Introduction to Bioinformatics

Clustering

Lecturer: Jan Baumbach
Teaching assistant(s): Diogo Marinho

Outline

- Intro;
- Clustering Methods;
- Cluster Validation;
- Homology Detection;
- Case Studies;
- Missing values.

Introduction

- Biomedical experiments result in ever growing data sets in terms of number of samples and dimensionality

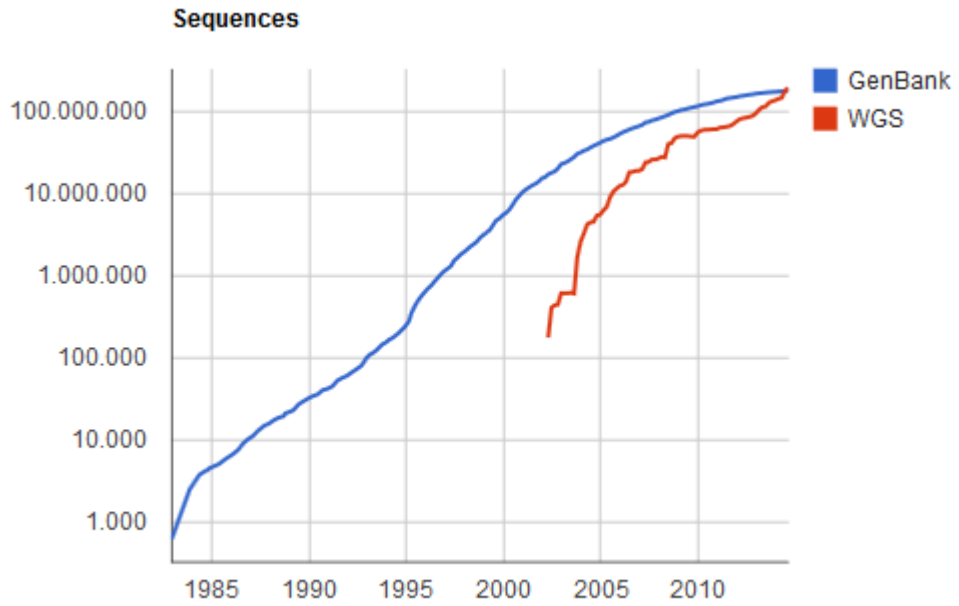
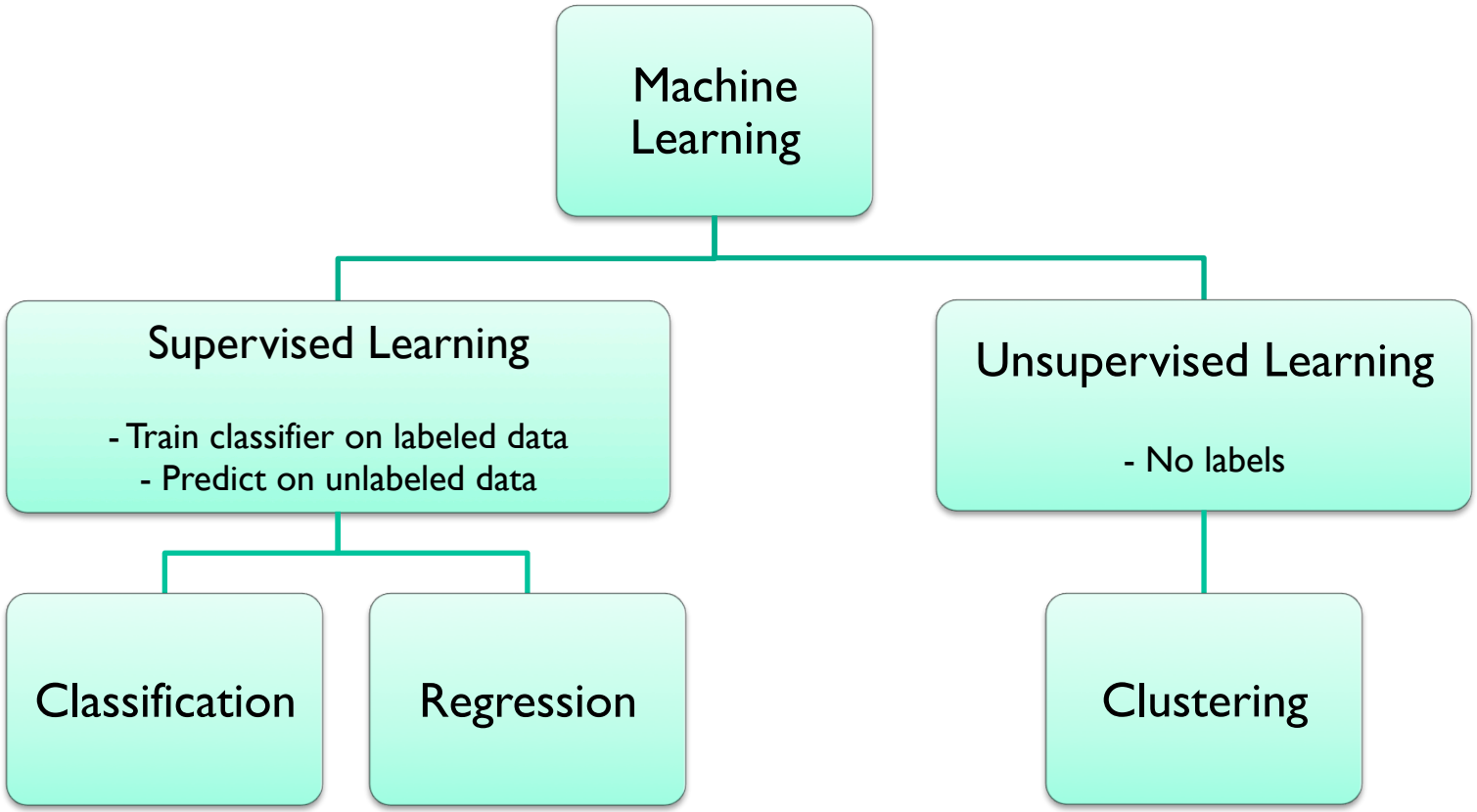


Image taken from http://www.ncbi.nlm.nih.gov/genbank/statistics

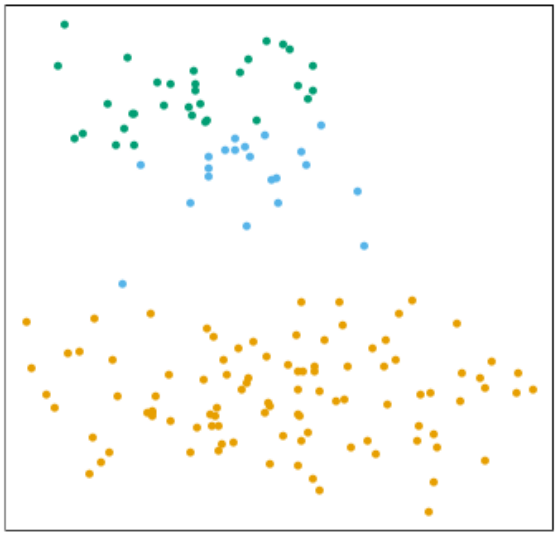
Introduction

- Manual analysis infeasible



Clustering

- Point representation of objects
- Proximity corresponds to similarity



- Identify groups of similar objects
  - Objects within same group highly similar
  - Objects of different groups highly dissimilar

Clustering Methods

- Long list of clustering methods
  - Each with pros and cons
  - No universally best performer
- Classification of methods
  - Density-based
  - Graph-based
  - Hierarchical
  - Model-based
  - Prototype-based

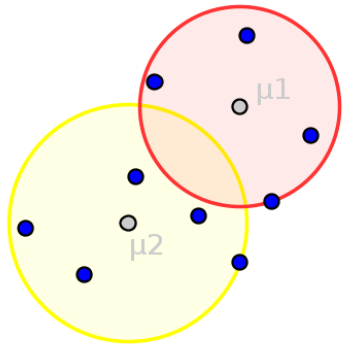
Clustering Methods

- Some example methods
- Density-based
- Graph-based
  - Affinity Propagation
  - Transitivity Clustering
- Hierarchical
- Model-based
- Prototype-based
  - k-Means

k-Means

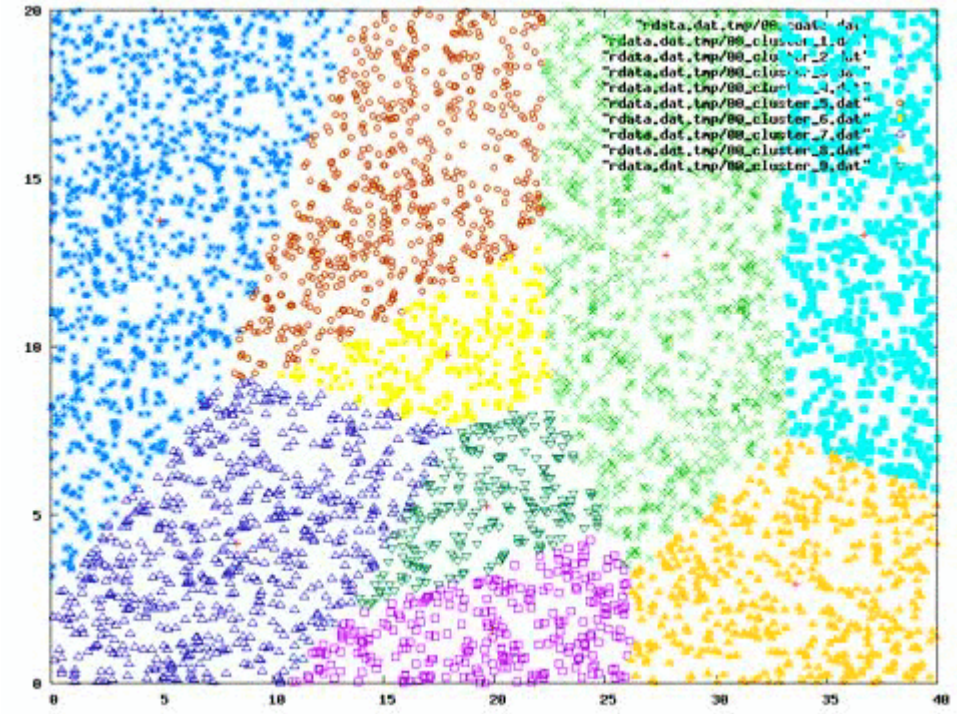
k-Means

- Prototype-based method
  - Cluster defined by a representative "prototype"
  - Assign objects to cluster with most similar prototype
- Prototypes are centroids (arithmetic mean) of clusters



k-Means

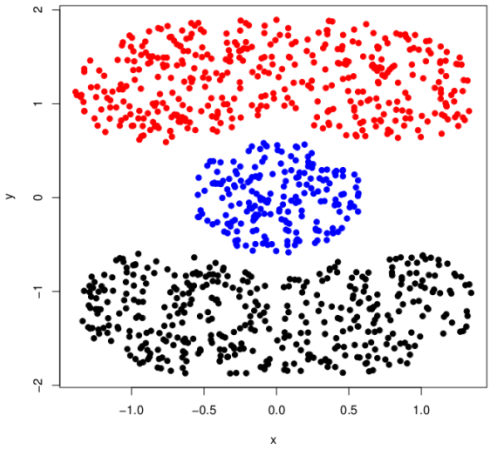
- Main parameter: k (number of clusters)
  - Also: number of random starts



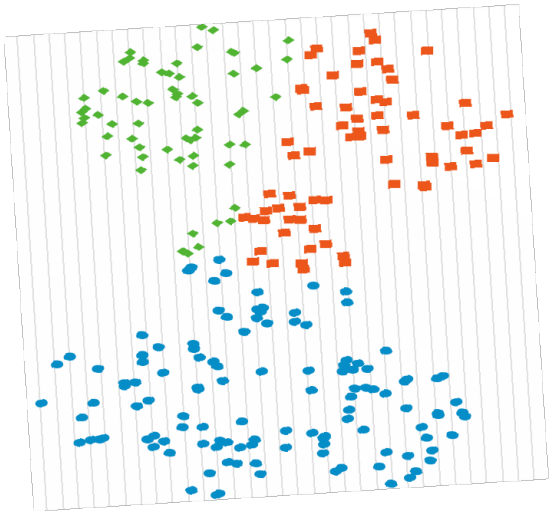
k-Means

- Non-deterministic, due to the random seed
- Assumption: Clusters are spherical
- Non-spherical/non-convex example:

3 classes:



K-Means with k=3 clusters:



Similarity Graph

- Clustering methods require input either as
  - Absolute coordinates or
  - Pairwise similarities s(x,y)
- Similarities can be converted to distances and vice versa
  - E.g. if $s(x, y) \in [0,1]$ then $d(x, y) = 1 - s(x, y)$
- Similarities can be interpreted as similarity graph
  - Nodes are objects
  - Edge weights are similarities

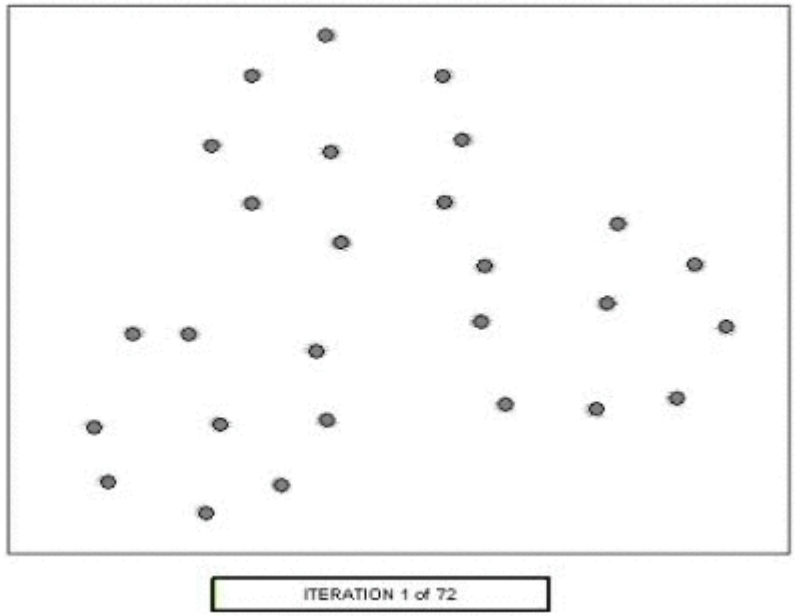
Affinity Propagation

Affinity Propagation

- Cluster prototypes called “exemplars”
- Send flow through similarity graph
  - More flow between exemplars and cluster members
- Availability
  - Candidate exemplars → objects
  - How available is object as exemplar for other objects
- Responsibility
  - Objects → candidate exemplars
  - Candidate exemplar favored over other exemplars?

