_1, y_1) + Score(x_2, y_2) + \ldots + Score(x_k, y_k)$ is $> T$.

Note: The *k*-mers produced are the *k*-mers that we should look for in the database that match well against *x* – so we're back to exact pattern matching!

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?

High Scoring *k*-mers Problem:

- **Input:** an amino acid *k*-mer *x*, a scoring matrix *Score*, and a threshold value *T*.
- **Output:** All amino acid *k*-mers *y* such that $Score(x_1, y_1) + Score(x_2, y_2) + \ldots + Score(x_k, y_k)$ is $> T$.

Checkpoint: Note that *Score* and *T* add more parameters to BLAST. What effect do you think increasing/decreasing *T* has on sensitivity?

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?

High Scoring *k*-mers Problem:

- **Input:** an amino acid *k*-mer *x*, a scoring matrix *Score*, and a threshold value *T*.
- **Output:** All amino acid *k*-mers *y* such that $Score(x_1, y_1) + Score(x_2, y_2) + \ldots + Score(x_k, y_k)$ is $> T$.

Exercise: Can you find an efficient algorithm solving this problem?

Step 2: For Each *k*-mer *x*, what other *k*-mers score well against it?

High Scoring *k*-mers Problem:

- **Input:** an amino acid *k*-mer *x*, a scoring matrix *Score*, and a threshold value *T*.
- **Output:** All amino acid *k*-mers *y* such that $Score(x_1, y_1) + Score(x_2, y_2) + \ldots + Score(x_k, y_k)$ is $> T$.

Note: Once we solve this problem, we organize all the resulting *k*-mers into a trie.

Step 3: Search for High-Scoring *k*-mers in Database

In summary, each read produces a collection of *k*-mers, and each *k*-mer produces a trie of high-scoring *k*-mers.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Step 3: Search for High-Scoring *k*-mers in Database

In summary, each read produces a collection of *k*-mers, and each *k*-mer produces a trie of high-scoring *k*-mers.

We then "slide" these tries across the database, looking for exact matches. Any exact matches become our **seeds**.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Step 3: Search for High-Scoring *k*-mers in Database

In summary, each read produces a collection of *k*-mers, and each *k*-mer produces a trie of high-scoring *k*-mers.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

We then "slide" these tries across the database, looking for exact matches. Any exact matches become our **seeds**.

(Recall from read mapping slides how pattern matching works with the trie.)

Step 3: Search for High-Scoring *k*-mers in Database

In summary, each read produces a collection of *k*-mers, and each *k*-mer produces a trie of high-scoring *k*-mers.

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

We then "slide" these tries across the database, looking for exact matches. Any exact matches become our **seeds**.

Aho-Corasick algorithm for trie pattern matching (recitation) is $O(|Database| + |Tries| + m)$; m = # of matches found, $|Tries|$ = # letters in all tries.

Step 4: Initial Extension into Maximal Segment Pairs

```
Database   ...CICDVQ...

Query           CDI
```

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Note: Just because a *k*-mer has a good score doesn't mean that it can't be *extended* into a longer match with the read.

Step 4: Initial Extension into Maximal Segment Pairs

```
Database    ...CICDVQ...

Query          CFCDIQ
```

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Note: Just because a *k*-mer has a good score doesn't mean that it can't be *extended* into a longer match with the read.

Step 4: Initial Extension into Maximal Segment Pairs

```
Database    ...CICDVQ...

Query          CFCDIQ
```

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

We extend each seed as far to the left and right as we can until the score w/r/t the database stops increasing. The result is a pair of substrings of the query and database called a **maximal segment pair**.

Step 4: Initial Extension into Maximal Segment Pairs

Database ...CICDVQ...

Query CFCDIQ

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

(Think of MSPs as high-scoring local alignments of query against database without gaps.)

Step 4: Initial Extension into Maximal Segment Pairs

```
Database    ...CICDVQ...

Query          CFCDIQ
```

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Checkpoint: Recall … what does "maximal" mean mathematically?

Step 4: Initial Extension into Maximal Segment Pairs

Database ...CICDVQ...

Query CFCDIQ

```
CFCDIQL
CFC
 FCD
  CDI
   DIQ
    IQL
```

Answer: A “local” maximum. In other words, the same query may correspond to many MSPs (or zero MSPs) throughout the database.

Step 5: Trim Maximal Segment Pairs

Yet another threshold parameter *S* is now used; we throw away any MSPs whose score (with respect to the scoring matrix) is < *S*.

