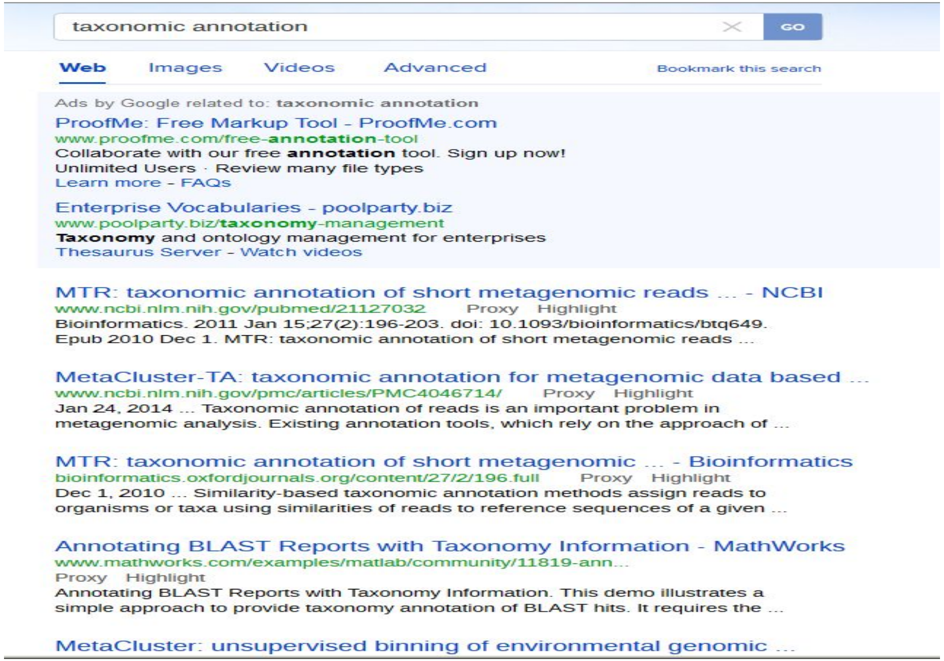
Database searching- Sequence alignment based methods

Nidhi Shah
M^3 Workshop
January 10, 2019

Google: "taxonomic annotation"



Ads by Google related to: **taxonomic annotation**

ProofMe: Free Markup Tool - ProofMe.com
www.proofme.com/free-**annotation**-tool
Collaborate with our free **annotation** tool. Sign up now!
Unlimited Users · Review many file types
Learn more - FAQs

Enterprise Vocabularies - poolparty.biz
www.poolparty.biz/**taxonomy**-management
Taxonomy and ontology management for enterprises
Thesaurus Server - Watch videos

MTR: taxonomic annotation of short metagenomic reads ... - NCBI
www.ncbi.nlm.nih.gov/pubmed/21127032 Proxy Highlight
Bioinformatics. 2011 Jan 15;27(2):196-203. doi: 10.1093/bioinformatics/btq649. Epub 2010 Dec 1. MTR: taxonomic annotation of short metagenomic reads ...

MetaCluster-TA: taxonomic annotation for metagenomic data based ...
www.ncbi.nlm.nih.gov/pmc/articles/PMC4046714/ Proxy Highlight
Jan 24, 2014 ... Taxonomic annotation of reads is an important problem in metagenomic analysis. Existing annotation tools, which rely on the approach of ...

MTR: taxonomic annotation of short metagenomic ... - Bioinformatics
bioinformatics.oxfordjournals.org/content/27/2/196.full Proxy Highlight
Dec 1, 2010 ... Similarity-based taxonomic annotation methods assign reads to organisms or taxa using similarities of reads to reference sequences of a given ...

Annotating BLAST Reports with Taxonomy Information - MathWorks
www.mathworks.com/examples/matlab/community/11819-ann...
Proxy Highlight
Annotating BLAST Reports with Taxonomy Information. This demo illustrates a simple approach to provide taxonomy annotation of BLAST hits. It requires the ...

MetaCluster: unsupervised binning of environmental genomic ...

- Database of known pages
- Report all that contain keyword

- Ranking important (which of the thousands is most relevant)

Slides adapted from Mihai Pop

Our "keywords"

```
>F4BT0V001CZSIM rank=0000138 x=1110.0 y=2700.0 length=
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACGT
>F4BT0V001BBJQS rank=0000155 x=424.0 y=1826.0 length=4
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAA
>F4BT0V001EDG35 rank=0000182 x=1676.0 y=2387.0 length=
ACTGACTGCATGCTGCCTCCCGTAGGAGTCGCCGTCCTCGACNC
>F4BT0V001D2HQQ rank=0000196 x=1551.0 y=1984.0 length=
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCGTCCCTCGAC
>F4BT0V001CM392 rank=0000206 x=966.0 y=1240.0 length=8
AANCAGCTCTCATGCTCGCCCTGACTTGGCATGTGTTAAGCCTGTAGGCTAGCG
>F4BT0V001EIMFX rank=0000250 x=1735.0 y=907.0 length=46
ACTGACTGCATGCTGCCTCCCGTAGGAGTGTCGCGCCATCAGACTG
>F4BT0V001ENDKR rank=0000262 x=1789.0 y=1513.0 length=56
GACACTGTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG
>F4BT0V001D91MI rank=0000288 x=1637.0 y=2088.0 length=56
ACTGCTCTCATGCTGCCTCCCGTAGGAGTGCCTCCCTGAGCCAGGATCAAACTCTG
>F4BT0V001D0Y5G rank=0000341 x=1534.0 y=866.0 length=75
GTCTGTGACATGCTGCCTCCCGTAGGAGTCTACACAAGTTGTGGCCCAGAACCACTGAGCCAGGATCAAACTCTG
>F4BT0V001EMLE1 rank=0000365 x=1780.0 y=1883.0 length=84
ACTGACTGCATGCTGCCTCCCGTAGGAGTGCCTCCCTGCGCCATCAATGCTGCATGCTGCTCCCTGAGCCAGGATCAAACTCTG
```

Taxonomic annotation algorithm

- Look through database for sequence
- Report all organisms that contain it
- Rank list by ?? (most relevant to least relevant)
- Handle sequencing errors
- What if the sequence is not in the database (Should handle evolutionary divergence)
  - Can we say anything about the data?

    e.g., google "taxnomic anntoation"

Taxonomic annotation algorithm

Solution:

- Organize the database (taxonomy)

  Kingdom;Phylum;Class;Order;Family;Genus;Species;Strain

- Use search procedure that can generalize from existing knowledge - ??