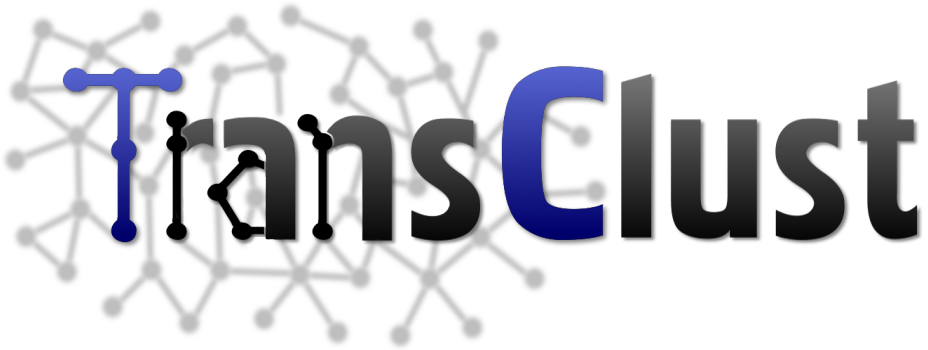
Affinity Propagation

- In each iteration update availabilities and responsibilities
  - $a(i,k) = \min\left\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}$
  - $r(i,k) = s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}$



B. J. Frey *et al.* Science 315, 972-976 (2007)

Transitivity Clustering



The problem of finding similar groups...

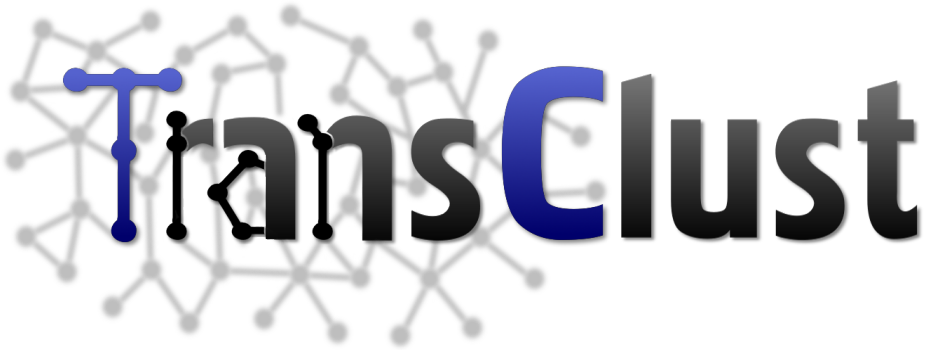
Given: - Set of objects M (e.g. protein sequences)

- Similarity function $s : \binom{M}{2} \rightarrow \mathbb{R}$

(e.g. all-vs.-all BLAST results)

Goal: - Find subsets (cluster) which are highly connected according to the similarity

→ Find clusters of homologous proteins



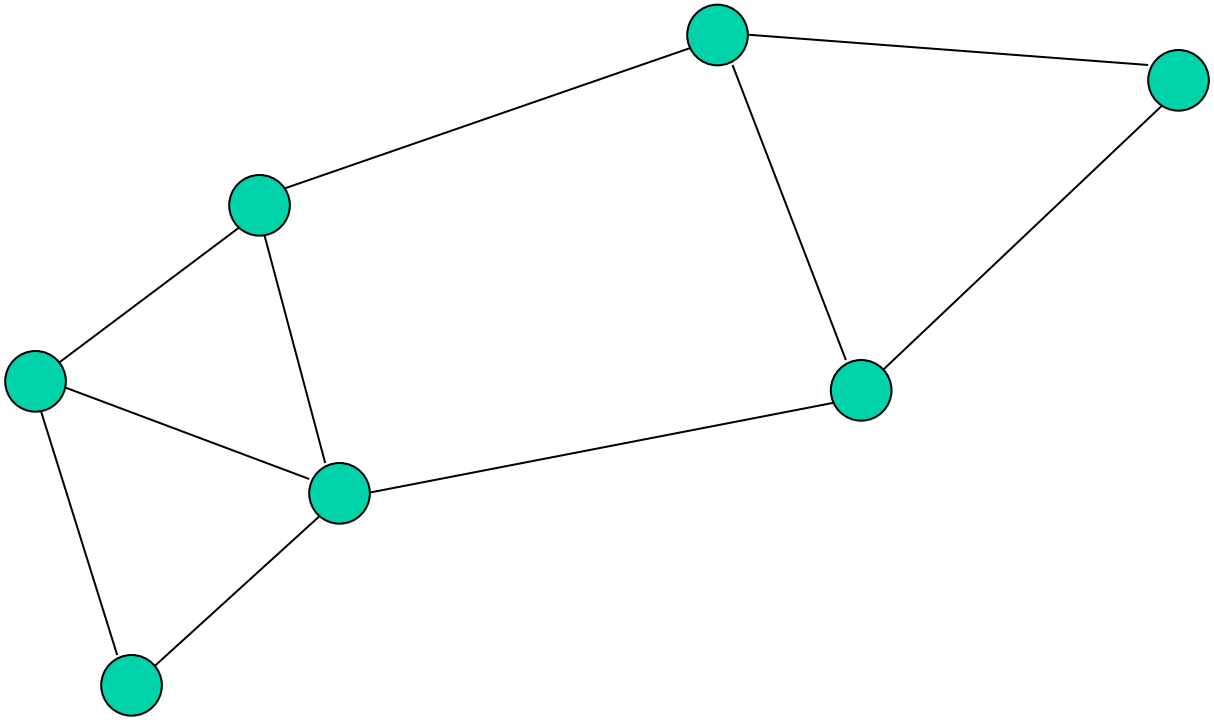
1. Similarity graph G
   - Nodes: proteins, edges: similarities, edge weights *w*

2. Edge weight threshold *t* to construct graph G‘
   - Existing edges: $w > t$
   - Not existing edges: $w < t$
   - Costs function for edge add/remove: e.g. $c = |w - t|$

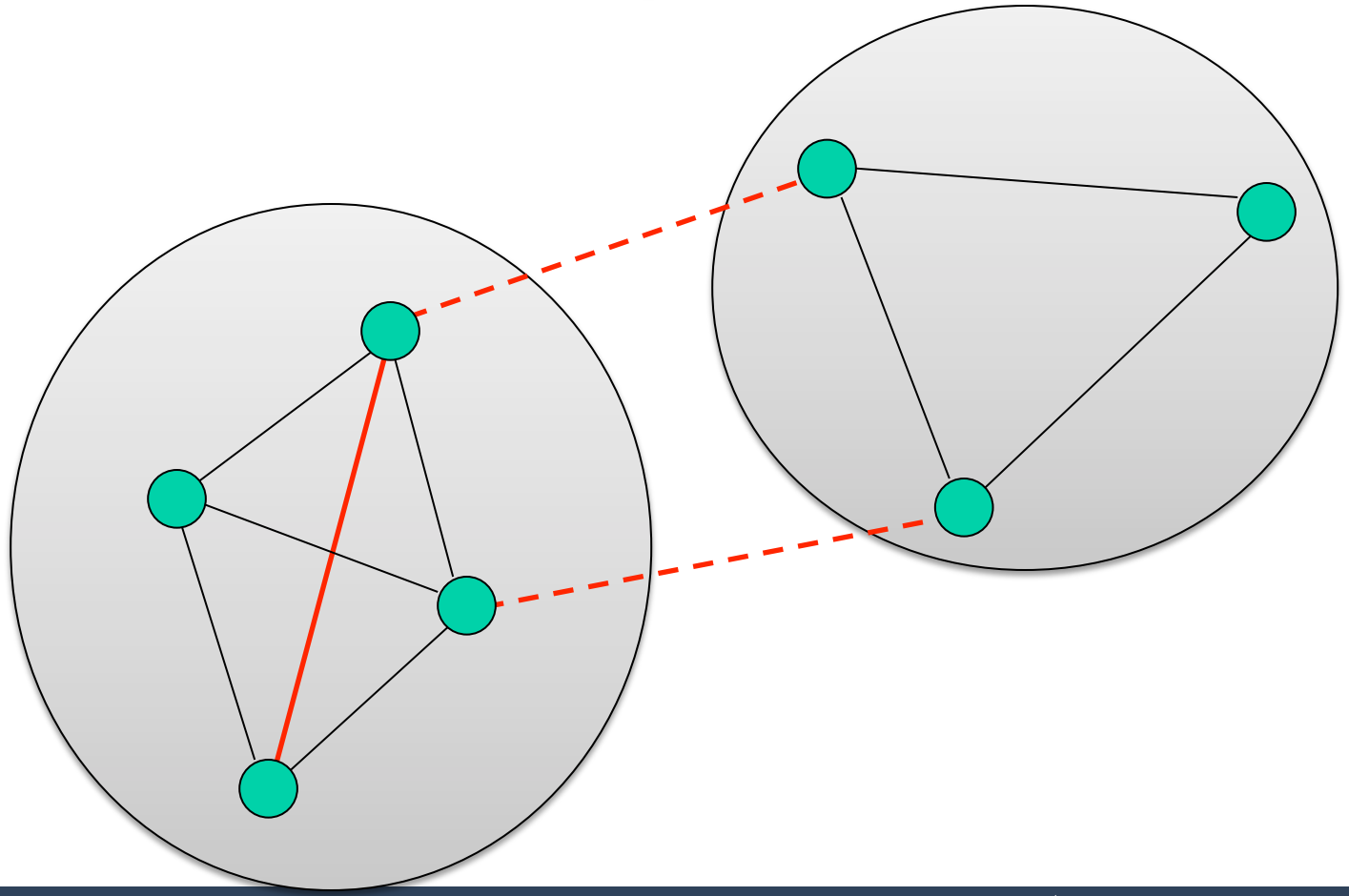
3. Transform graph G‘ into a transitive graph G*
   - with minimal costs *c* for edge additions/removals

Note: The problem is NP-hard and APX-hard!