Step 5: Trim Maximal Segment Pairs

Yet another threshold parameter S is now used; we throw away any MSPs whose score (with respect to the scoring matrix) is $< S$.

Checkpoint: Does increasing S increase or decrease the sensitivity of BLAST?

Step 6: How "Good" is Each MSP?

Checkpoint: Say we have one MSP of length 12 with score 56 and another MSP of length 9 with score 45. Which one is better?

Step 6: How “Good” is Each MSP?

Checkpoint: Say we have one MSP of length 12 with score 56 and another MSP of length 9 with score 45. Which one is better?

Answer: The way to answer this question is to not say “which is better” but to say “which would be less likely in a random environment?”

Step 6: How “Good” is Each MSP?

p-value: The probability that an event we observe would have occurred in a random reconstruction of whatever we are working with. (This is a very bad statistical definition.)

Step 6: How "Good" is Each MSP?

p-value: The probability that an event we observe would have occurred in a random reconstruction of whatever we are working with. (This is a very bad statistical definition.)

In this case, the p-value we want to compute is the probability Pr(s >= Q) that we would observe an MSP of score s at least some threshold Q in a *random* database.

Step 6: How "Good" is Each MSP?

p-value: The probability that an event we observe would have occurred in a random reconstruction of whatever we are working with. (This is a very bad statistical definition.)

In this case, the p-value we want to compute is the probability Pr($s >= Q$) that we would observe an MSP of score s at least some threshold Q in a *random* database.

(Constructing the database and computing this p-value would take two lectures of statistics.)

Step 6: How "Good" is Each MSP?

p-value: The probability that an event we observe would have occurred in a random reconstruction of whatever we are working with. (This is a very bad statistical definition.)

Checkpoint: Are we hoping for MSPs with lower p-values or higher p-values?

Step 7: Combine Nearby MSPs

Database ...XXAAAAAAAXXXCCCCCCXX...

Query ZZAAAAAAAZZZCCCCCCZZ

MSP1 MSP2

Step 7: Combine Nearby MSPs

```
Database   ...XXAAAAAAAXXXCCCCCCXX...

Query         ZZAAAAAAAZZZCCCCCCZZ
```

Combined MSP

We will merge two nearby MSPs into one MSP if their combined p-value is still significant (i.e., below yet another threshold parameter).

Step 8: Align Step in BLAST

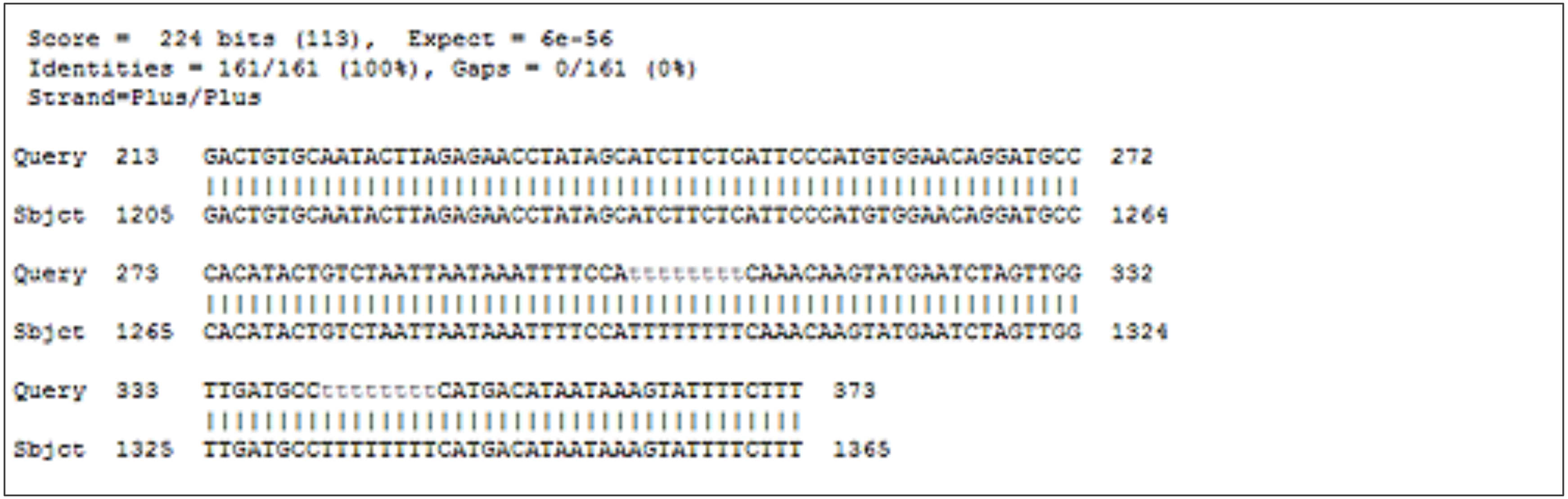
We *still* haven’t added any gaps to our alignments, but we have trimmed away everything but MSP regions that we know are very interesting.

