SWARM

Robust and fast clustering method for amplicon-based studies

Presented by: Michael Khaitov

Course Instructor: Prof Elhanan Borenstein

TAU CS 0368-3116-01: Seminar in Computational Methods in Metagenomics and Microbiome Research

ביה"ס למדעי המחשב ע"ש בלווטניק
הפקולטה למדעים מדויקים
ע"ש ריימונד ובברלי סאקלר
אוניברסיטת תל אביב

Background

Clustering Methods

SWARM

SWARM v2

Summary & Discussion points

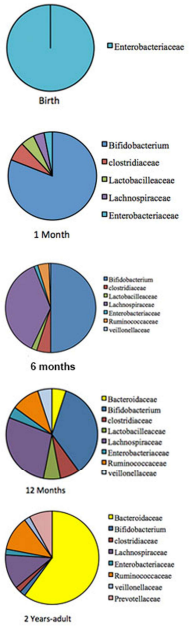
Background

cttcatgataccacatgatact
cagtttacacacagatacaca
cagatacaattcagatacaat
cagatagagacagatagaga
gatagattttcgattagattttc
cttcatgataccacatgatact
cagtttacacacagatacaca
cagatacaattcagatacaat
cagatagagacagatagaga
gatatgattttcgatagattttc
cttcatgataccacatgatact
cagtttacacacagatacaca
cagatacaattcagatacaat
cagatagagacagatagaga
gatagattttcgatatgattttc

cacatgatgg
cacgtgatac
cacgagatac

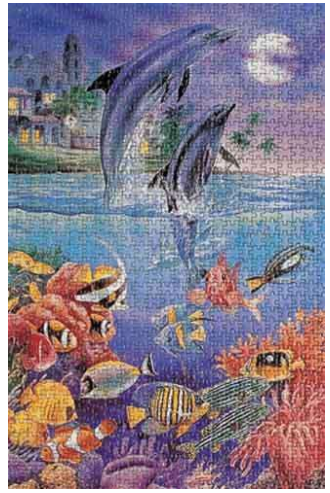
gatatattttc
gatagataat
gagggatttc

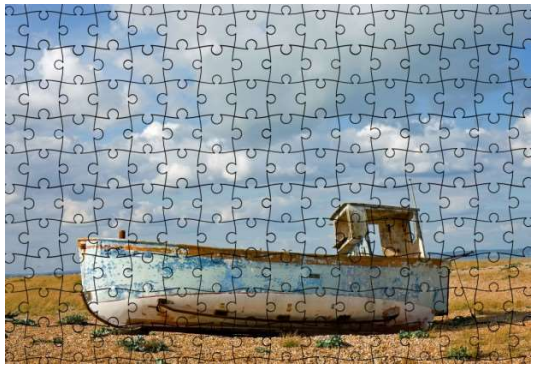
atttgtattttct
aatagataagt
aagggattgtc



Background

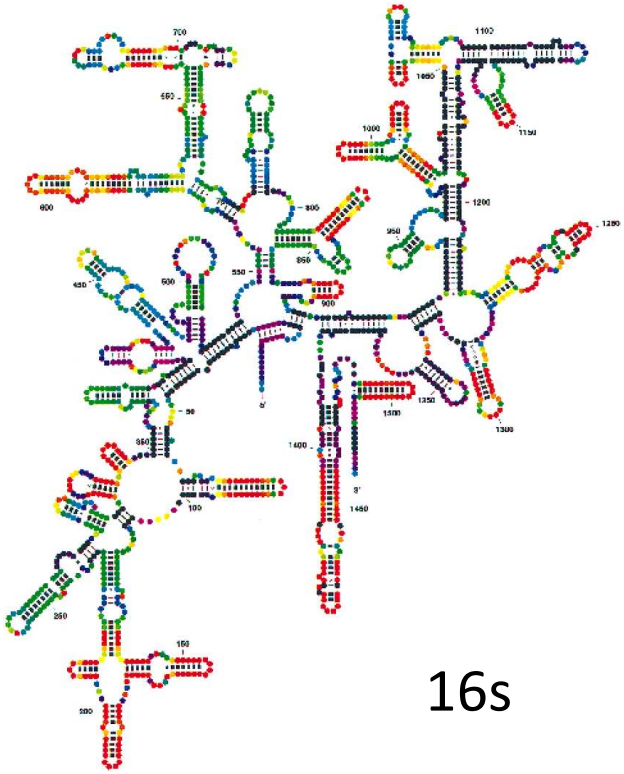
Targeted Sequencing





Background

Targeted Sequencing



16s

Clustering Methods

Clustering Methods

Two major types:

1. reference based

2. de-novo

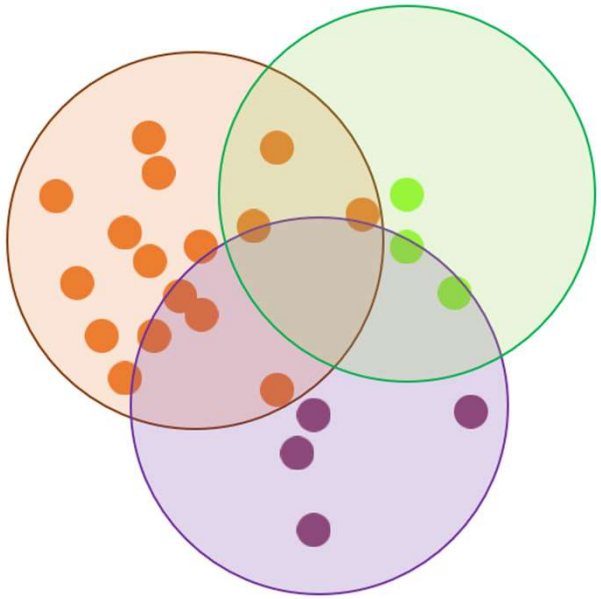
Greedy Clustering Methods

Two main properties:

1. Global threshold t (usually 97%)

2. One-shot

Greedy Clustering Methods



Greedy Clustering Methods

Two major problems with the greedy approach:

1. Order matters

2. Fixed global threshold

SWARM

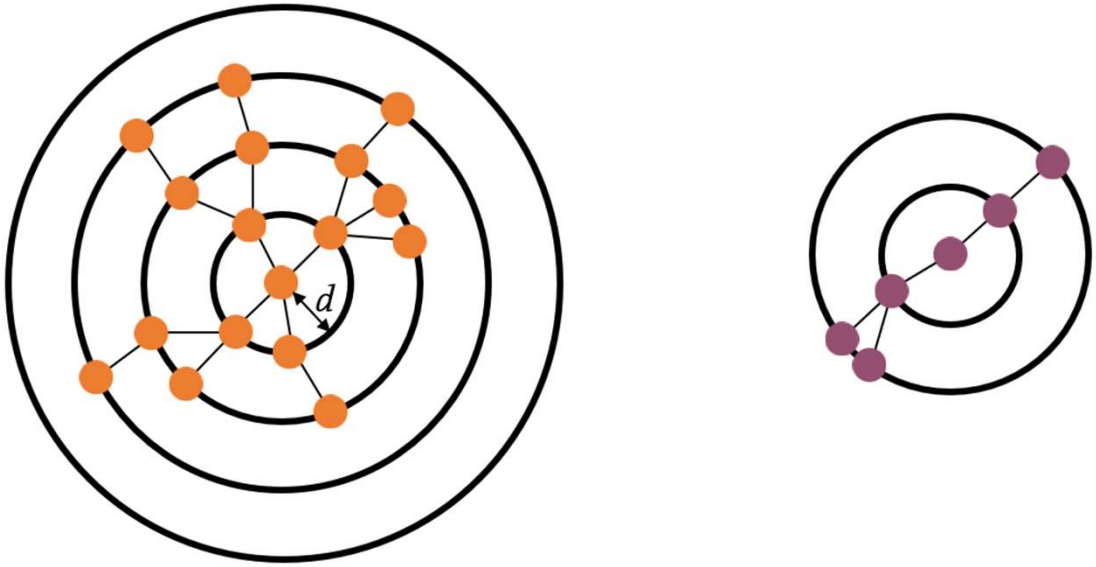
SWARM tackles these two problems:

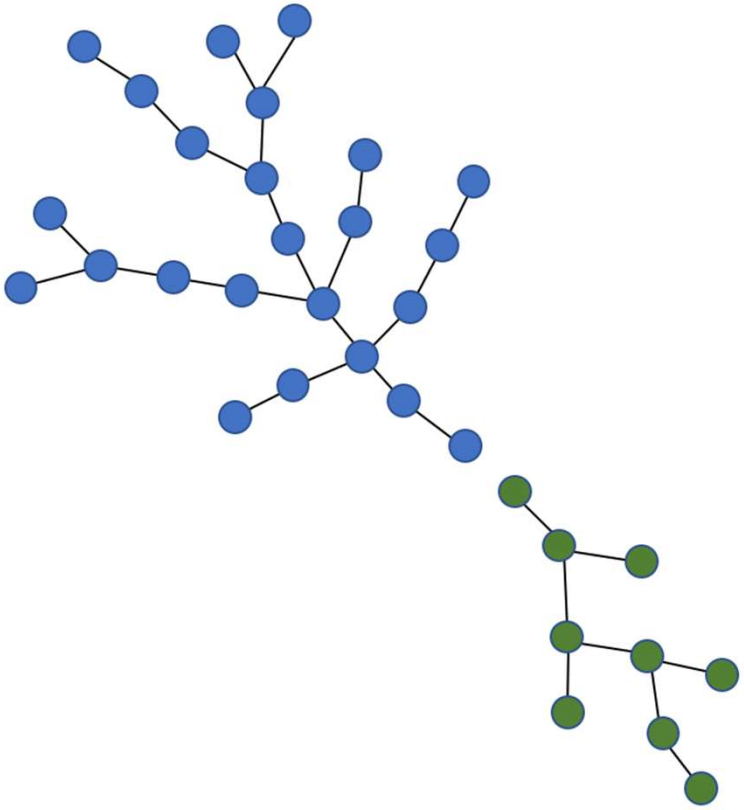
1. Order **doesn't** matter

2. **No** fixed global threshold

   (rather – it uses a local threshold d)