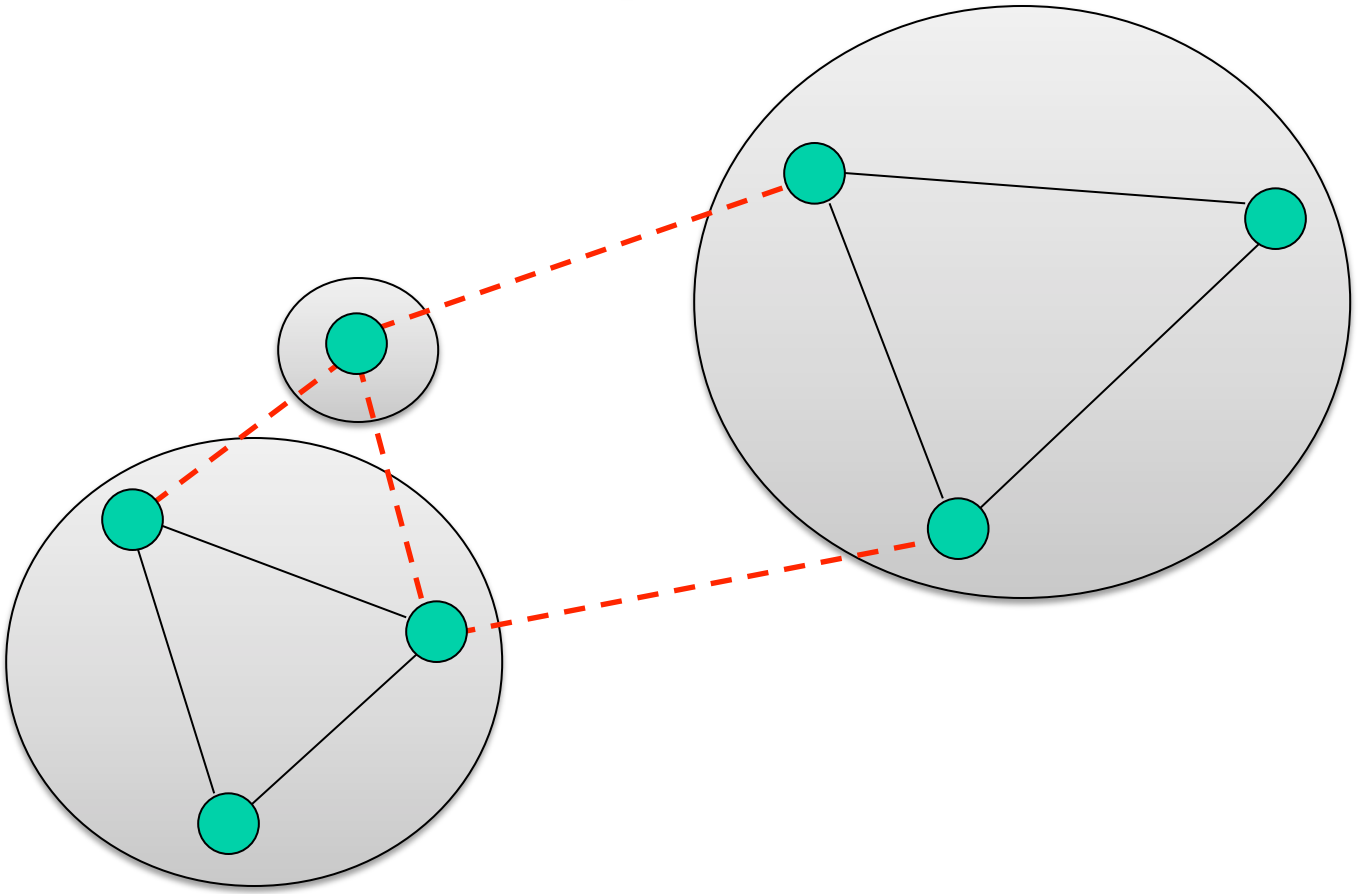
TransClust



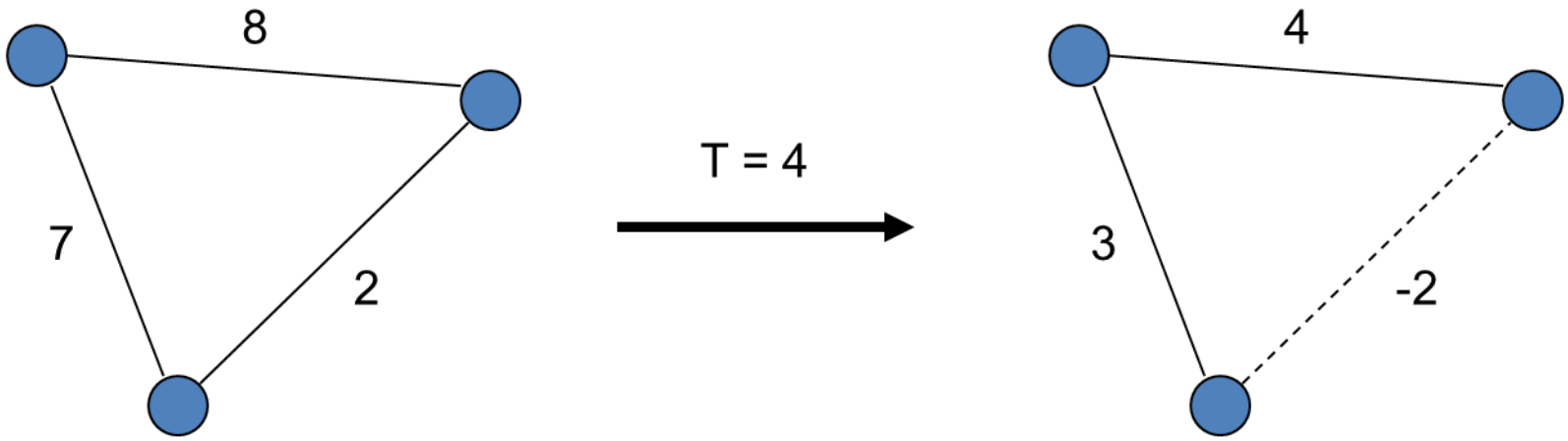
TransClust



TransClust



Weighted Transitive Graph Projection



Similarity graph

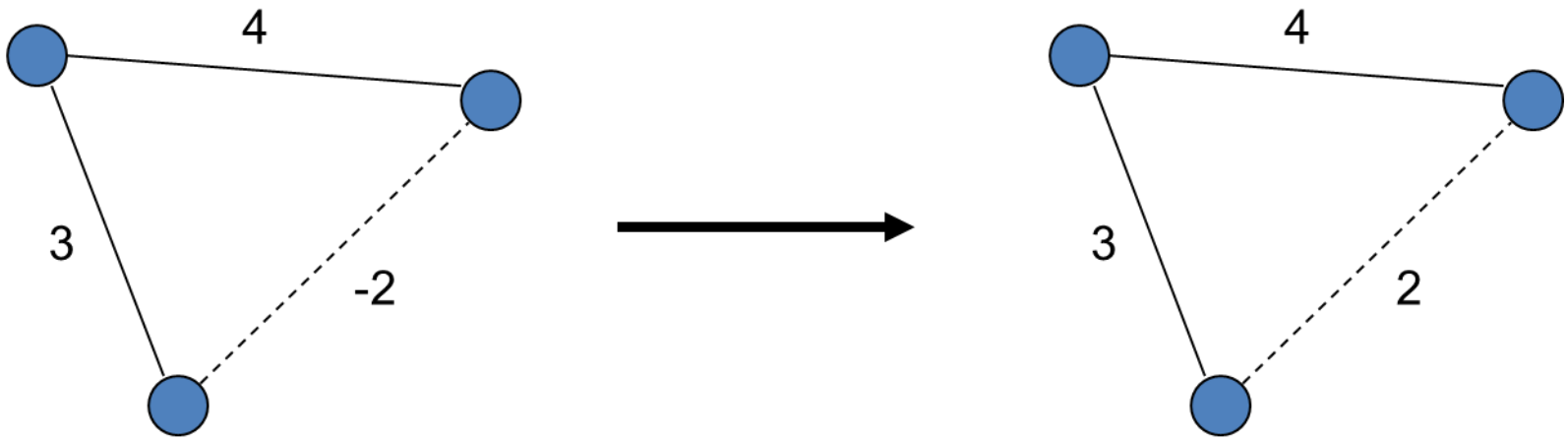
(e.g. –log10 of BLAST BBH E-value: E)

Modified similarity graph

s(x,y) = E - T

T. Wittkop *et al.* Nat Methods 7, 419-420 (2010)

Weighted Transitive Graph Projection



Modified similarity graph

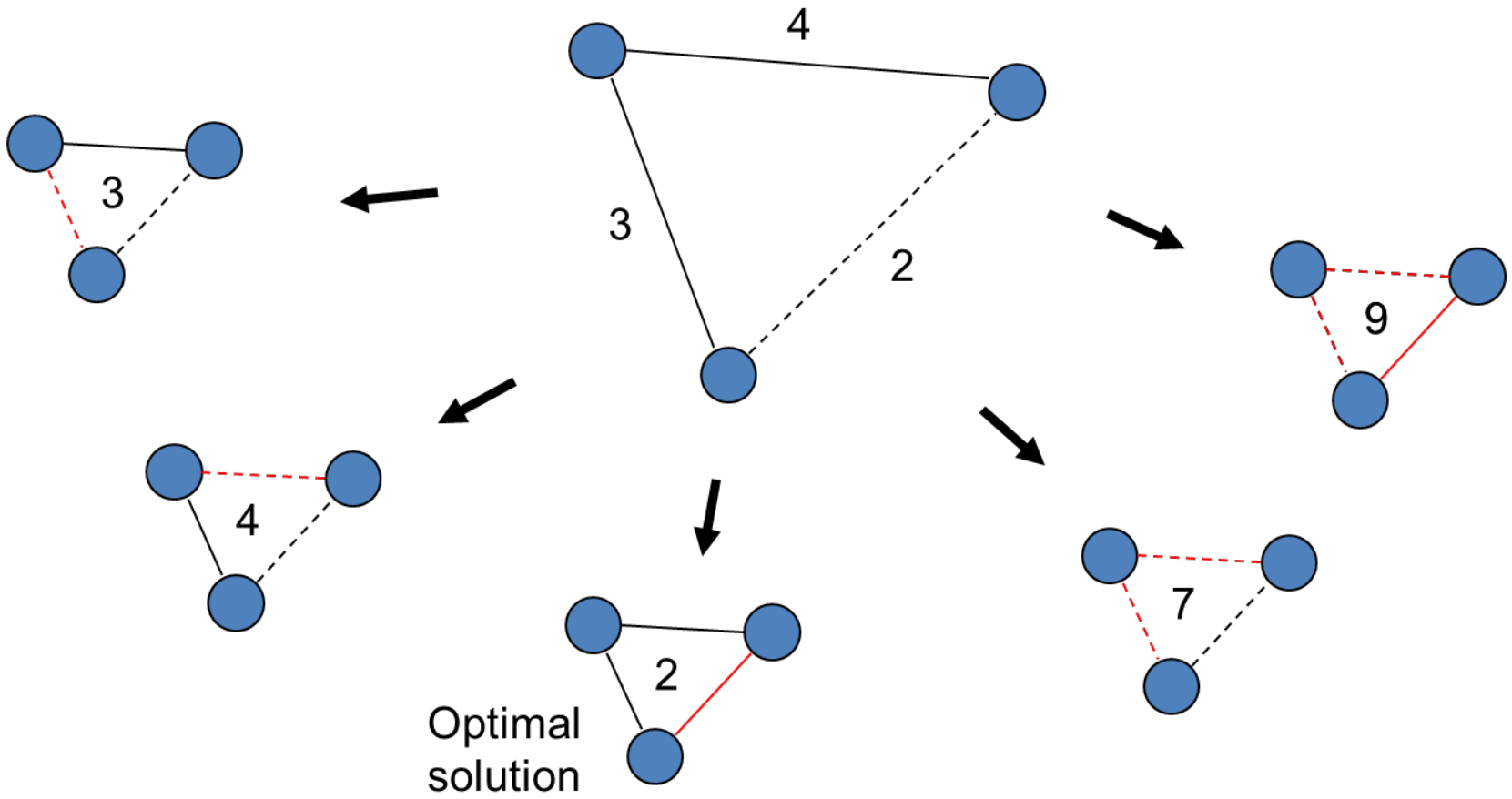
$s(x,y) = E - T$

Cost graph

$c(x,y) = | s(x,y) |$

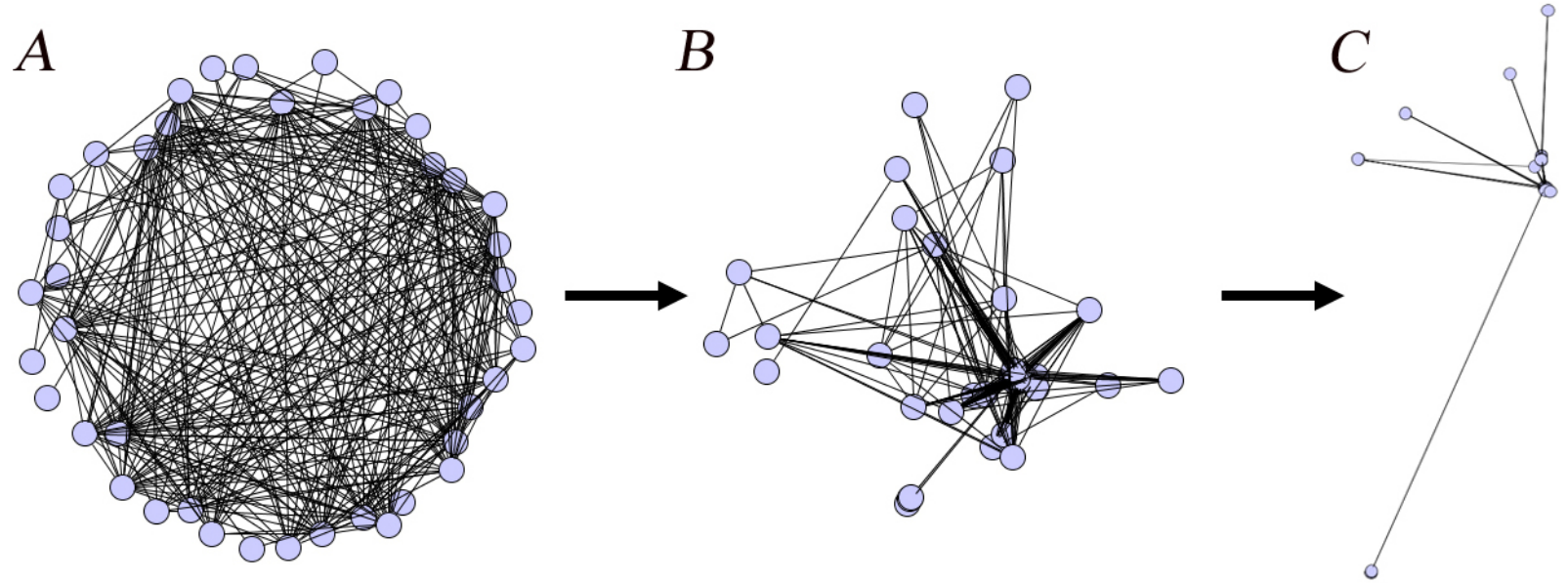
T. Wittkop *et al.* Nat Methods 7, 419-420 (2010)

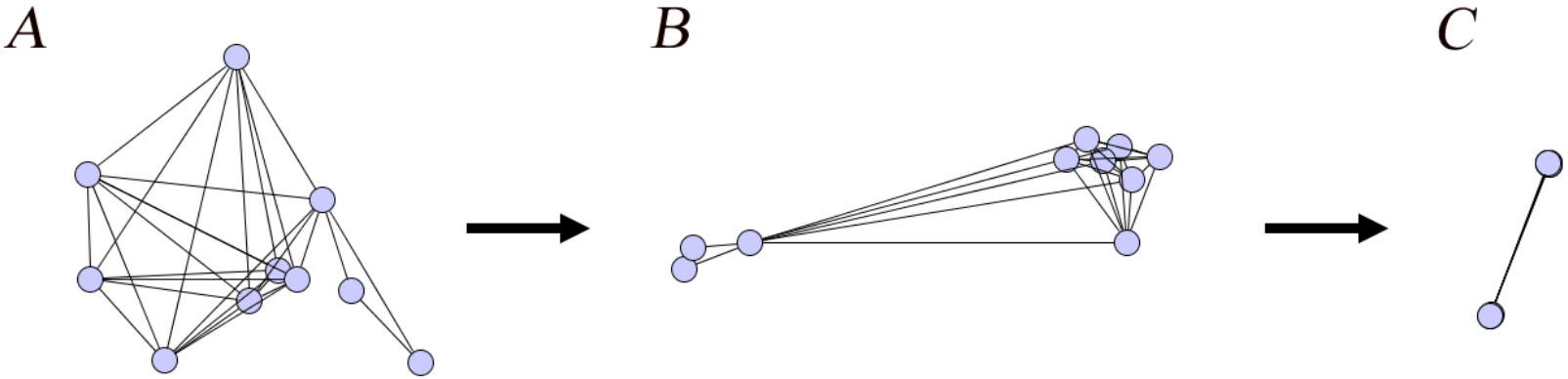
Weighted Transitive Graph Projection



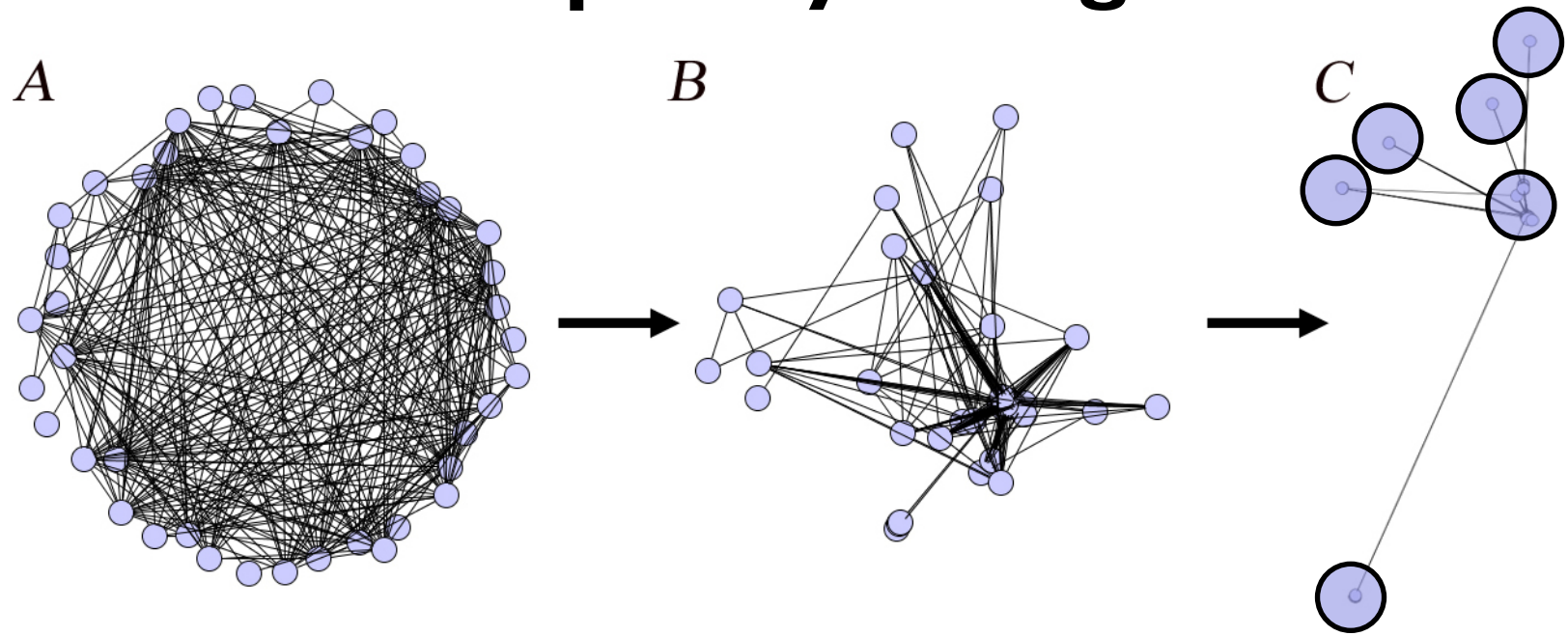
T. Wittkop *et al.* Nat Methods 7, 419-420 (2010)