Step 8: Align Step in BLAST

We *still* haven't added any gaps to our alignments, but we have trimmed away everything but MSP regions that we know are very interesting.

We now can perform a (Smith-Waterman) alignment of every MSP we found and report the resulting alignment (with p-value).

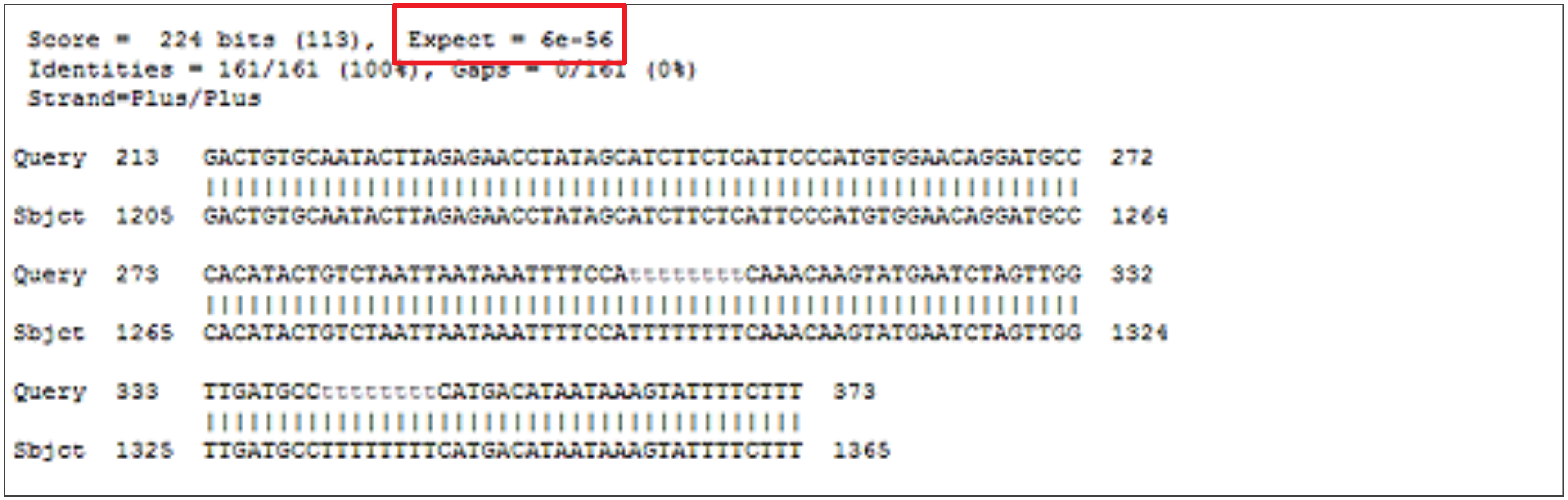
Sample BLAST Output



```
 Score =  224 bits (113),  Expect = 6e-56
 Identities = 161/161 (100%), Gaps = 0/161 (0%)
 Strand=Plus/Plus

Query  213   GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  272
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1205  GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  1264

Query  273   CACATACTGTCTAATTAATAAATTTTCCAttttttttCAAACAAGTATGAATCTAGTTGG  332
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1265  CACATACTGTCTAATTAATAAATTTTCCATTTTTTTTCAAACAAGTATGAATCTAGTTGG  1324

Query  333   TTGATGCCttttttttCATGACATAATAAAGTATTTTCTTT  373
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1325  TTGATGCCTTTTTTTTCATGACATAATAAAGTATTTTCTTT  1365
```