SWARM





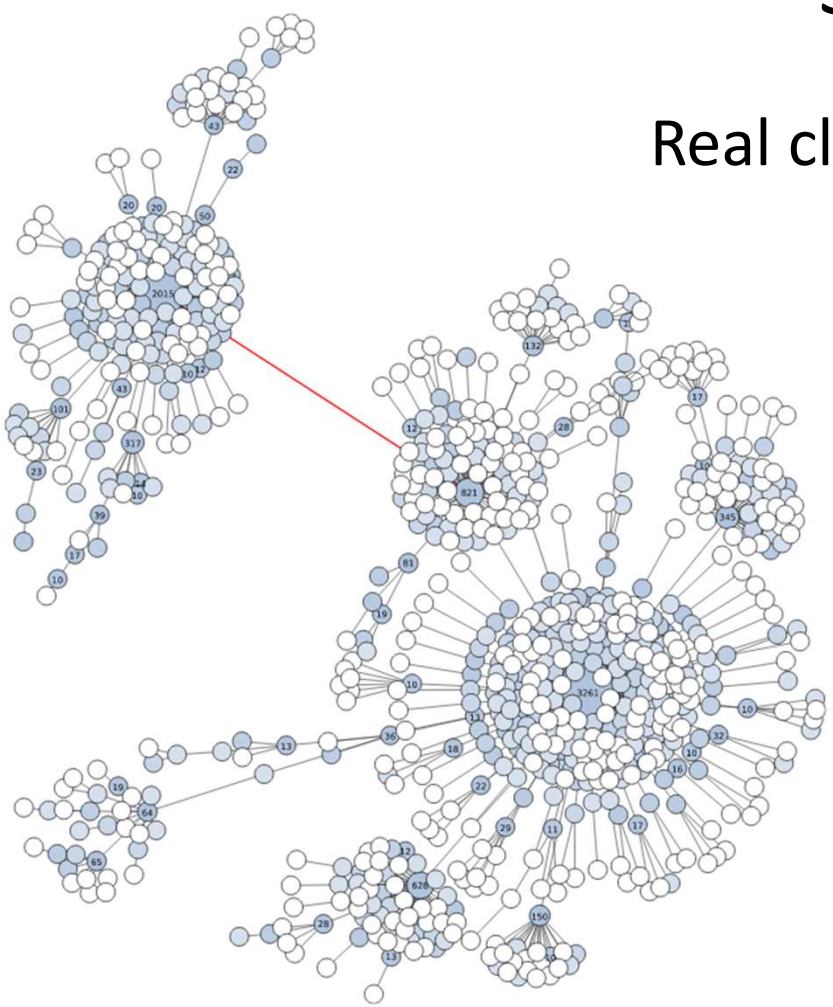
SWARM

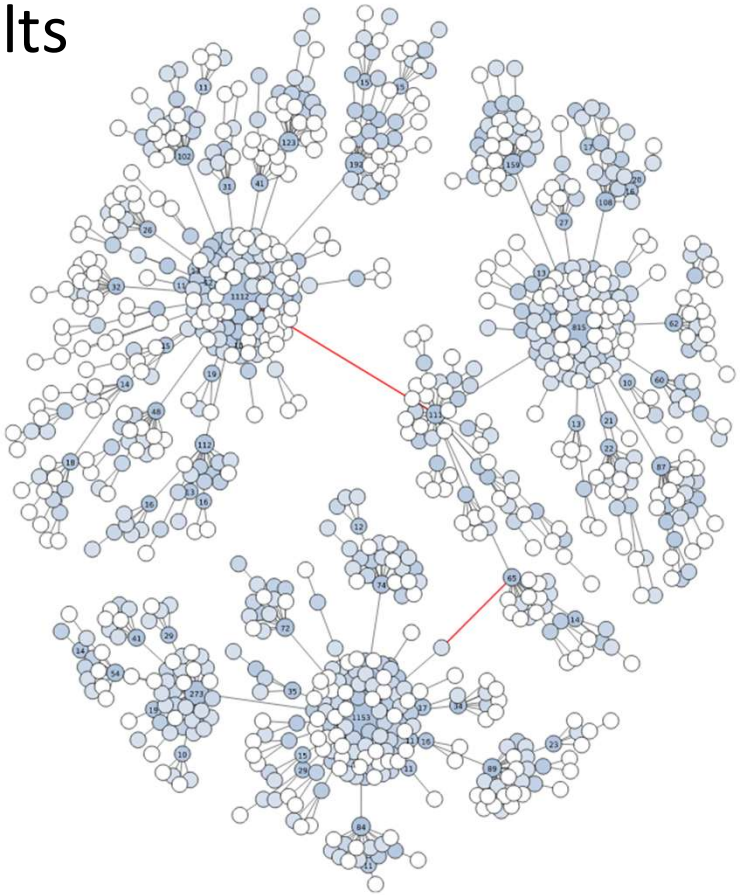
Solution:
Refine the clusters
(break up chains)

1. Detect “Peaks”
2. Go towards the minimum
3. Detect the “Valleys”
4. If ratio (between peak and valley) is high, break the chain