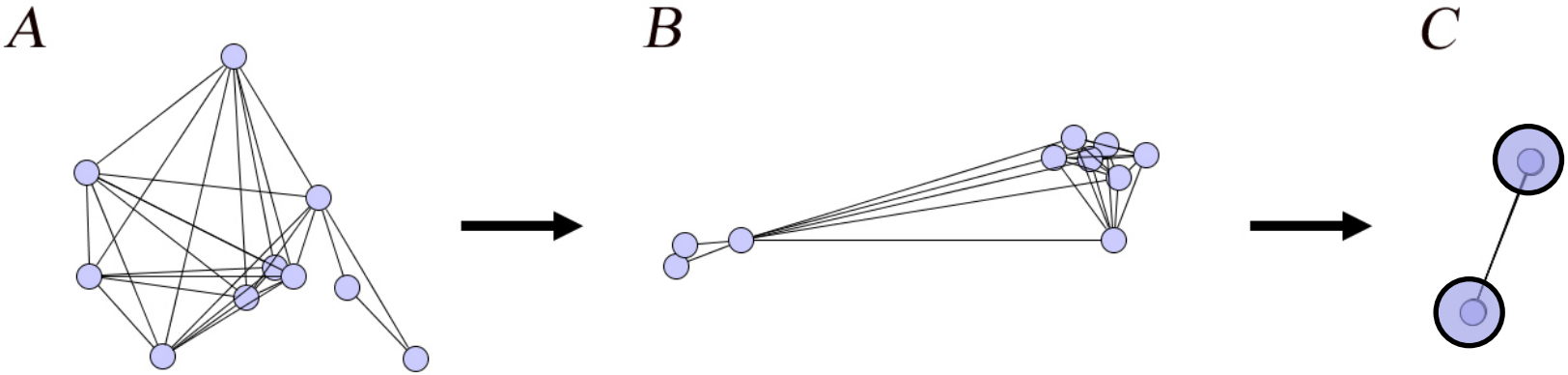
Graph Layouting



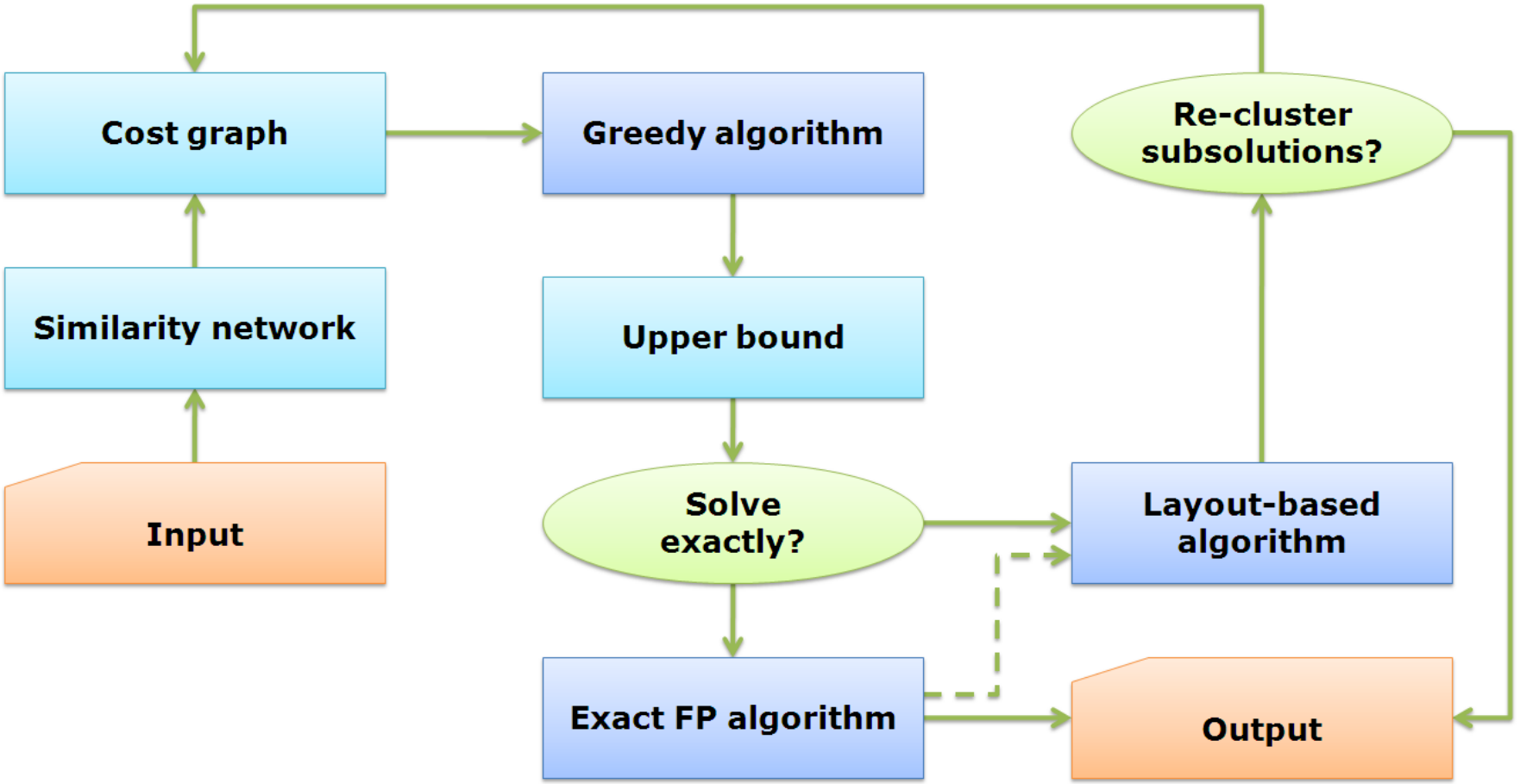


Graph Layouting

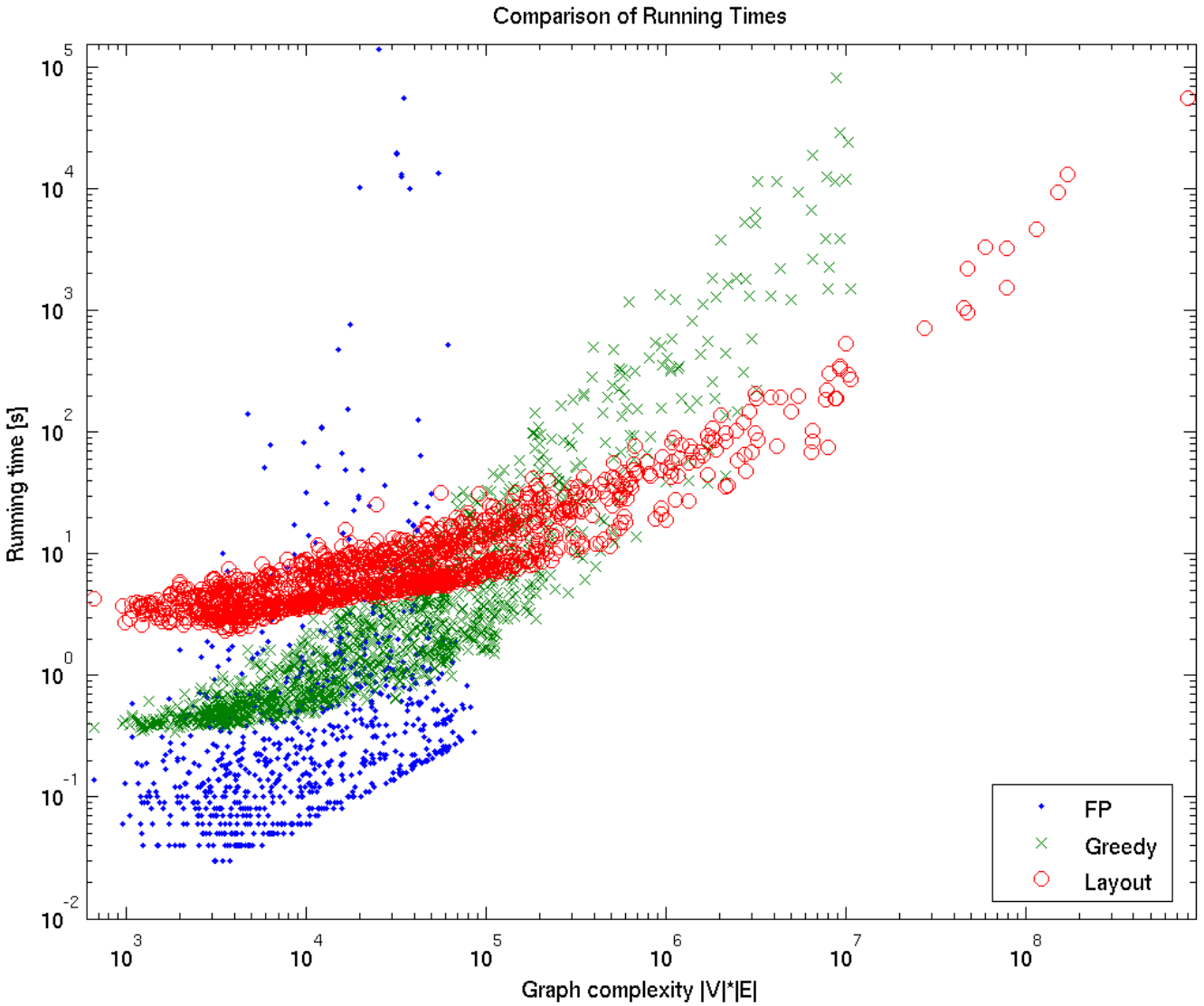




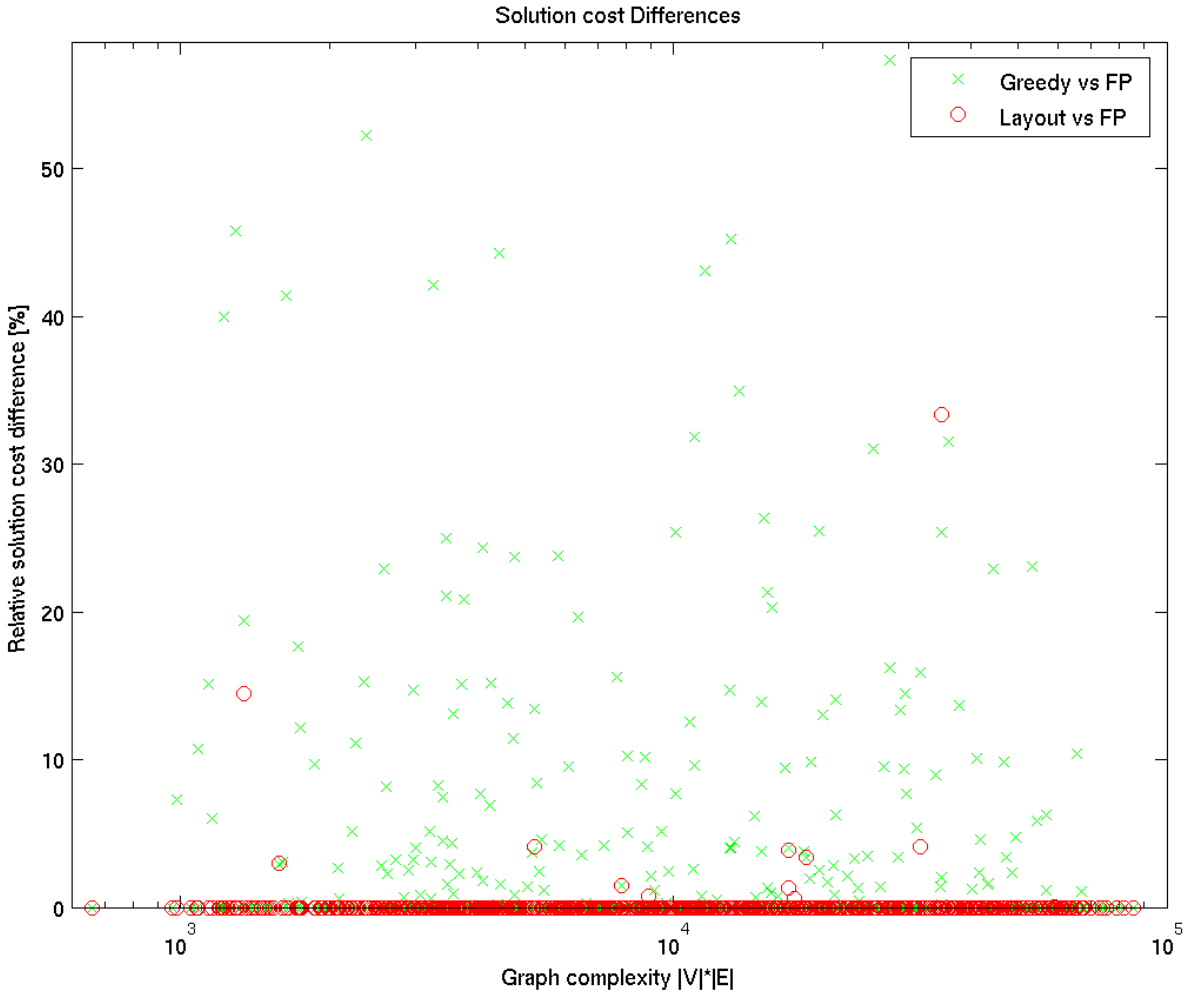
Weighted Transitive Graph Projection (Integration)



Weighted Transitive Graph Projection (Runtime)



Weighted Transitive Graph Projection (Runtime)

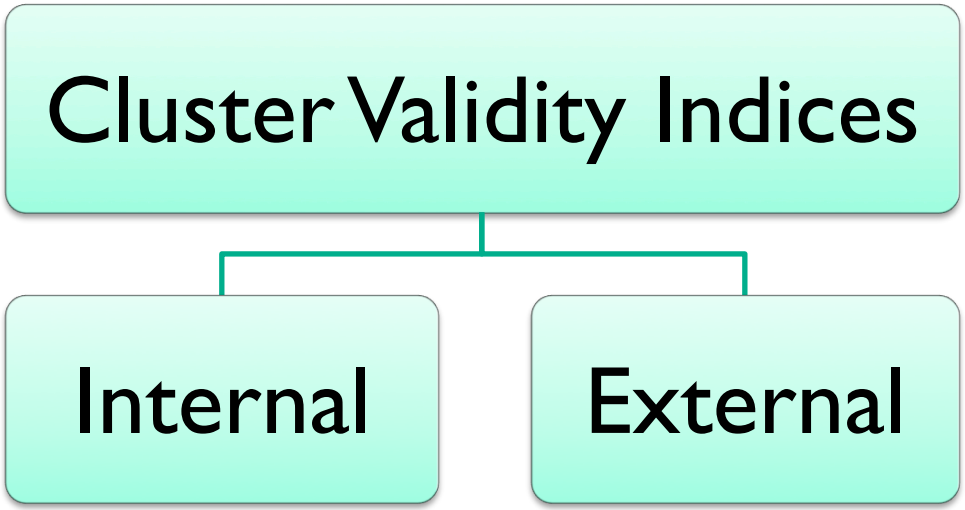


- Optimal cost for all 825 connected components: 171,986.8
- Solution cost found by Layout Heuristic: 172,244.6 (difference: 0.15%)

Cluster Validation

Cluster Validation

- A data set can always be clustered, that does not mean the clustering makes sense