© 2018 Phillip Compeau

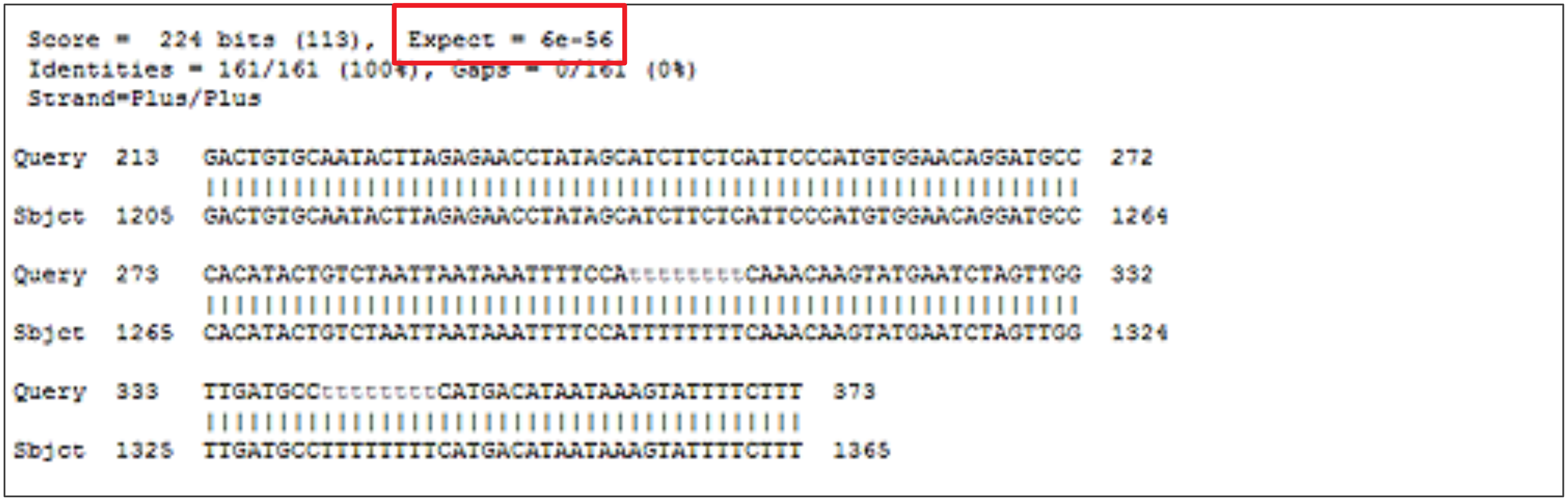
Sample BLAST Output



```
 Score =  224 bits (113),  Expect = 6e-56
 Identities = 161/161 (100%), Gaps = 0/161 (0%)
 Strand=Plus/Plus

Query  213   GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  272
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1205  GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  1264

Query  273   CACATACTGTCTAATTAATAAATTTTCCAttttttttCAAACAAGTATGAATCTAGTTGG  332
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1265  CACATACTGTCTAATTAATAAATTTTCCATTTTTTTTCAAACAAGTATGAATCTAGTTGG  1324

Query  333   TTGATGCCttttttttCATGACATAATAAAGTATTTTCTTT  373
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1325  TTGATGCCTTTTTTTTCATGACATAATAAAGTATTTTCTTT  1365
```

In practice, BLAST returns an **E-value:** the expected number of hits of comparable score in a random database.