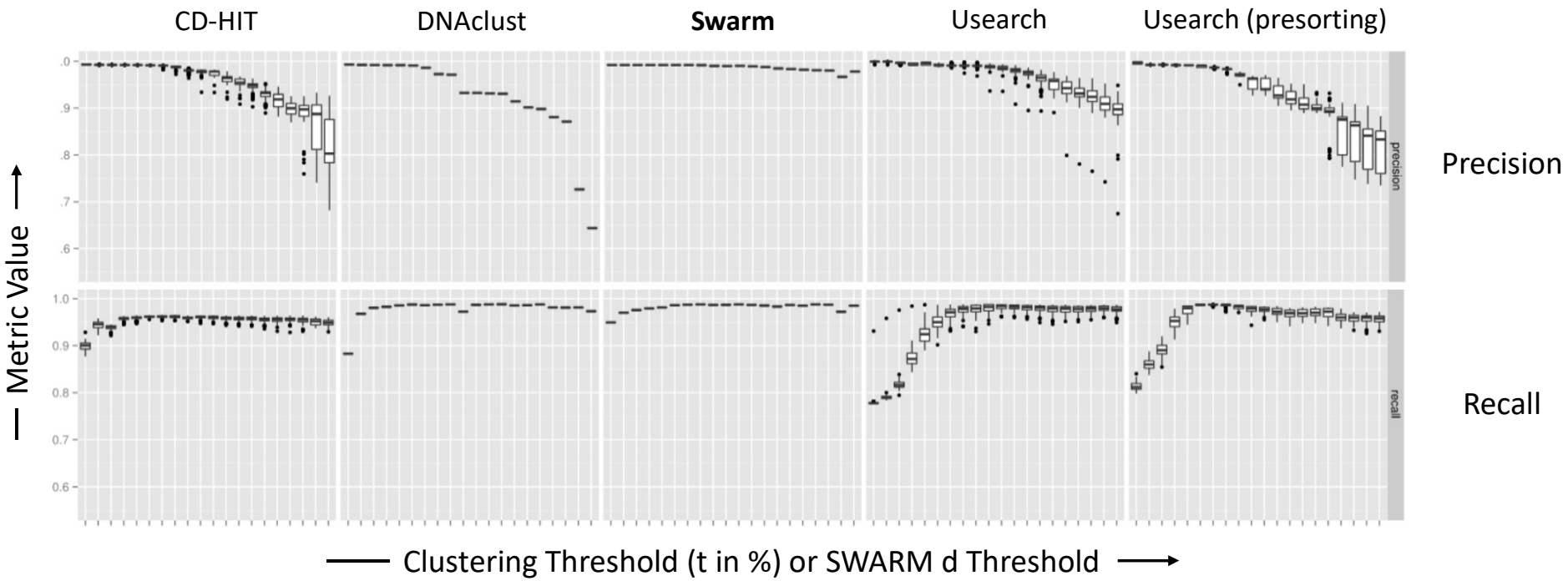
SWARM

Real clustering results





Numerical Results



Precision: % of amplicons (sequences) assigned to the same OTU actually of the same species.

Recall: % of amplicons assigned to the same species grouped in the same OTU.

SWARM

More reasons we like SWARM:

Open Source

Parallelizable

Cool Name

(https://github.com/torognes/swarm)

SWARM v2

Two major improvements:

1. Big speed improvement ($O(n^2) \rightarrow O(nL)$)

2. Avoiding small (low abundant) OTUs

SWARM v2

1. Speed Improvement (only for $d = 1$)

Microvariants

Substitution:

AGAGAATCAGTATAGCCGAGACTAGAG

AGAGAATCAGTCTAGCCGAGACTAGAG

$3L$

Insertion:

AGAGAATCAGTATAGCCGAGACTAGAG

AGAGAATCAGTATAGCCGATGACTAGAG

$3(4L(L+1)1)+1$

Deletion:

AGAGAATCAGTATAGCCGAGACTAGAG

AGAGATCAGTATAGCCGAGACTAGAG

L

SWARM v2

1. Speed Improvement

Total number of microvariants are linear in length: maximum of $7L + 4$

Instead of comparing each sequence to all sequences, use a **hash table**!