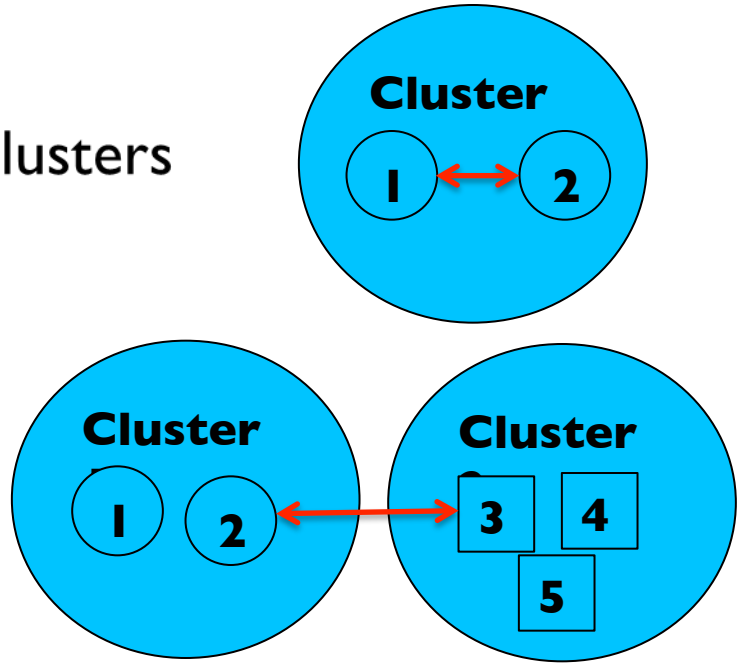


Based on intrinsic properties of the data set

Based on gold standard

Silhouette Value

- Internal index
- Also called Silhouette Width
- Defined on distances $d(x, y)$
  - Rewards small distances **within** clusters
  - Rewards large distances **between** clusters



Silhouette Value

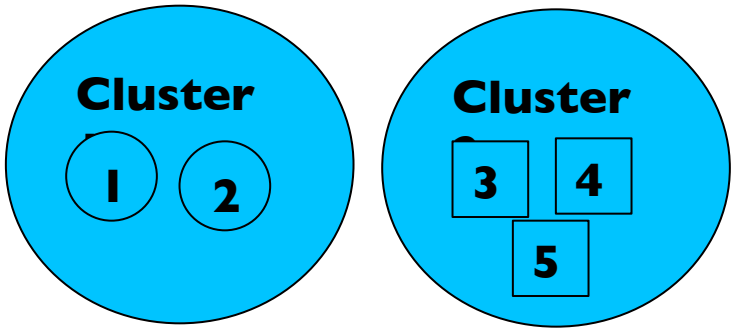
- c_i: cluster of object i
- o_i: closest other cluster to cluster of object i
- $d(i,j)$: distance of objects i and j
- $S = \frac{1}{n}\sum s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$
  - $a_i = \begin{cases} 0, & \text{if } i = j \\ \frac{1}{|c_i|-1}\sum_{j \in c_i} d(i,j), & else \end{cases}$
  - $b_i = \frac{1}{|o_i|}\sum_{j \in o_i} d(i,j)$
- $S \in [-1,1]$
- Undefined for clusterings with 1 cluster and for empty

Average distance of object i to other objects j of same cluster

Average distance of object i to objects j of closest other cluster

Silhouette Value Example

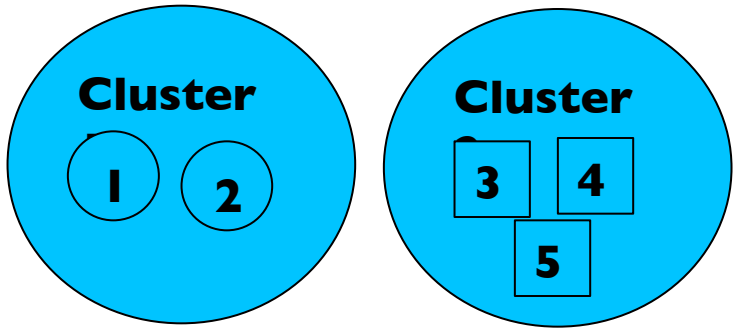
- Clustering C



Symmetric distances d(x,y):

	1	2	3	4	5
1	0.0	0.1	0.8	0.7	0.9
2	0.1	0.0	0.85	0.95	0.9
3	0.8	0.85	0.0	0.5	0.15
4	0.75	0.95	0.5	0.0	0.1
5	0.9	0.9	0.15	0.1	0.0

Silhouette Value Example



	1	2	3	4	5
1	0.0	0.1	0.8	0.7	0.9
2	0.1	0.0	0.85	0.95	0.9
3	0.8	0.85	0.0	0.5	0.15
4	0.75	0.95	0.5	0.0	0.1
5	0.9	0.9	0.15	0.1	0.0

- $a_1 = 0.1, b_1 = \frac{1}{3}(0.8 + 0.75 + 0.9) \approx 0.817, s_1 = \frac{0.817 - 0.1}{0.817} \approx 0.878$
- $a_2 = 0.1, b_2 = \frac{1}{3}(0.85 + 0.95 + 0.9) = 0.9, s_2 = \frac{0.9 - 0.1}{0.9} \approx 0.889$
- $a_3 = \frac{1}{2}(0.5 + 0.15) = 0.325, b_3 = \frac{1}{2}(0.8 + 0.85) = 0.825, s_3 = \frac{0.825 - 0.325}{0.825} \approx 0.606$
- $s_4 \approx 0.647, s_5 \approx 0.861$

  $\Rightarrow S \approx 0.778$

F-Score

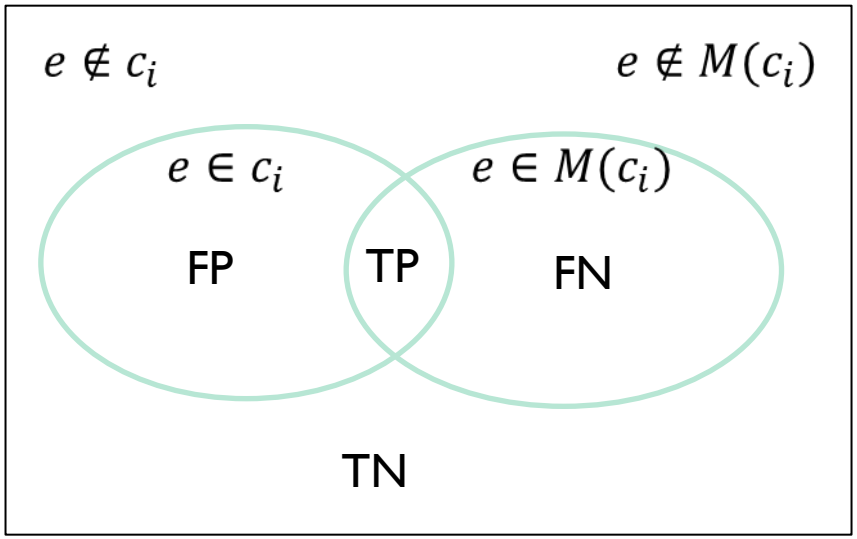
- Originates from binary classification
- Based on
  - True Positives (TP)
  - True Negatives (TN)
  - False Positives (FP)
  - False Negatives (FN)

Reality / Classification	0	1
0	TN	FN
1	FP	TP

- $Precision = \frac{TP}{TP+FP}$ $\quad Recall = \frac{TP}{TP+FN}$
- $F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * TP}{2 * TP+FN+FP}$

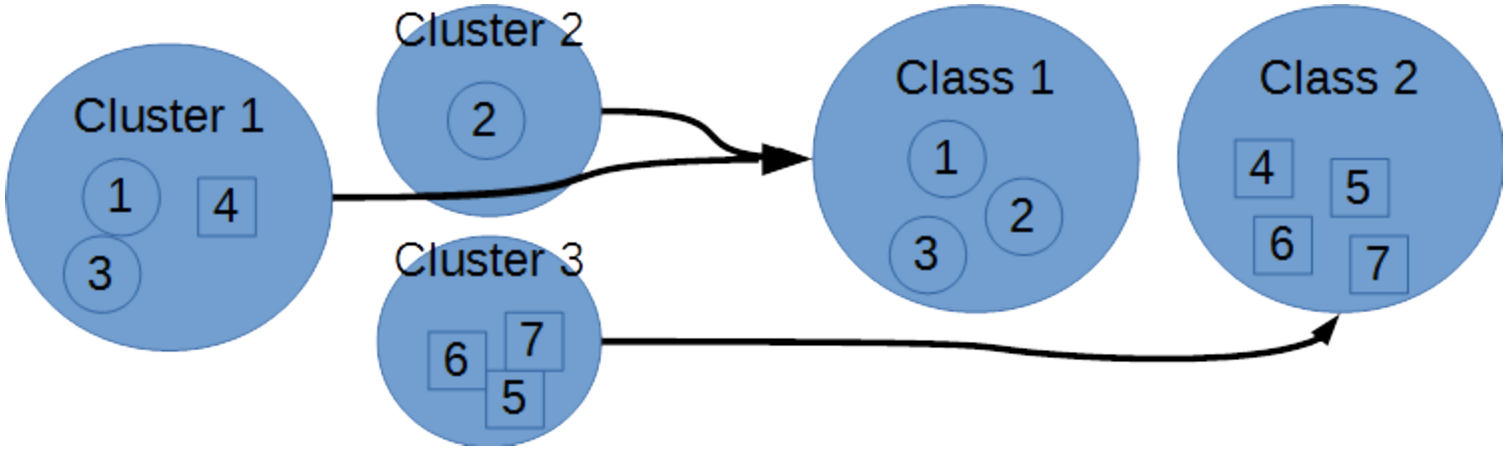
F-Score for Clustering

- Given clustering $C = \{c_i\}$ and gold standard $G = \{g_i\}$
- Find total mapping M: $C \rightarrow G$, such that
  - $\forall\, c_i \in C : c_i \cap M(c_i) \geq c_i \cap g_j$
  - "Map each cluster to the class with most common objects"
- Object e is
  - TP, if $e \in c_i \wedge e \in M(c_i)$
  - FN: if $e \notin c_i \wedge e \in M(c_i)$
  - FP: if $e \in c_i \wedge e \notin M(c_i)$

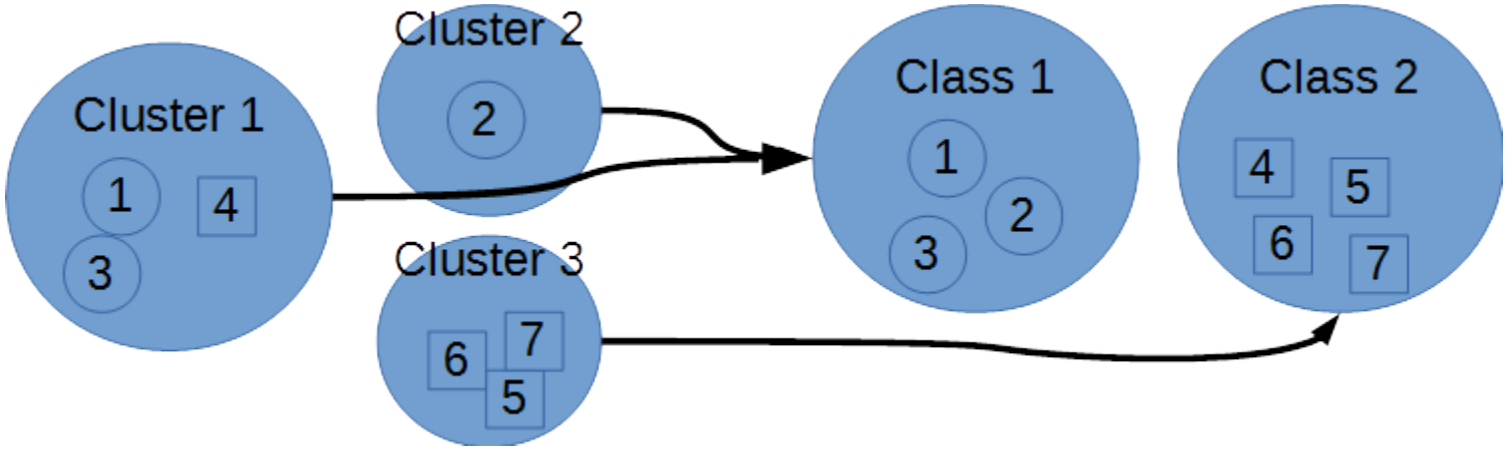


F-Score Example

- Clustering C Gold-Standard G



F-Score Example



- Cluster 1: TP = 2, FP = 1, FN = 1
- Cluster 2: TP = 1, FP = 0, FN = 2
- Cluster 3: TP = 3, FP = 0, FN= 1

$$\Rightarrow F = \frac{2 * 6}{2 * 6+4+1} = 0.706$$

Homology Detection

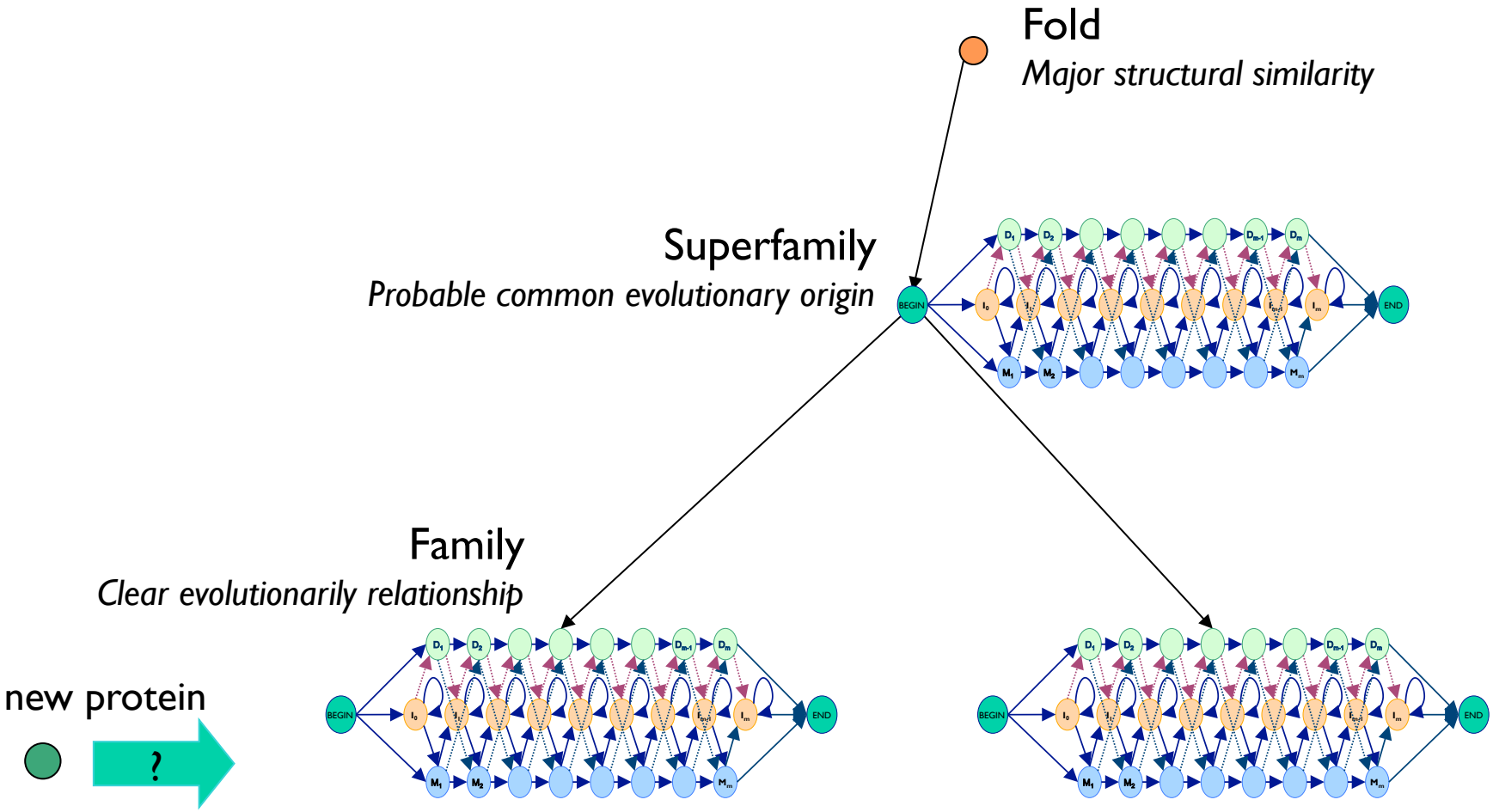
Homology

- Homologous proteins present similar:
  - Sequences,
  - Structures,
  -Functions.