Taxonomic annotation algorithm

Solution:

- Organize the database (taxonomy)

  Kingdom;Phylum;Class;Order;Family;Genus;Species;Strain

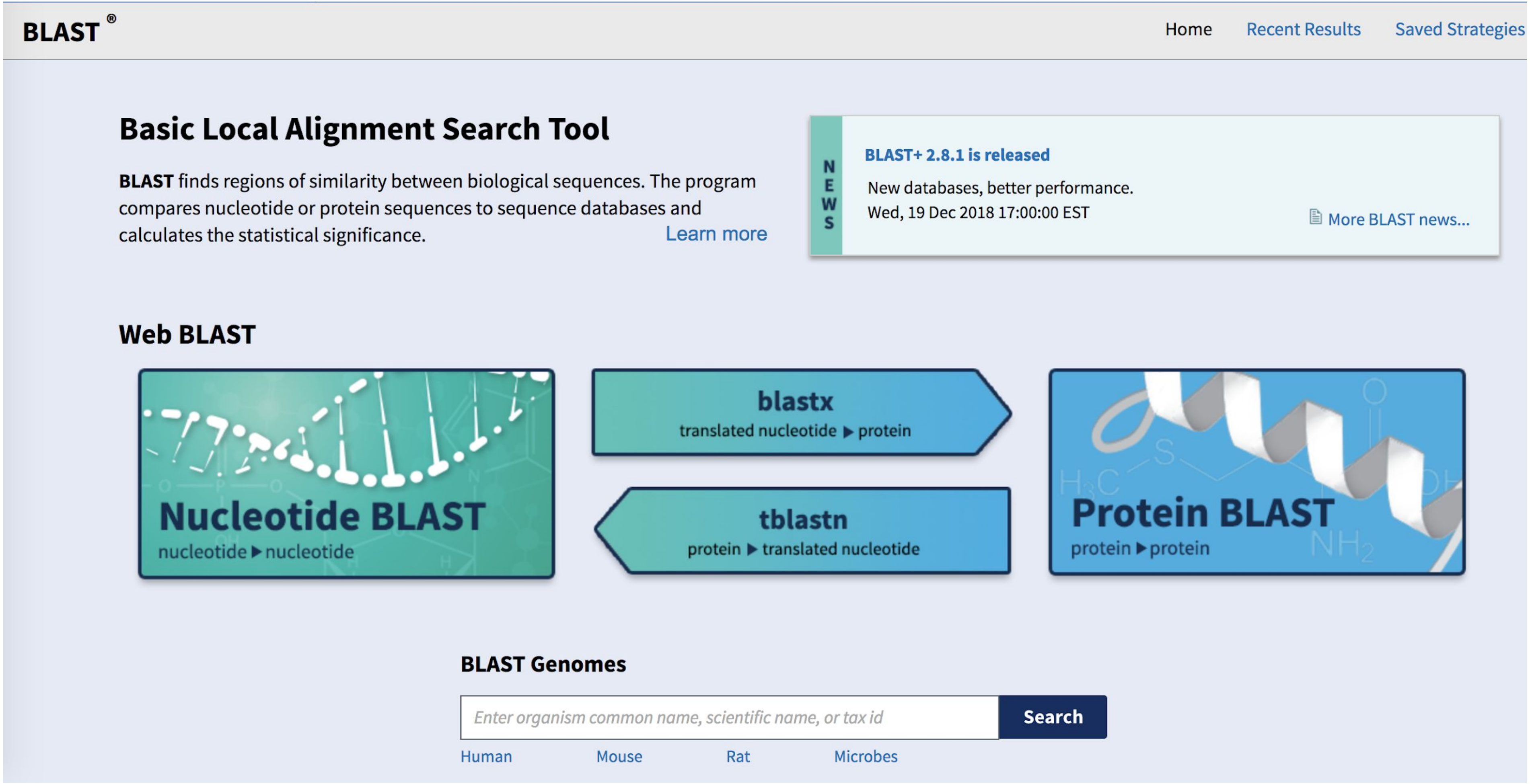
- Use search procedure that can generalize from existing knowledge
  - Sequence similarity search
  - Assume taxonomically related sequences are similar

Similarity search

```
Query  ACCATAG-GCCGTCAGACCTAGACTAGA
DB     AC-ATAGAGCCGTCAGA-CTATACTAGA
```

- Finds exact matches
- Handles sequencing errors
- May handle evolutionary divergence
- May provide statistical guarantees (is this a random hit?) -- can help with ranking results!

- MANY tools exist for doing the search! (e.g. BLAST, Diamond, Usearch, BLAT, etc)
- Differ by
  - assumptions about data
  - similarity cutoff
  - heuristics to speed up search (incl. memory/speed trade-off)