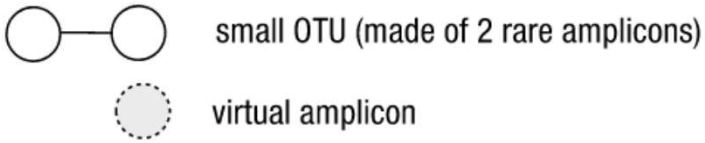
Time complexity drops from $O(n^2)$ to $\boldsymbol{O(nL)}$

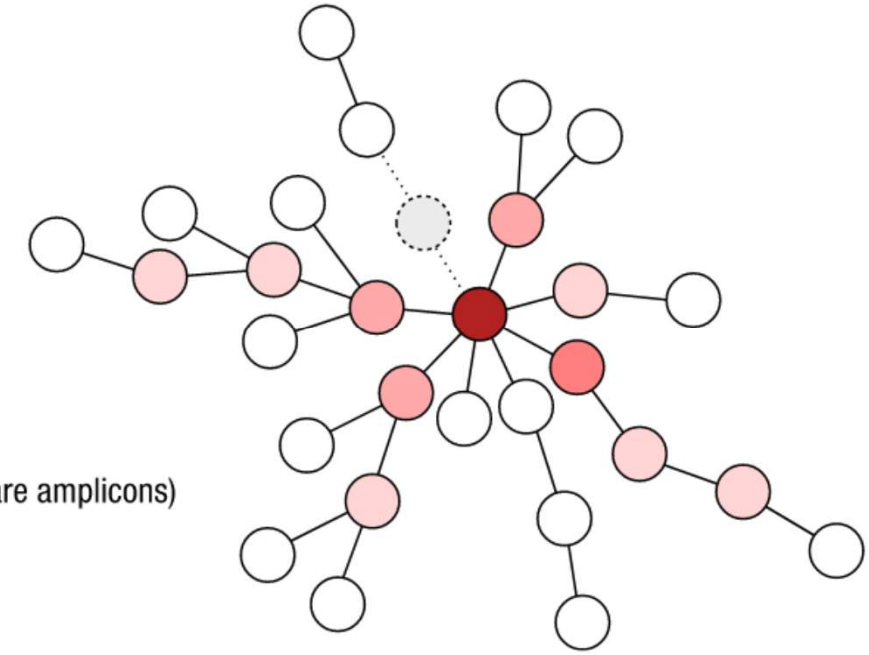
(Increased memory complexity due to hash table, but shown to be linear)

SWARM v2

2. Avoiding small (low abundant) OTUs

Low abundant OTUs (< 3 amplicons) are checked against large abundant OTUs and merged if share a microvariant.

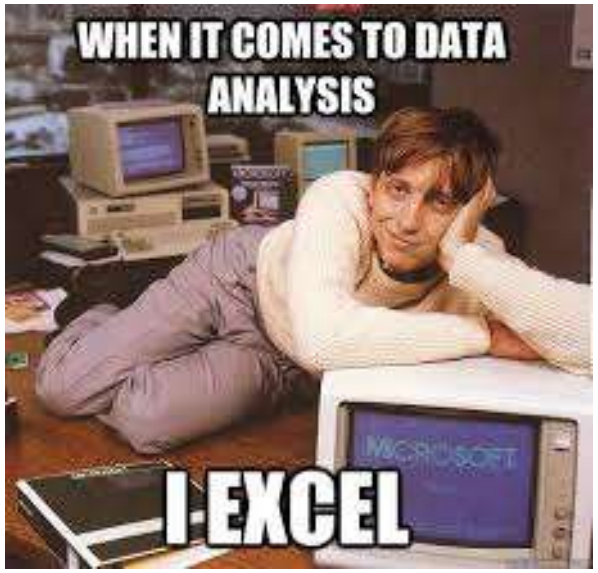




Summary

SWARM (v2) offers improved and more natural clustering method:

1. Uses a local, more natural, threshold
2. Not sensitive to user choices and ordering
3. Speed improvements on second version



Now we can take the data (OTUs) and do science!

Discussion Points

- We've seen precision vs recall in the context of OTU clustering.

If they are 'one on the account of the other' (assuming evenly), do we have a clear preference on which one is more important?

***Precision:** % of amplicons assigned to the same OTU actually of the same species.*
***Recall:** % of amplicons assigned to the same species grouped in the same OTU.*

- How can open source algorithms, libraries and pipelines

affect private and recreational research?
Will this field always remain accessible to researches only?