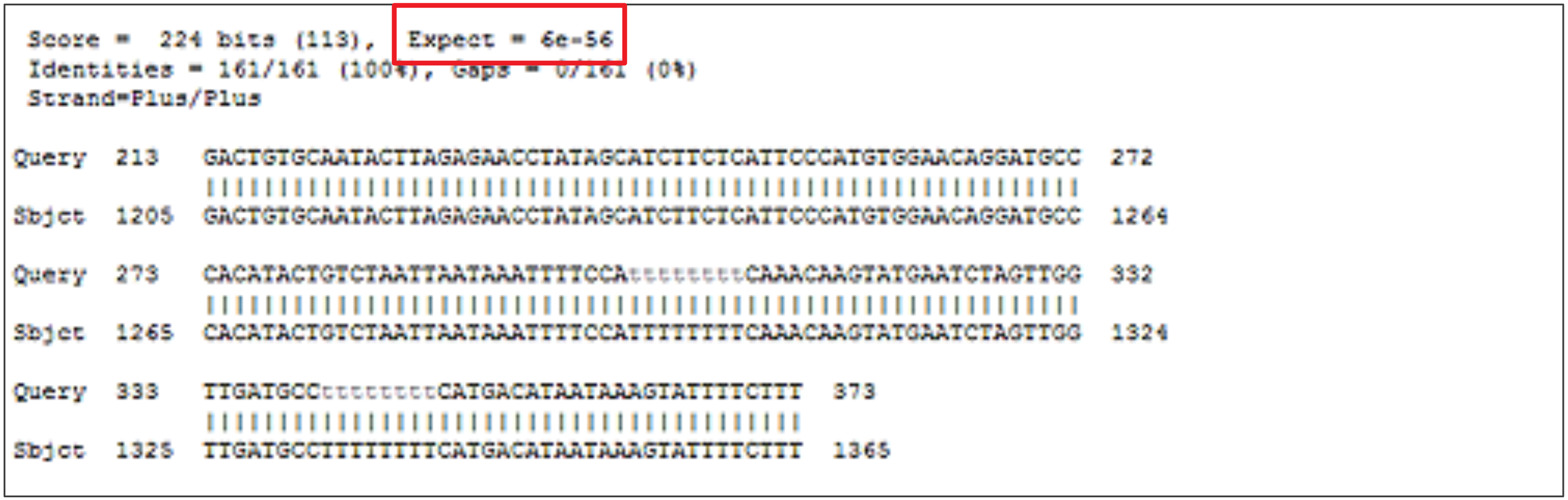
Sample BLAST Output



```
 Score =  224 bits (113),  Expect = 6e-56
 Identities = 161/161 (100%), Gaps = 0/161 (0%)
 Strand=Plus/Plus

Query  213   GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  272
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1205  GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  1264

Query  273   CACATACTGTCTAATTAATAAATTTTCCAttttttttCAAACAAGTATGAATCTAGTTGG  332
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1265  CACATACTGTCTAATTAATAAATTTTCCATTTTTTTTCAAACAAGTATGAATCTAGTTGG  1324

Query  333   TTGATGCCttttttttCATGACATAATAAAGTATTTTCTTT  373
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1325  TTGATGCCTTTTTTTTCATGACATAATAAAGTATTTTCTTT  1365
```

Note: Very small p-values are approximately equal to their corresponding E-values.

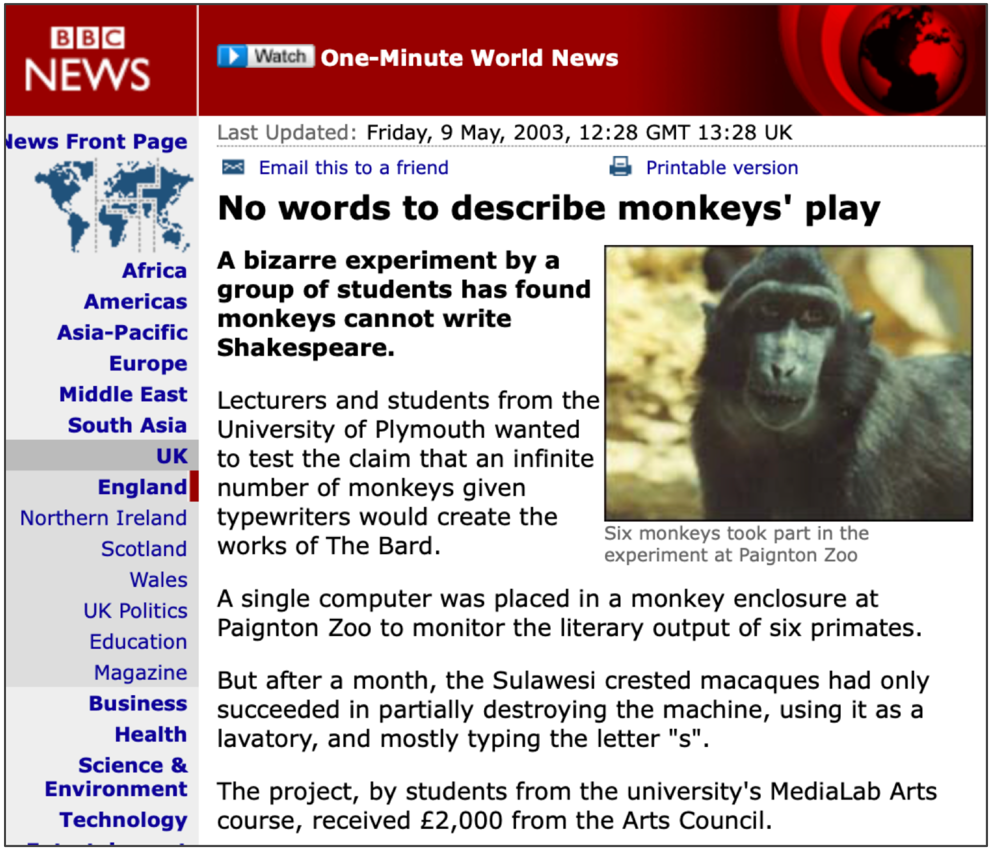
Sample BLAST Output



```
 Score =  224 bits (113),  Expect = 6e-56
 Identities = 161/161 (100%), Gaps = 0/161 (0%)
 Strand=Plus/Plus

Query  213   GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  272
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1205  GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  1264

Query  273   CACATACTGTCTAATTAATAAATTTTCCAttttttttCAAACAAGTATGAATCTAGTTGG  332
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1265  CACATACTGTCTAATTAATAAATTTTCCATTTTTTTTCAAACAAGTATGAATCTAGTTGG  1324

Query  333   TTGATGCCttttttttCATGACATAATAAAGTATTTTCTTT  373
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1325  TTGATGCCTTTTTTTTCATGACATAATAAAGTATTTTCTTT  1365
```