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

BLAST+ 2.8.1 is released

New databases, better performance.
Wed, 19 Dec 2018 17:00:00 EST

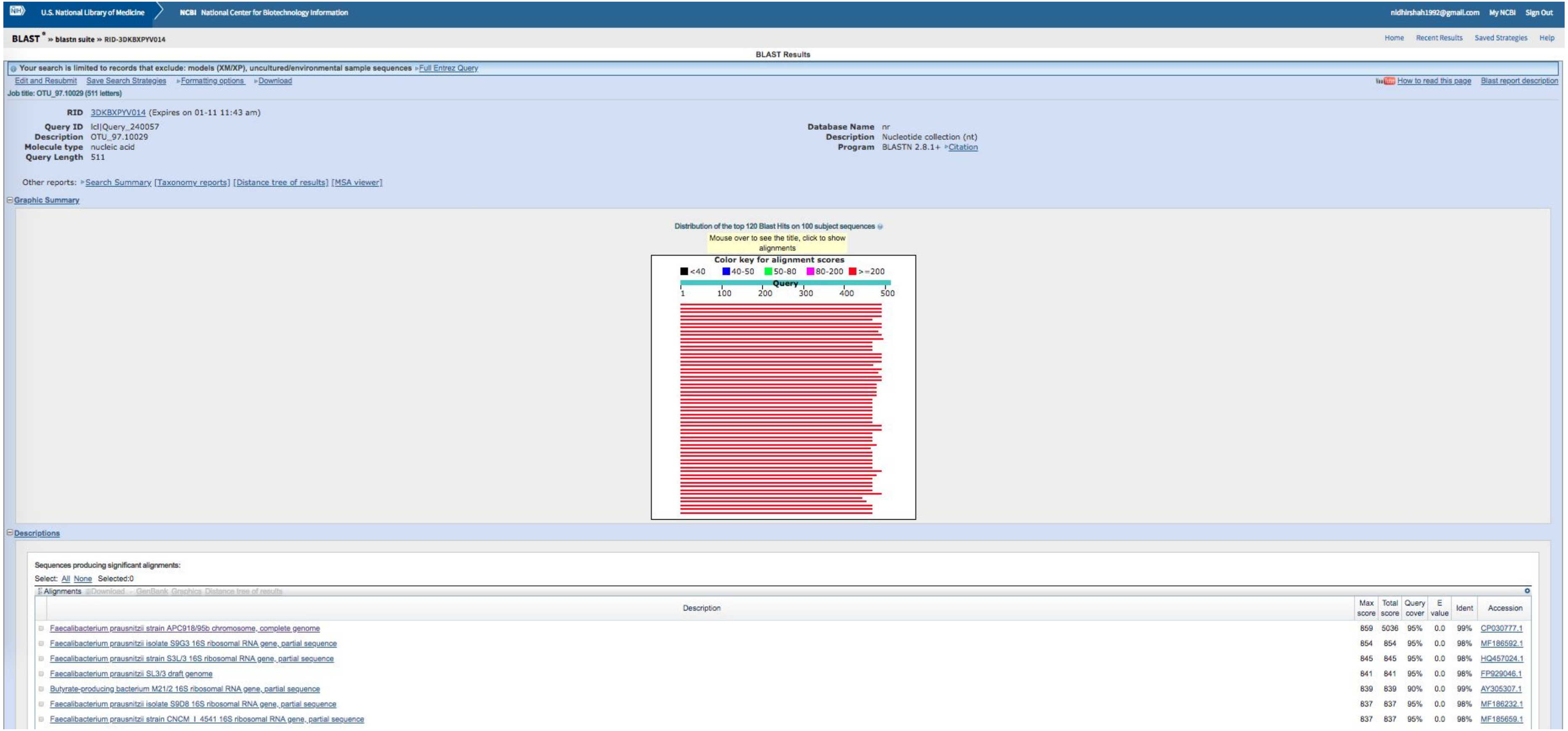
More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id

BLAST web result example



U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
nidhirshah1992@gmail.com
My NCBI
Sign Out
Home
Recent Results
Saved Strategies
Help

BLAST® » blastn suite » RID-3DKBXPYV014

BLAST Results

Your search is limited to records that exclude: models (XM/XP), uncultured/environmental sample sequences ▸Full Entrez Query

Job title: OTU_97.10029 (511 letters)

RID 3DKBXPYV014 (Expires on 01-11 11:43 am)

Query ID lcl|Query_240057
Description OTU_97.10029
Molecule type nucleic acid
Query Length 511

Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.8.1+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

⊟Graphic Summary

Distribution of the top 120 Blast Hits on 100 subject sequences

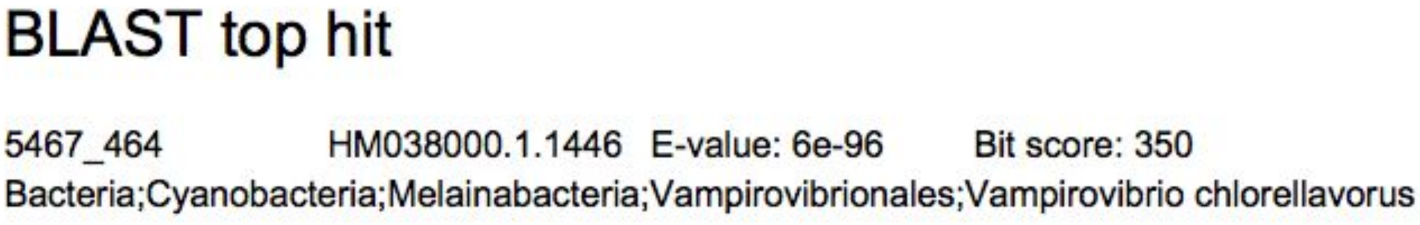
Mouse over to see the title, click to show alignments

⊟Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Faecalibacterium prausnitzii strain APC918/95b chromosome, complete genome	859	5036	95%	0.0	99%	CP030777.1
□	Faecalibacterium prausnitzii isolate S9G3 16S ribosomal RNA gene, partial sequence	854	854	95%	0.0	98%	MF186592.1
□	Faecalibacterium prausnitzii strain S3L/3 16S ribosomal RNA gene, partial sequence	845	845	95%	0.0	98%	HQ457024.1
□	Faecalibacterium prausnitzii SL3/3 draft genome	841	841	95%	0.0	98%	FP929046.1
□	Butyrate-producing bacterium M21/2 16S ribosomal RNA gene, partial sequence	839	839	90%	0.0	99%	AY305307.1
□	Faecalibacterium prausnitzii isolate S9D8 16S ribosomal RNA gene, partial sequence	837	837	95%	0.0	98%	MF186232.1
□	Faecalibacterium prausnitzii strain CNCM_I_4541 16S ribosomal RNA gene, partial sequence	837	837	95%	0.0	98%	MF185659.1