- They most likely evolved from a common ancestor;

- Classification: there are three major levels in the hierarchy…

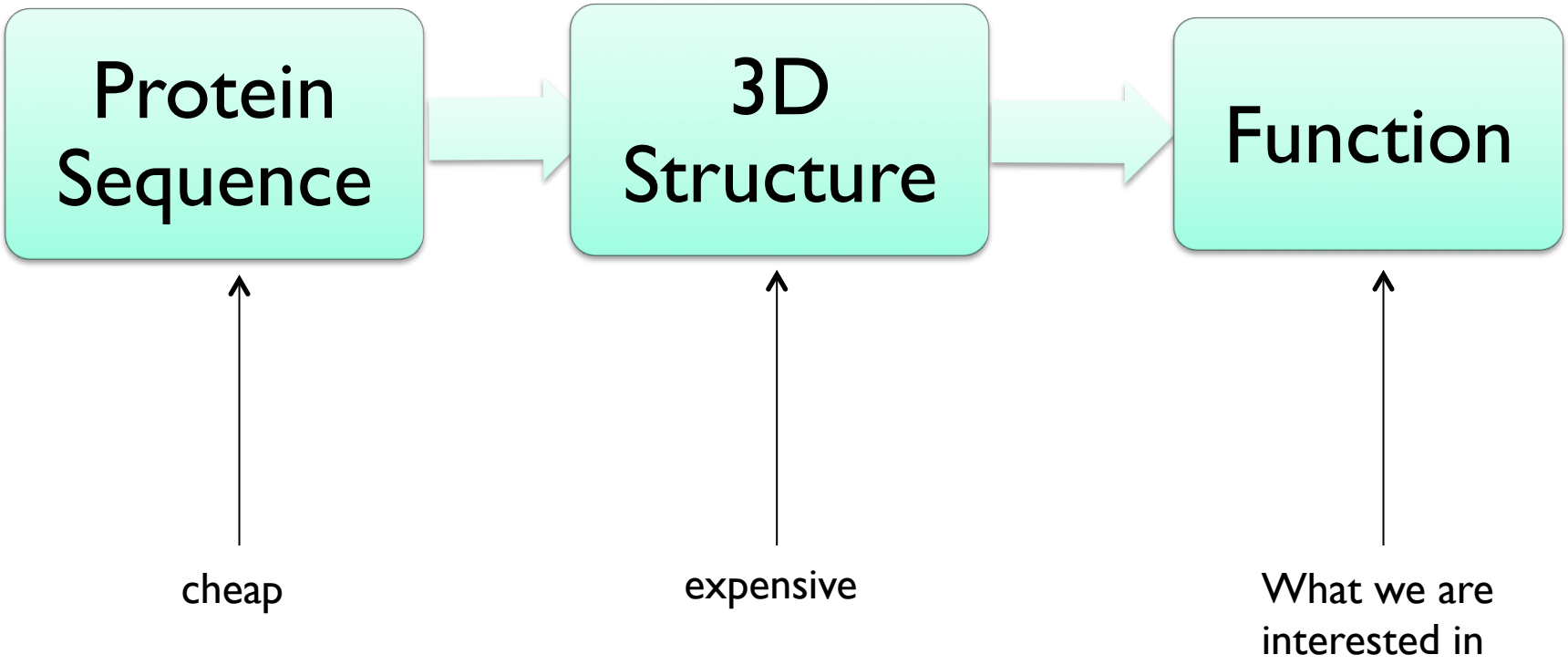
Homology
(Classification)



Case study I

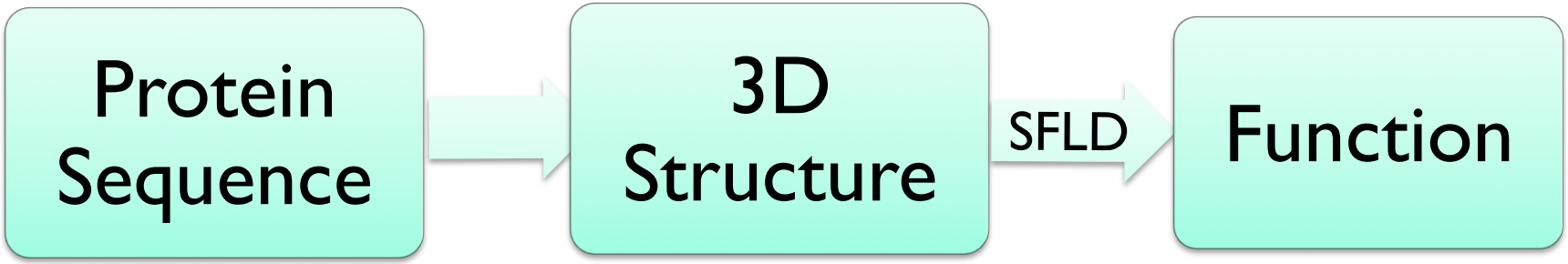
Homology Detection with Gold Standard

Case study: Amidohydrolases



Case study: Amidohydrolases

- We know the function of some proteins (gold standard)
- Stored in SFLD




```
-MPREDRATWKSNYFLKIIQLLDDYPKCFIVG
-MPREDRATWKSNYFLKIIQLLDDYPKCFIVG
-MPREDRATWKSNYFLKIIQLLDDYPKCFIVG
-MPREDRATWKSNYFLKIIQLLDDYPKCFIVG
-MPREDRATWKSNYFMKIIQLLDDYPKCFVVG
-MPREDRATWKSNYFLKIIQLLDDYPKCFIVG
-MPREDRATWKSNYFLKIIQLLDDYPKCFIVG
-MPREDRATWKSNYFLKIIQLLNDYPKCFIVG
-MVRENKAAWKAQYFIKVVELFDEFPKCFIVG
-MSGAG-SKRKKLFIEKATKLFTTYDKMIVAE
-MSGAG-SKRKNVFIEKATKLFTTYDKMIVAE
```

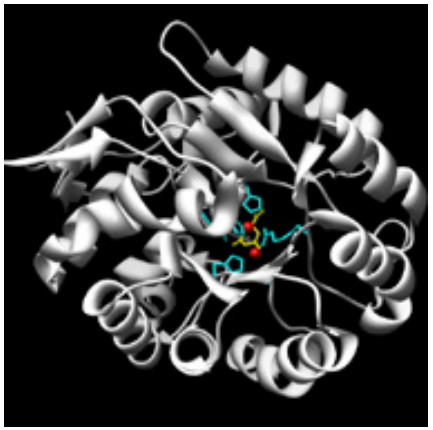


Image taken from http://sfld.rbvi.ucsf.edu/django/superfamily/4/

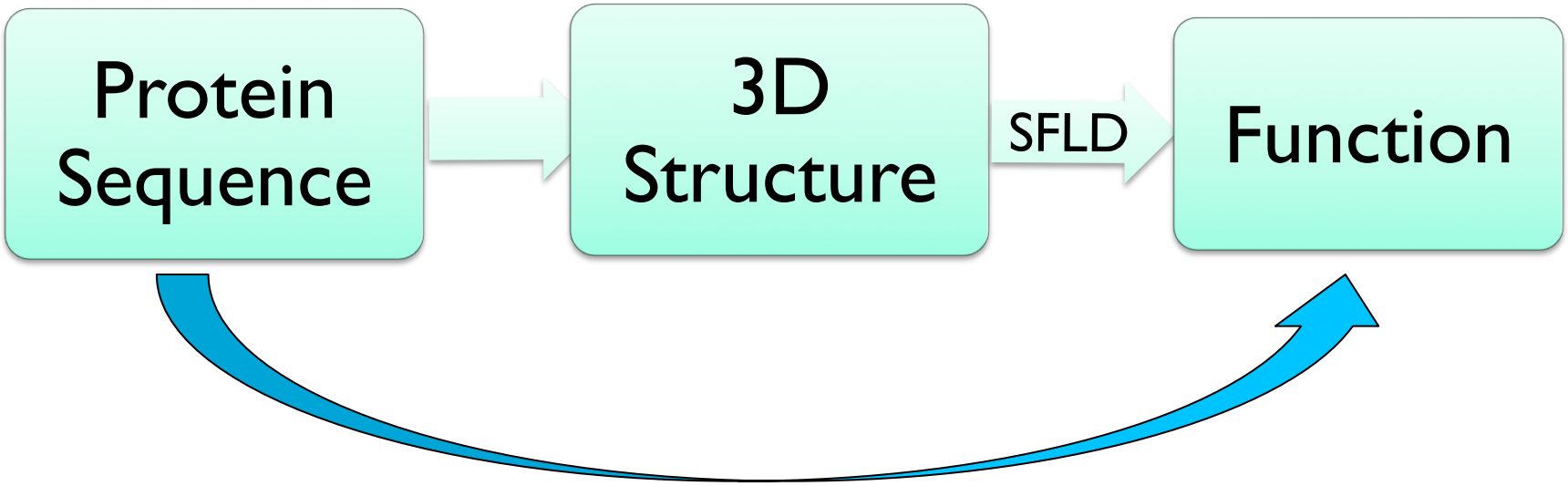
Case study: Amidohydrolases

- SFLD (Structure Function Linkage Database)
  - Hierarchical classification of proteins

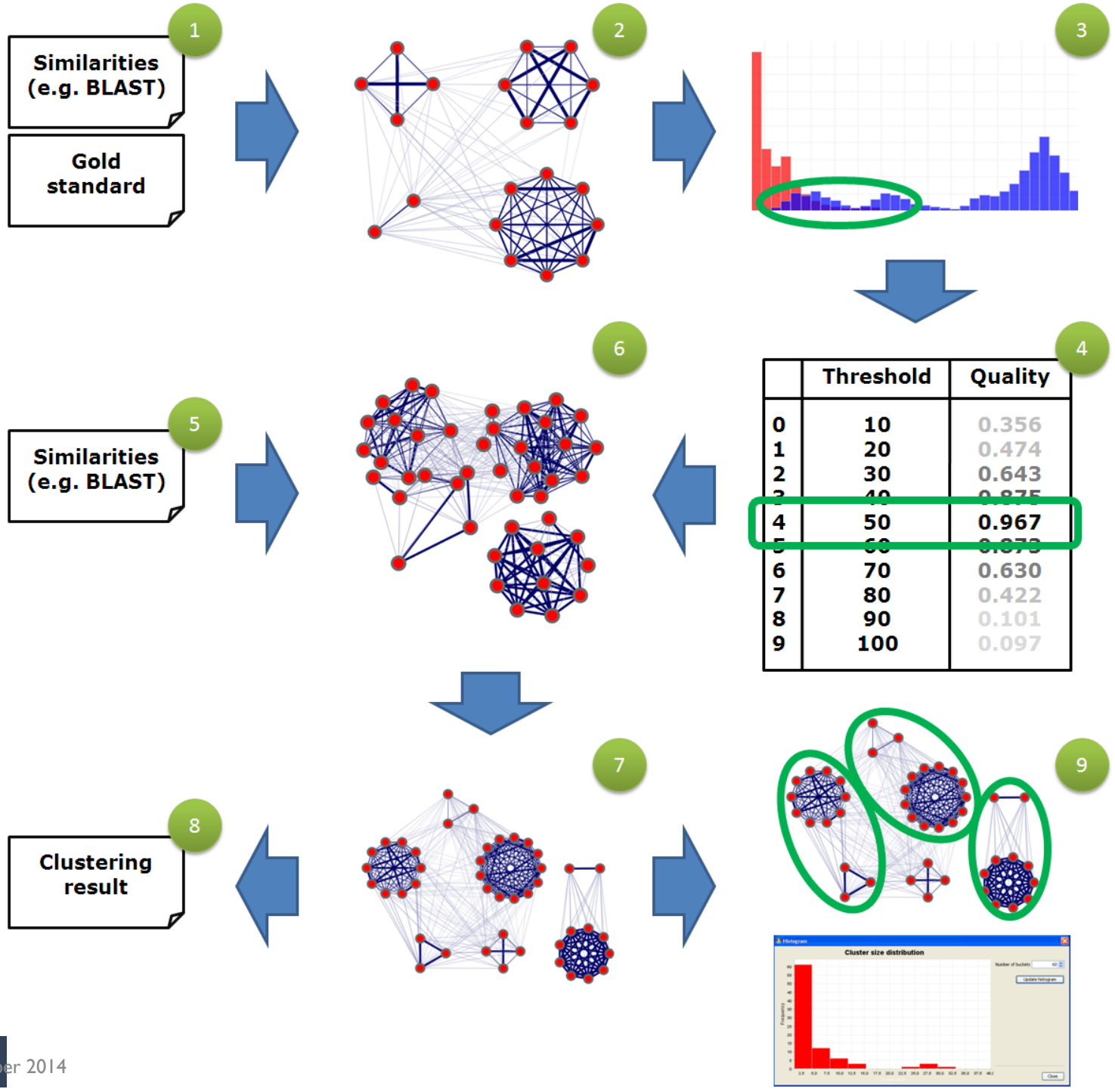
- **Amidohydrolase** is a large superfamily containing more than 20,000 member proteins

Case study: Amidohydrolases

- Problem: SFLD is incomplete
- How can we easily determine function of protein?