Checkpoint: We report all E-values below some threshold; it is just one more parameter affecting sensitivity.

Be Careful with P-Values!



Last Updated: Friday, 9 May, 2003, 12:28 GMT 13:28 UK

No words to describe monkeys' play

A bizarre experiment by a group of students has found monkeys cannot write Shakespeare.

Lecturers and students from the University of Plymouth wanted to test the claim that an infinite number of monkeys given typewriters would create the works of The Bard.

Six monkeys took part in the experiment at Paignton Zoo

A single computer was placed in a monkey enclosure at Paignton Zoo to monitor the literary output of six primates.

But after a month, the Sulawesi crested macaques had only succeeded in partially destroying the machine, using it as a lavatory, and mostly typing the letter "s".

The project, by students from the university's MediaLab Arts course, received £2,000 from the Arts Council.

Overview of Metagenomics Process

1. Find six reading frames of every read.
2. Apply BLAST to resulting query database to produce alignments.
3. Use alignments to quantify certainty that a metagenome came from different species (the correct term is “organizational taxonomic unit”).

Overview of Metagenomics Process

1. Find six reading frames of every read.
2. Apply BLAST to resulting query database to produce alignments.
3. Use alignments to quantify certainty that a metagenome came from different species (the correct term is "organizational taxonomic unit").

Checkpoint: We haven't discussed #3 … how might we do it?

Why We Need Fast Heuristics Like BLAST for "Big Data"

"But why do we care about a 30 year-old algorithm applied to metagenomic read binning? Is it still relevant?"

Why We Need Fast Heuristics Like BLAST for "Big Data"

"But why do we care about a 30 year-old algorithm applied to metagenomic read binning? Is it still relevant?"

Short answer: computers are getting faster, but we're also sequencing more organisms, so the database we're consulting is growing too. (This is a common phenomenon across fields.)

A Longer Answer

Say the runtime of an old and new algorithm are T_{old} and T_{new} , respectively. The speedup provided by T_{new} is the ratio T_{old} / T_{new} .

A Longer Answer

Say the runtime of an old and new algorithm are T_{old} and T_{new}, respectively. The speedup provided by T_{new} is the ratio $T_{\text{old}} / T_{\text{new}}$.

Exercise: Say T_{old} is n^2 and T_{new} is 16*n*lg(n). What is the speedup for an arbitrary n? What happens for $n = 2^5$, 2^{10}, 2^{20}, and 2^{40}?

A Longer Answer

Say the runtime of an old and new algorithm are T_{old} and T_{new}, respectively. The speedup provided by T_{new} is the ratio $T_{\text{old}} / T_{\text{new}}$.

Exercise: Say T_{old} is n^2 and T_{new} is 16*n*lg(n). What is the speedup for an arbitrary n? What happens for $n = 2^5$, 2^{10}, 2^{20}, and 2^{40}?

Key Point 1: A slight improvement to algorithm runtime can compound for larger datasets.

A Longer Answer

Say the runtime of an old and new algorithm are T_{old} and T_{new}, respectively. The speedup provided by T_{new} is the ratio $T_{\text{old}} / T_{\text{new}}$.

Exercise: Say T_{old} is n^2 and T_{new} is 16*n*lg(n). What is the speedup for an arbitrary n? What happens for $n = 2^5$, 2^{10}, 2^{20}, and 2^{40}?

Key Point 2: Optimizing read aligners is still an active area of research for this reason!