BLAST top hit

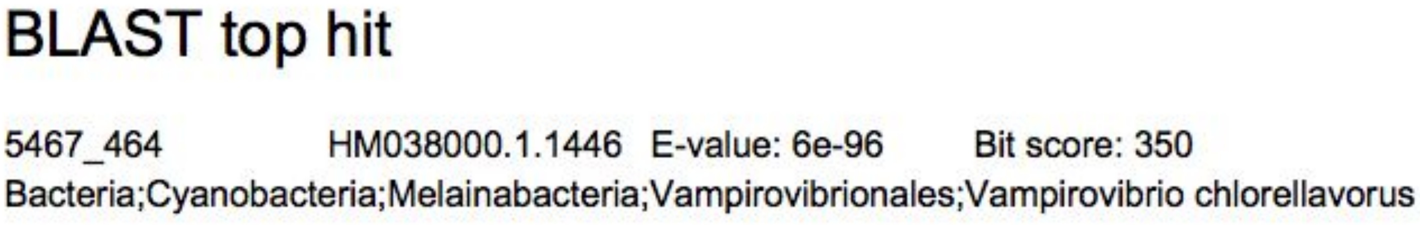
```
5467_464            HM038000.1.1446   E-value: 6e-96        Bit score: 350
Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus
```

Bit score - the “information” contained in the alignment

E-value - how many random alignments one expects for the same bit score

The lower the E-value, the more significant the alignment score of the sequence match

Run BLAST on command line



BLAST top hit

```
5467_464            HM038000.1.1446   E-value: 6e-96        Bit score: 350
Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus
```

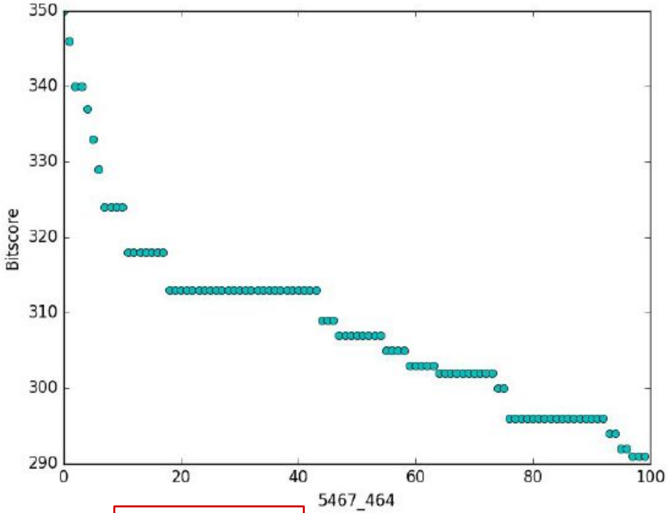
Bit score - the “information” contained in the alignment

E-value - how many random alignments one expects for the same bit score

The lower the E-value, the more significant the alignment score of the sequence match

BLAST...more hits

```
5467_464        HM038000.1.1446 Identity: 80.00%  E-value: 6e-96 Bit score: 350
```



top 100 hits sorted by bit score

Bacteria;Cyanobacteria;Melainabacteria;Vampirovibrionales;Vampirovibrio chlorellavorus

Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas;Brevundimonas mediterranea

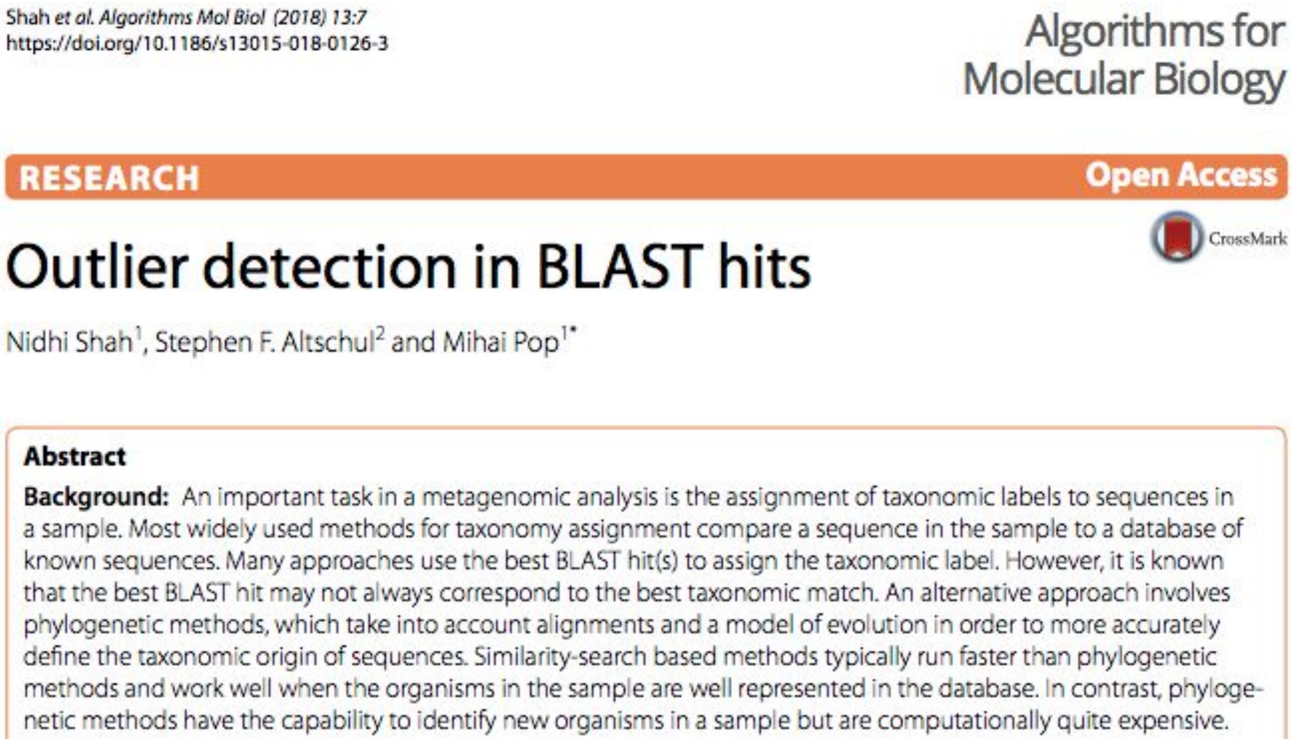
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas; Brevundimonas bacteroides

Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Butyricicoccus;Butyricicoccus pullicaecorum

How to interpret these results?