1) Cluster protein sequences
2) Use known functions from SFLD to validate clustering


1
Similarities (e.g. BLAST)
Gold standard
2
3
4
Threshold
Quality
0 10 0.356
1 20 0.474
2 30 0.643
4 50 0.967
6 70 0.630
7 80 0.422
8 90 0.101
9 100 0.097
5
Similarities (e.g. BLAST)
6
7
8
Clustering result
9
Cluster size distribution

Brown *et al.* gold standard of 866 enzymes assigned to (super) families

GS1 – Amidohydrolases superfamily:

232 proteins, 29 families

TS/GS2 – All other (four) superfamilies:

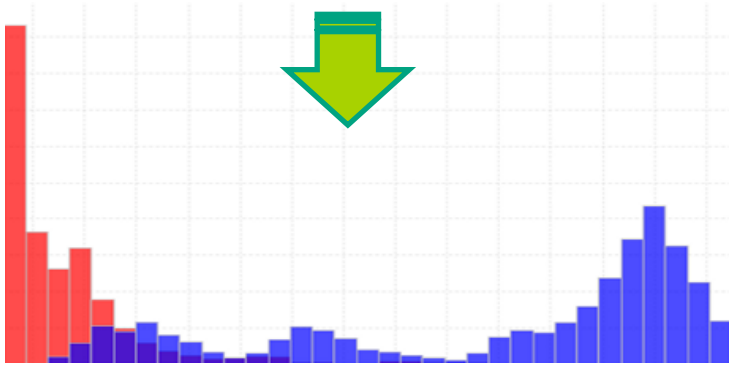
634 proteins, 62 families

Brown SD, Gerlt JA, Seffernick JL, Babbitt PC (2006) **A gold standard set of mechanistically diverse enzyme superfamilies.** *Genome Biol.* 2006;7(1):R8.

Brown *et al.* gold standard of 866 enzymes assigned to (super) families

GS1 – Amidohydrolases superfamily:

232 proteins, 29 families



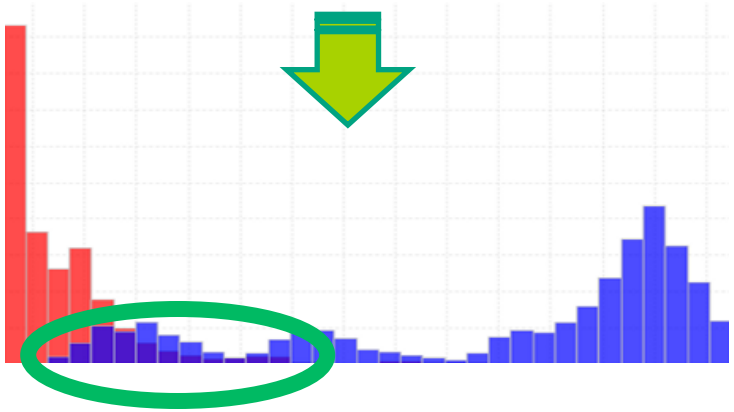
TS/GS2 – All other (four) superfamilies:

634 proteins, 62 families

Brown *et al.* gold standard of 866 enzymes assigned to (super) families

GS1 – Amidohydrolases superfamily:

232 proteins, 29 families



TS/GS2 – All other (four) superfamilies:

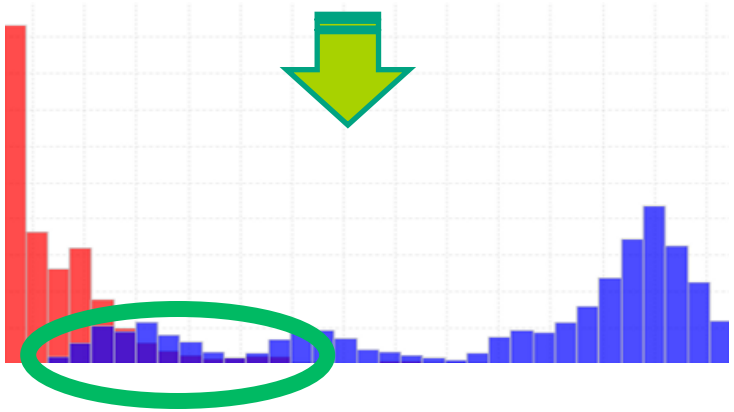
634 proteins, 62 families

Iterative clustering for thresholds between 45 to 75 → Best F-measure (0.97) with threshold 67

Brown *et al.* gold standard of 866 enzymes assigned to (super) families

GS1 – Amidohydrolases superfamily:

232 proteins, 29 families



Iterative clustering for thresholds between 45 to 75 → Best F-measure (0.97) with threshold 67

TS/GS2 – All other (four) superfamilies:

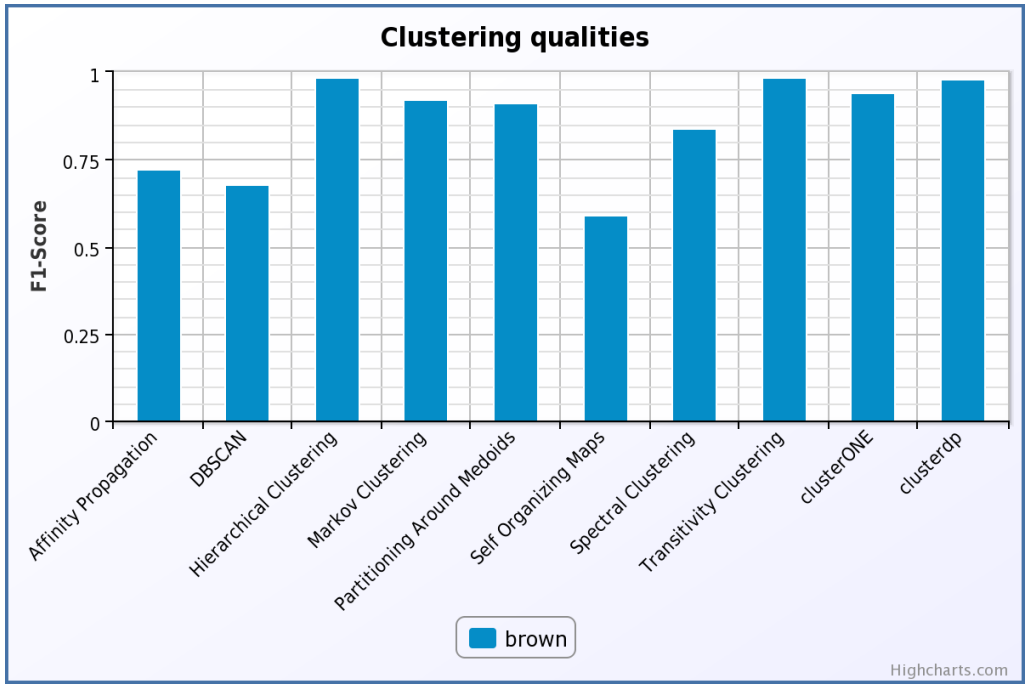
634 proteins, 62 families

Cluster TS/GS2 with threshold 67 →

F-measure: 0.84

Note: Best F-Measure (0.89) with threshold 58

Case study: Amidohydrolases



Method	F-Score
Hierarchical Clustering	0.987
Transitivity Clustering	0.986
clusterdp	0.979
clusterONE	0.942
Markov Clustering	0.923
PAM	0.912
Spectral Clustering	0.837
Affinity Propagation	0.724
DBSCAN	0.68
SOM	0.589

Case study: Amidohydrolases

Method	Silhouette Value
Hierarchical Clustering	0.806
Transitivity Clustering	0.805
clusterdp	0.795
Markov Clustering	0.772
PAM	0.764
clusterONE	0.74
Spectral Clustering	0.635
Affinity Propagation	0.483
DBSCAN	0.335
SOM	0.335

Case study II

Homology Detection
without Gold Standard

Case study: Actinobacteria

- One of the biggest clades of bacteria and with several genome sequences deposited (309 complete/3,419 drafts);
- High diversity throughout different lifestyles;
- Cope with a variety of different habitats;
- CMNR group.

Corynebacterium pseudotuberculosis
(Transmission electron microscopy).

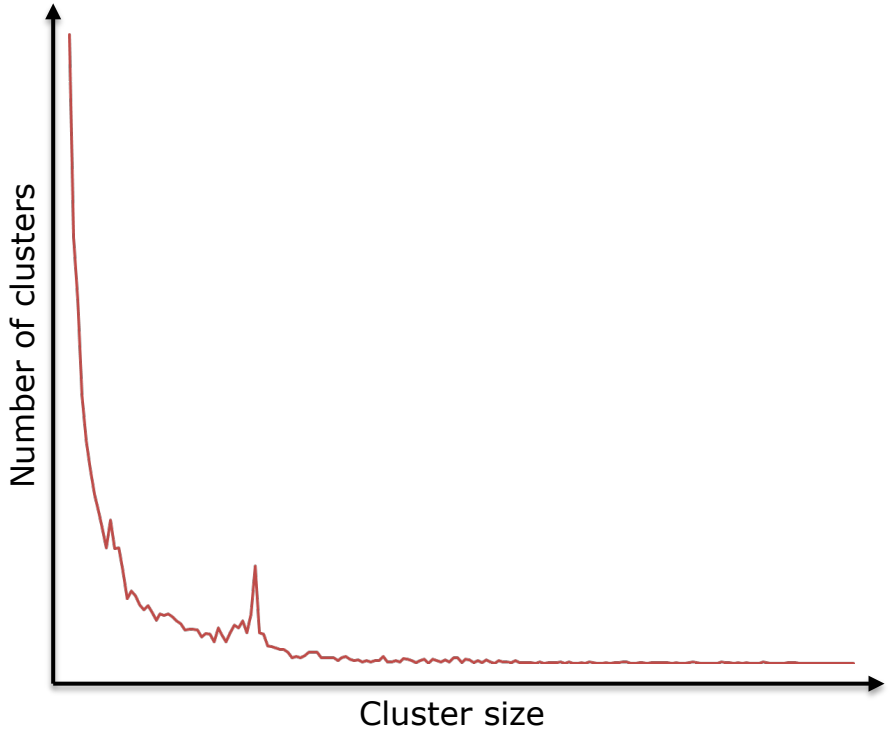
Case study: Actinobacteria

Problem/objectives

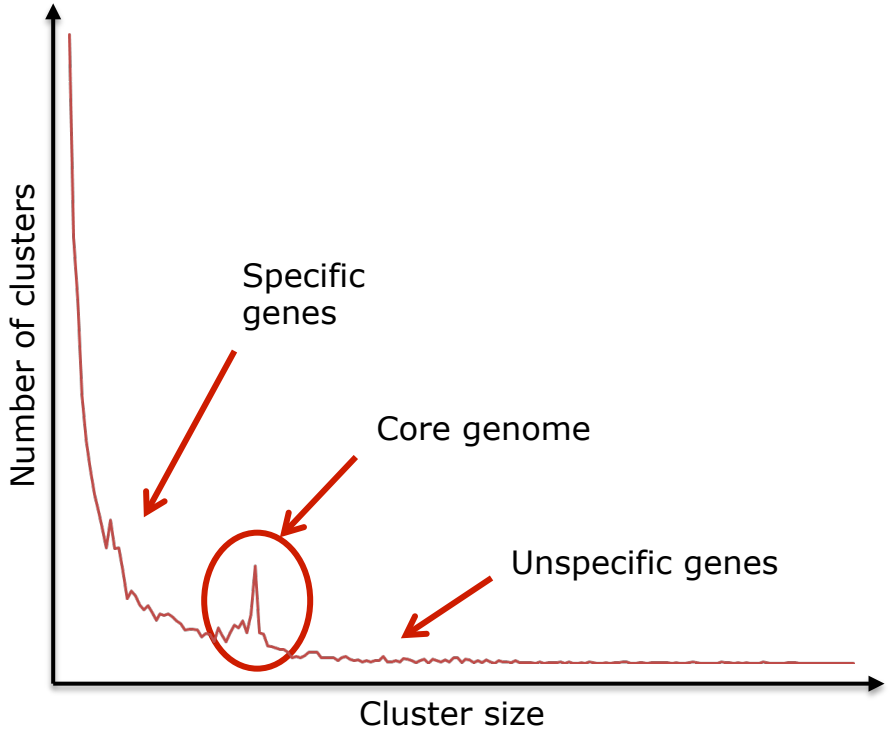
- Estimate a robust density parameter without gold standard data;

- Better understanding the genetic repertoire of bacteria with different lifestyles.