- Why does this happen??

- Should we use the best hit and transfer annotation to the query sequence?
- Cutoffs on percent identity? Bit score? E-value?
- For e.g. Megan, PhymBl
- What about query coverage?



Shah *et al. Algorithms Mol Biol* (2018) 13:7
https://doi.org/10.1186/s13015-018-0126-3

Algorithms for
Molecular Biology

RESEARCH Open Access

Outlier detection in BLAST hits

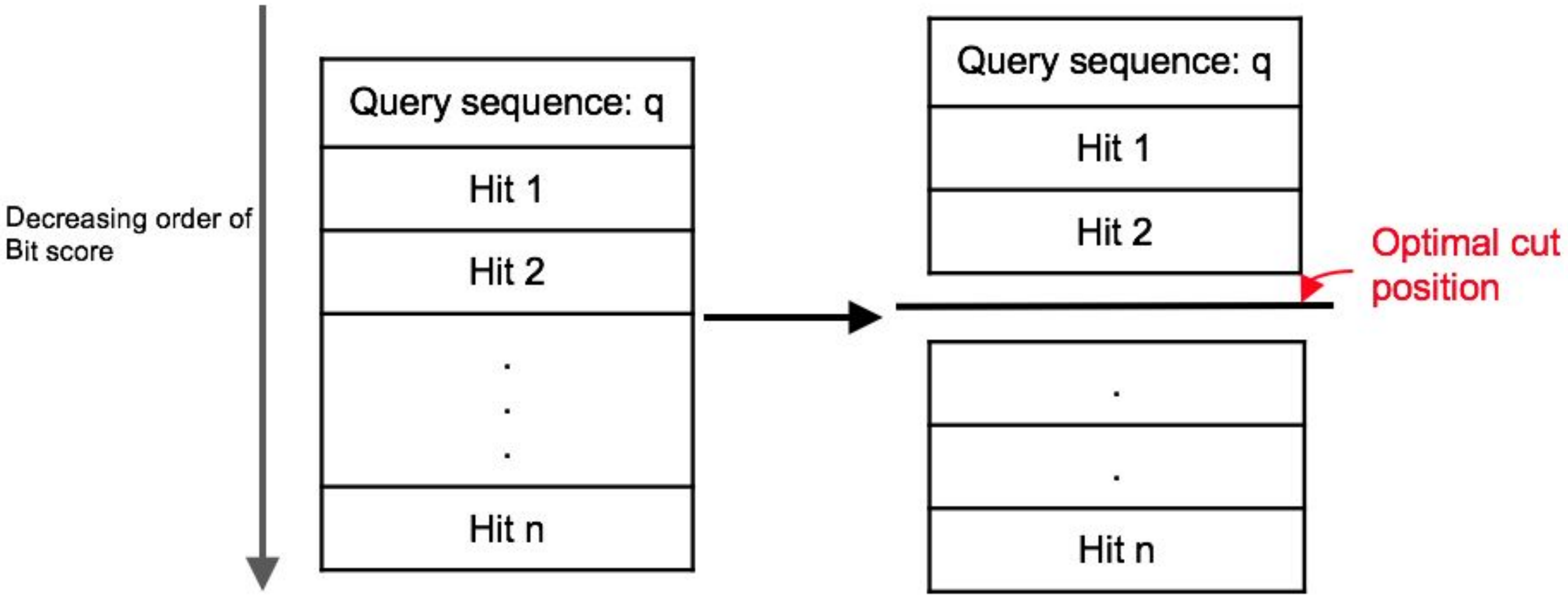
Nidhi Shah[1], Stephen F. Altschul[2] and Mihai Pop[1*]

Abstract

Background: An important task in a metagenomic analysis is the assignment of taxonomic labels to sequences in a sample. Most widely used methods for taxonomy assignment compare a sequence in the sample to a database of known sequences. Many approaches use the best BLAST hit(s) to assign the taxonomic label. However, it is known that the best BLAST hit may not always correspond to the best taxonomic match. An alternative approach involves phylogenetic methods, which take into account alignments and a model of evolution in order to more accurately define the taxonomic origin of sequences. Similarity-search based methods typically run faster than phylogenetic methods and work well when the organisms in the sample are well represented in the database. In contrast, phylogenetic methods have the capability to identify new organisms in a sample but are computationally quite expensive.

Github: https://github.com/shahnidhi/outlier_in_BLAST_hits

Methods



Step 1: Construct a multiple sequence alignment (MSA)

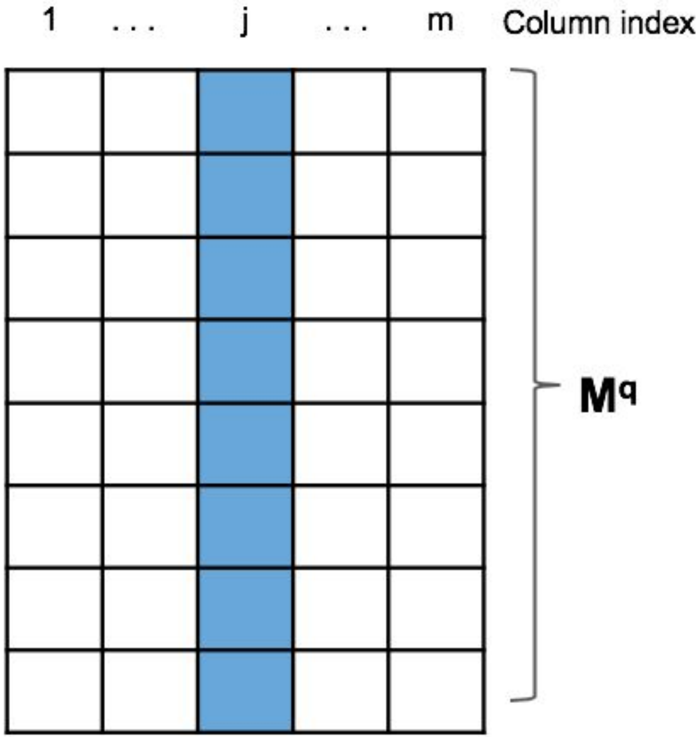
Step 2: Find a cut that splits MSA into two disjoint sets

Methods - Score of a Multiple Sequence Alignment

- Bayesian Integral Log Odds (BILD) score
- Dirichlet distribution for nucleotide probabilities prior

$$L(M_j^q) = \log\left[\frac{\Gamma(\alpha^*)}{\Gamma(\alpha^* + c_j^*)}\prod_{k=1}^{4}\frac{\Gamma(\alpha_k + c_{jk})}{\Gamma(\alpha_k)}\right]$$

- Jeffrey's prior: $\alpha^* = 2$, $\alpha^k = 0.5$
- C^{jk} = total count of base k (A, C, T, G)
- $C^{j*} = C^{jA} + C^{jC} + C^{jT} + C^{jG}$



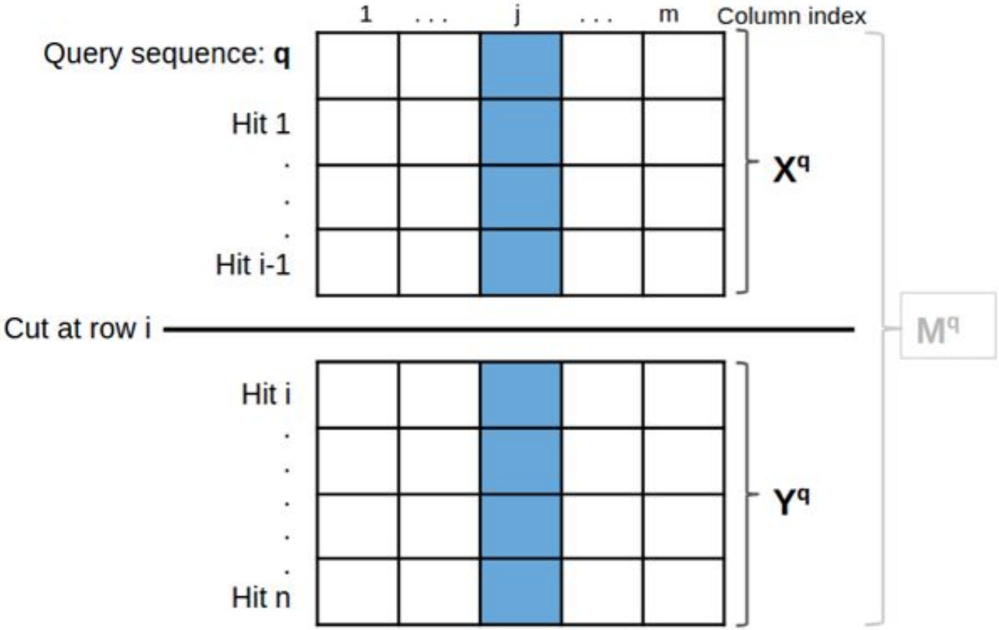
Methods - Scoring a split in an MSA

Score of a cut at row i : V_i^q

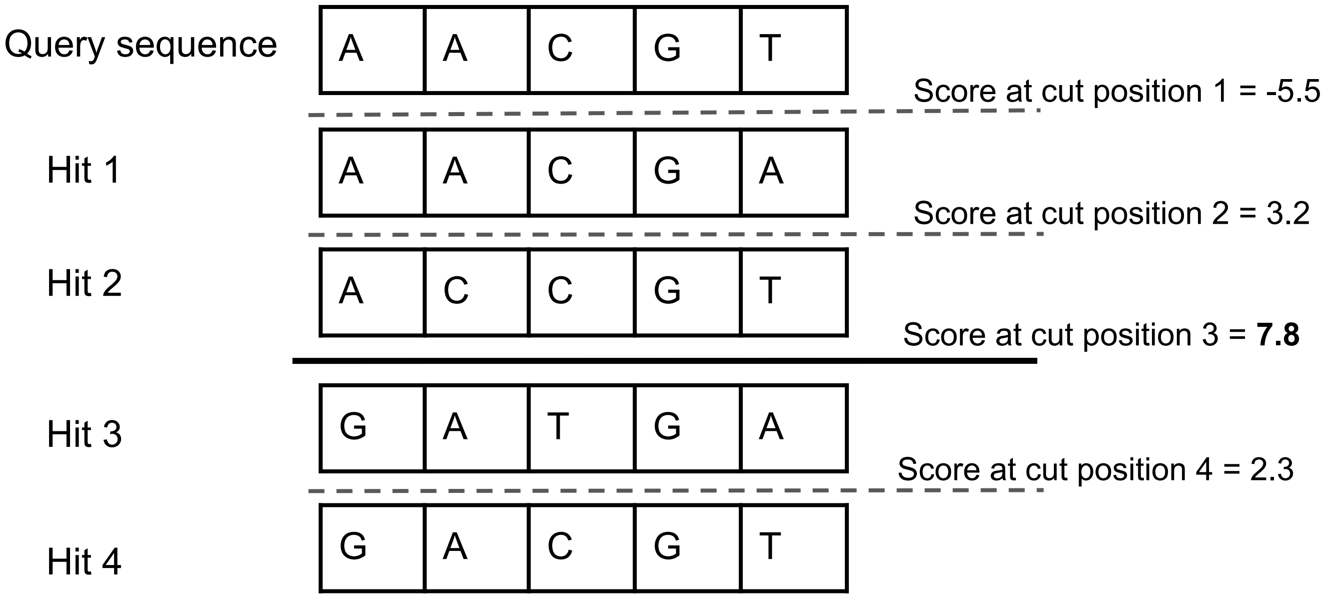
$$V_{ji}^q = L(X_{ji}^q) + L(Y_{ji}^q) - L(M_j^q)$$

$$V_i^q = \sum_{j=1}^{m} e_j^a V_{ji}^q$$

- e_j is the entropy of column j



Methods - Detecting outliers



Query sequence	A	A	C	G	T
Hit 1	A	A	C	G	A
Hit 2	A	C	C	G	T
Hit 3	G	A	T	G	A
Hit 4	G	A	C	G	T

Score at cut position 1 = -5.5

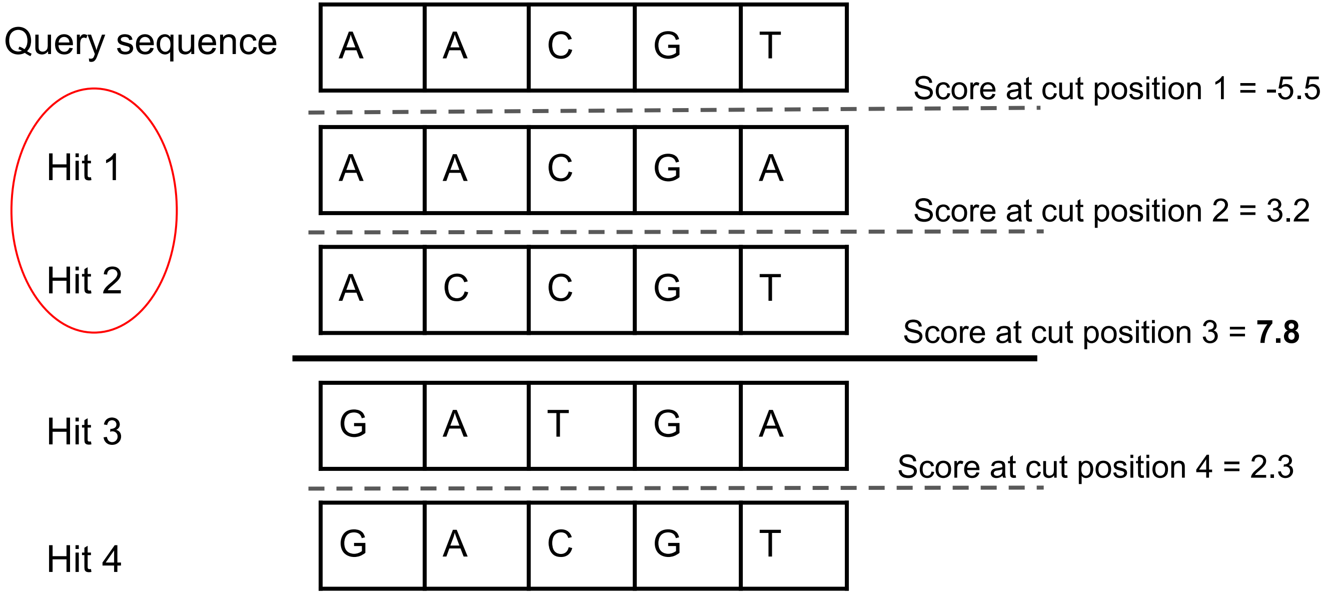
Score at cut position 2 = 3.2

Score at cut position 3 = **7.8**

Score at cut position 4 = 2.3

Note: We are looking for the first positive peak in the score

Methods - Detecting outliers



Query sequence	A	A	C	G	T
Hit 1	A	A	C	G	A
Hit 2	A	C	C	G	T
Hit 3	G	A	T	G	A
Hit 4	G	A	C	G	T

Score at cut position 1 = -5.5

Score at cut position 2 = 3.2

Score at cut position 3 = **7.8**

Score at cut position 4 = 2.3

Note: We are looking for the first positive peak in the score

Run outlier detection pipeline on command line

How to get taxonomic annotation?

- Use Most recent common ancestor of candidate DB sequences (outliers)
- Look at output files

- Can we do something better?

Taxonomic resolution

Family	Genus	Species
Streptococcaceae	Streptococcus	salivarius
Streptococcaceae	Streptococcus	
Streptococcaceae	Streptococcus	salivarius
Streptococcaceae	Streptococcus	
Streptococcaceae	Streptococcus	thermophilus

↑ Lowest Common Ancestor

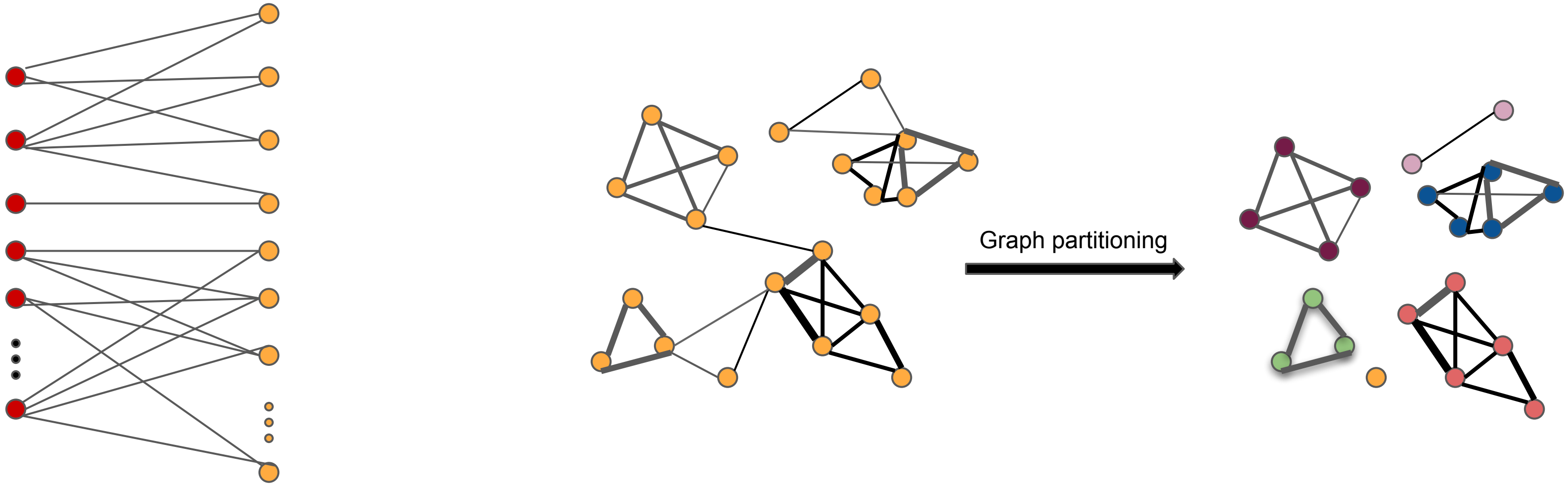
Non-Pathogenic

S. salivarius
S. thermophilus

Pathogenic

S. mitis
S. pyogenes

Creating taxonomic-agnostic clusters



Graph partitioning

Query sequences

Database sequences

“Confusion graph” on database sequences

Database partitions/clusters

Look at partition files

Why certain query sequences have no outliers?

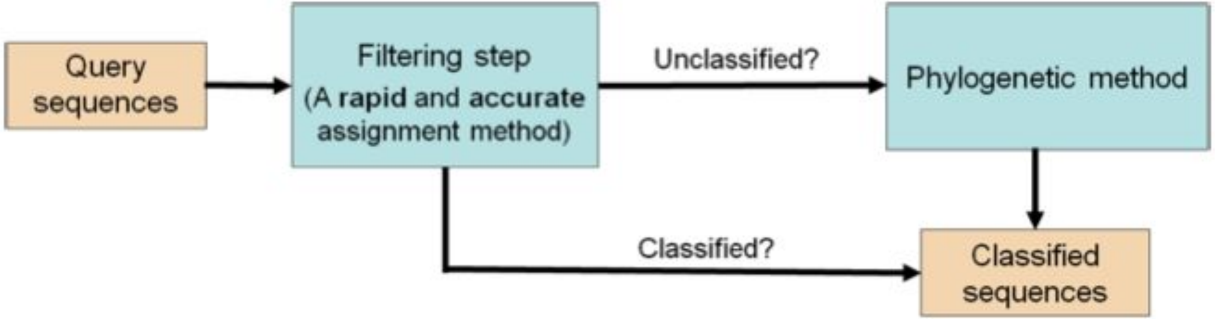
Why certain query sequences have no outliers?

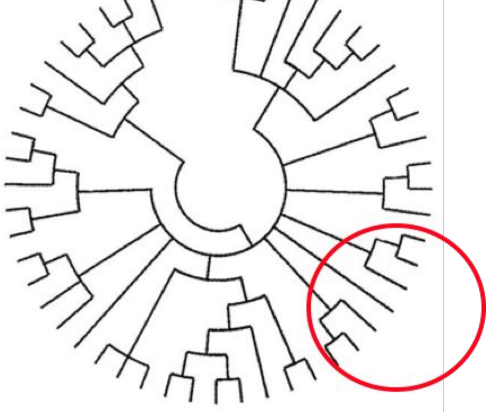
- Query sequence is very different from everything in the database

- The algorithm cannot find any cut position with positive score (peak)

Thanks!

A two-step approach





Suggestions for WMS

Assembly

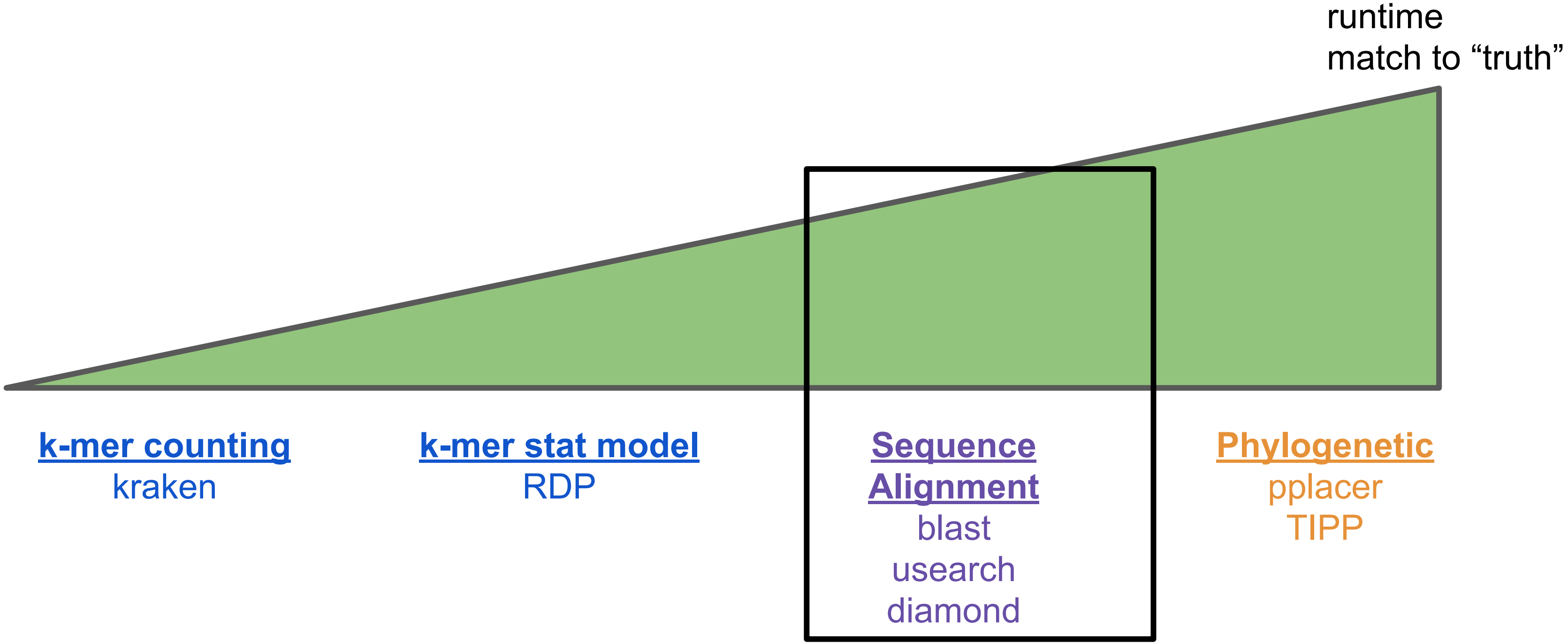
Variant detection

Look at assembly graphs

Read level classification

Abundance profiling - marker gene based methods

Landscape of Taxonomic Approaches



k-mer counting
kraken

k-mer stat model
RDP

Sequence Alignment
blast
usearch
diamond

Phylogenetic
pplacer
TIPP

Assume taxonomically related sequences have similar k-mer frequencies

Assume taxonomically related sequences are similar

Simulate evolution