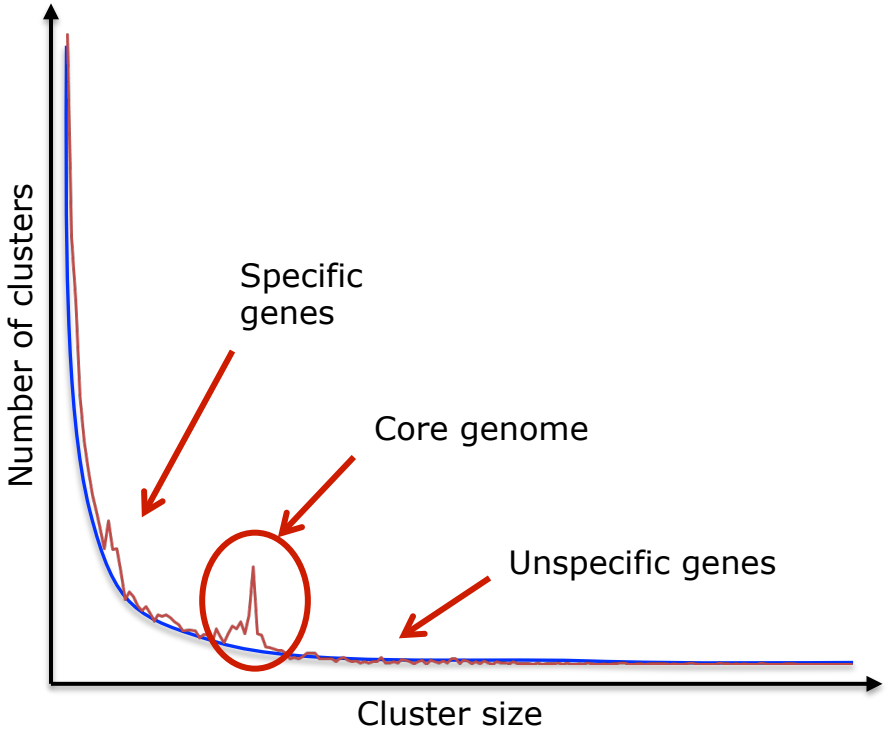
Case study: Actinobacteria



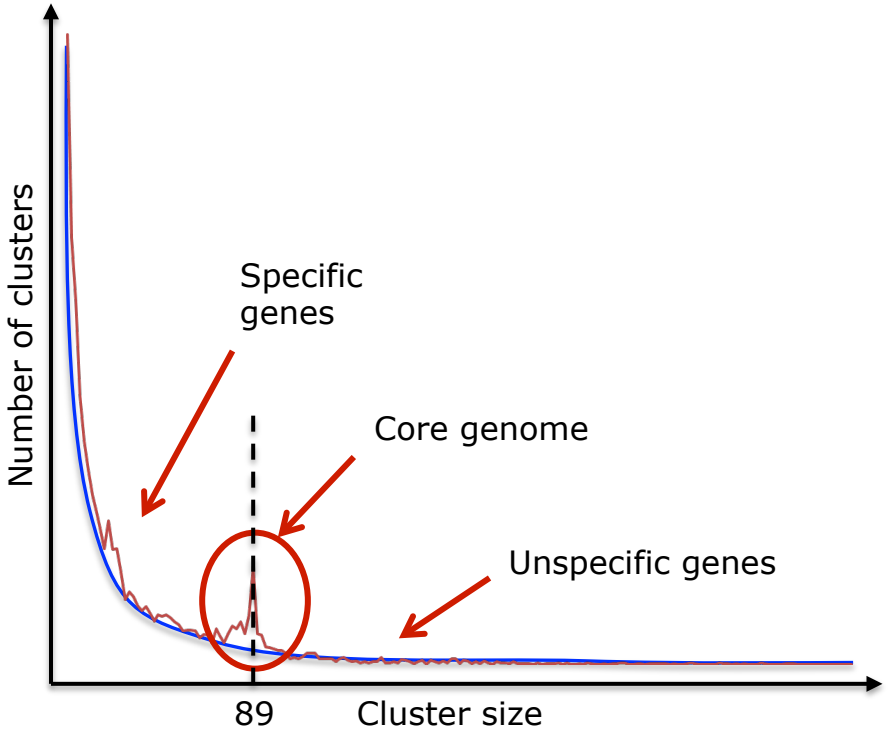
Case study: Actinobacteria



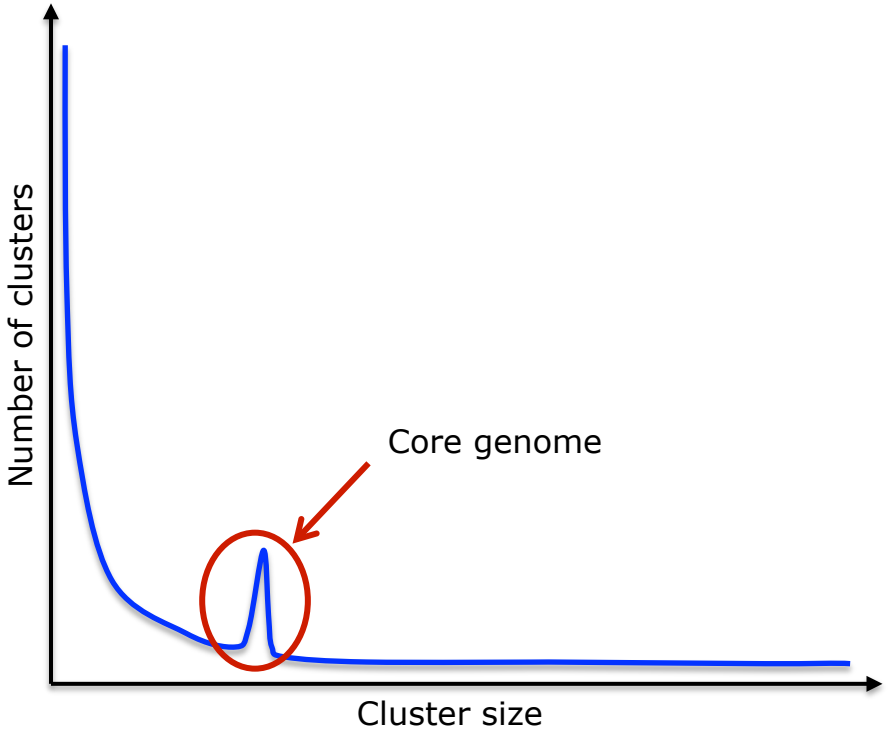
Case study: Actinobacteria



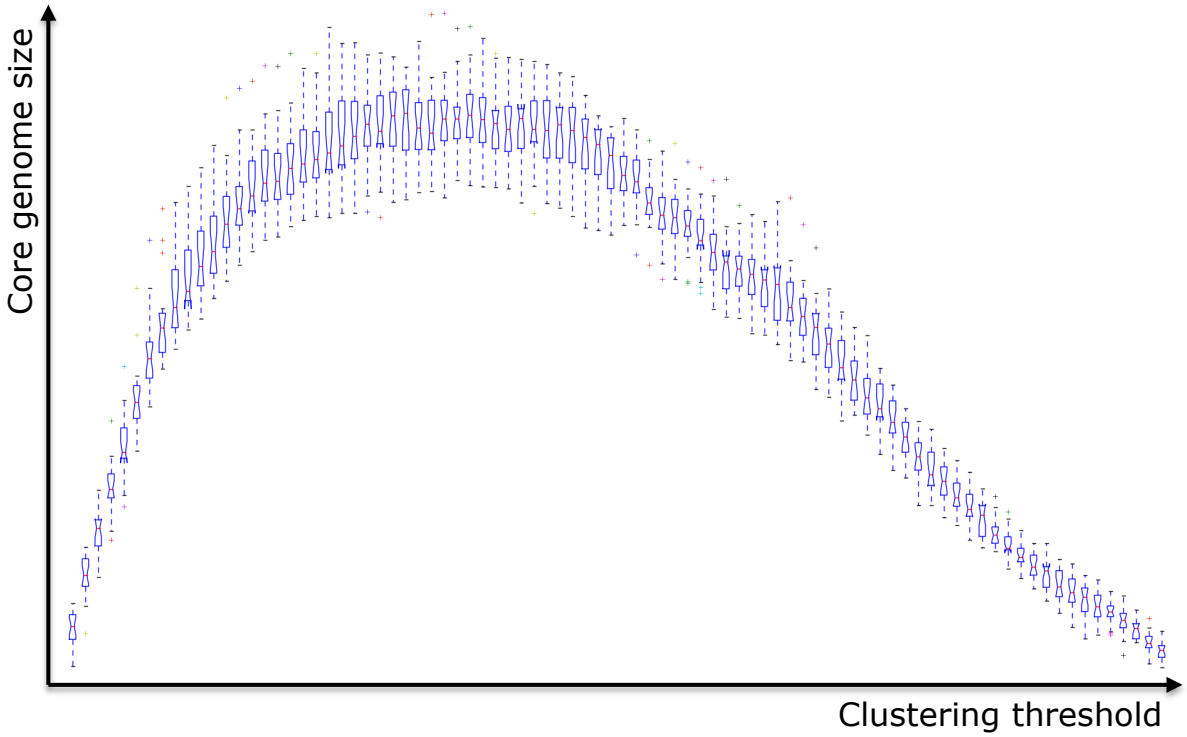
Case study: Actinobacteria



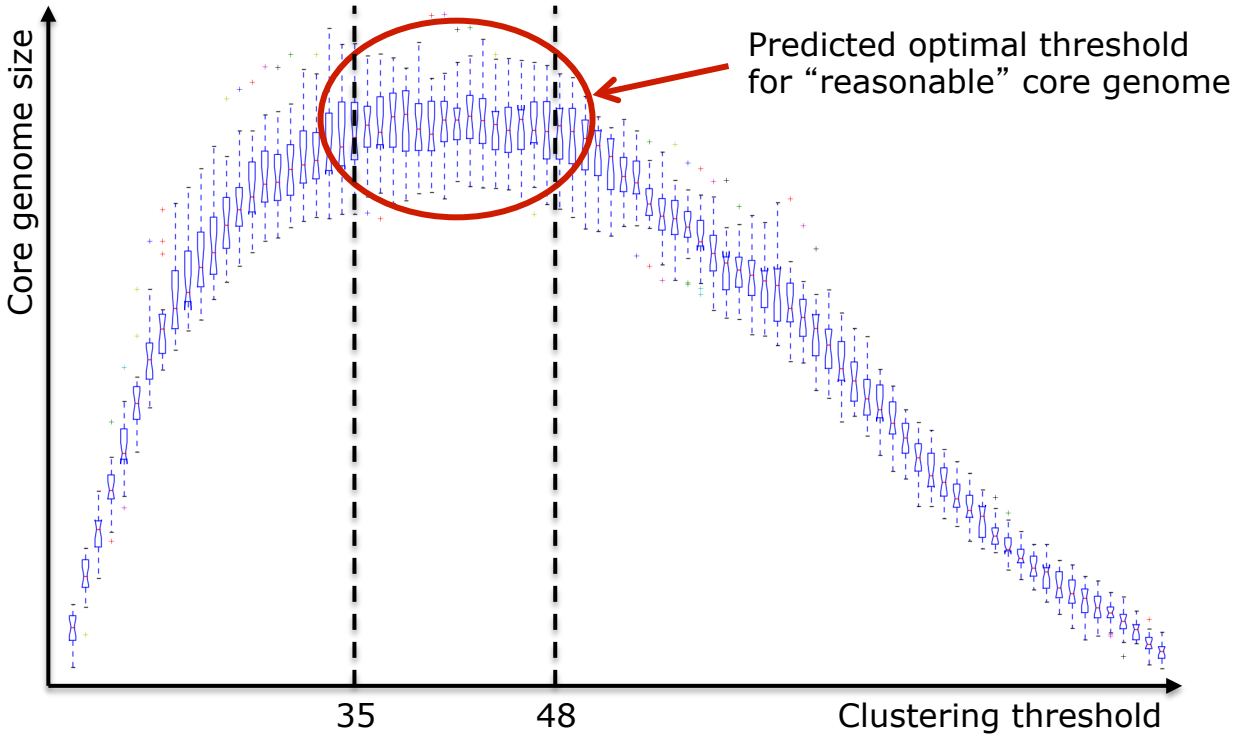
Case study: Actinobacteria



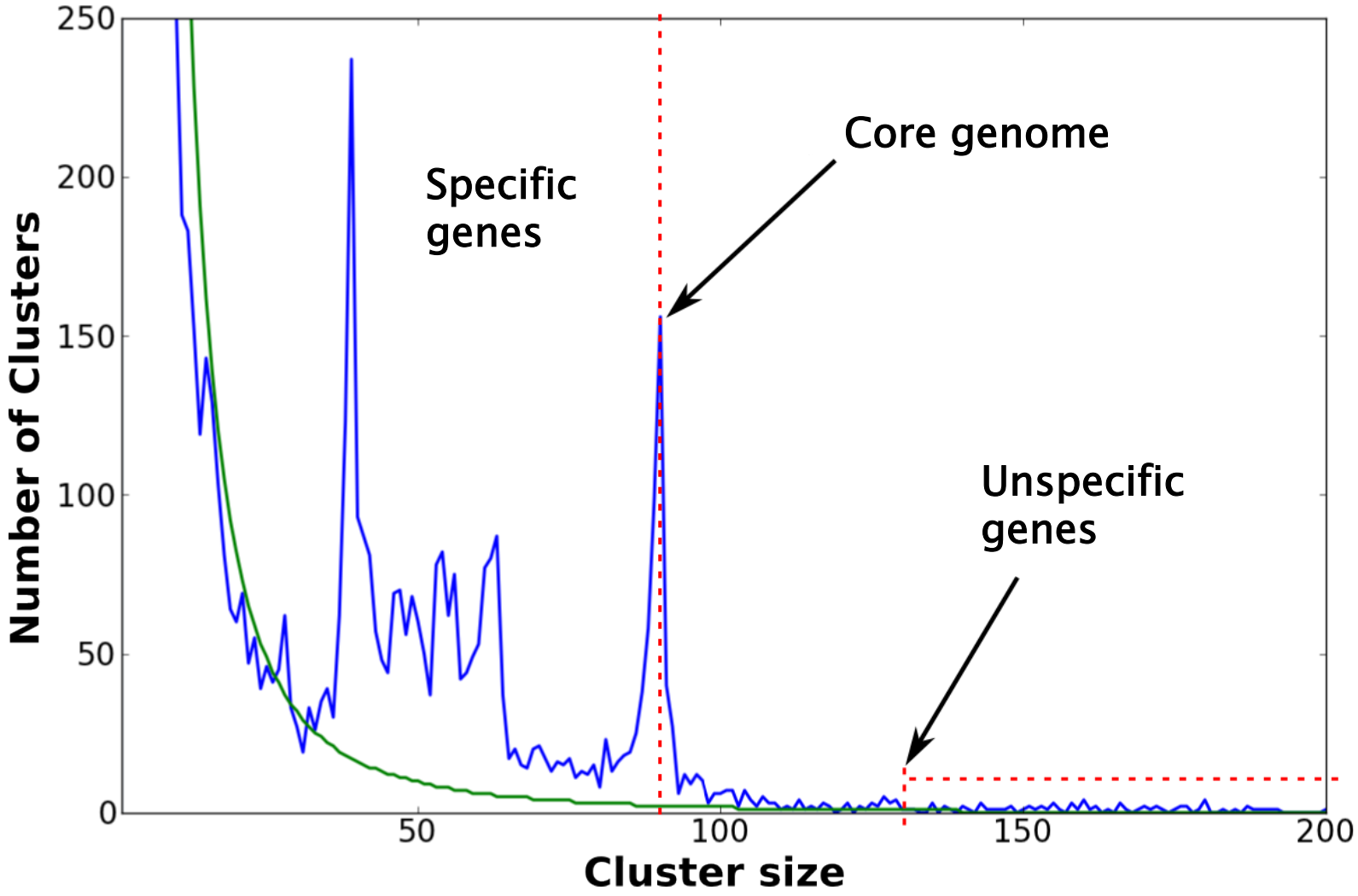
Case study: Actinobacteria



Case study: Actinobacteria



Case study: Actinobacteria



Cluster size distribution of the 89 actinobacteria for similarity threshold 48.

Clustering with missing values

Clustering with missing values

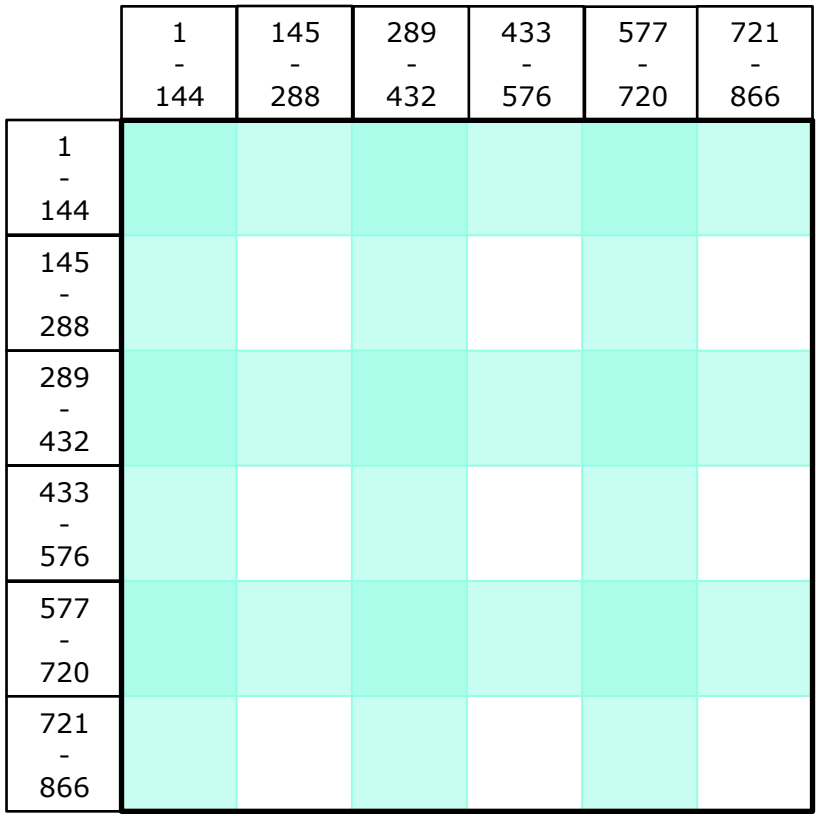
Situation

- Data not complete;
- Performance improvement.

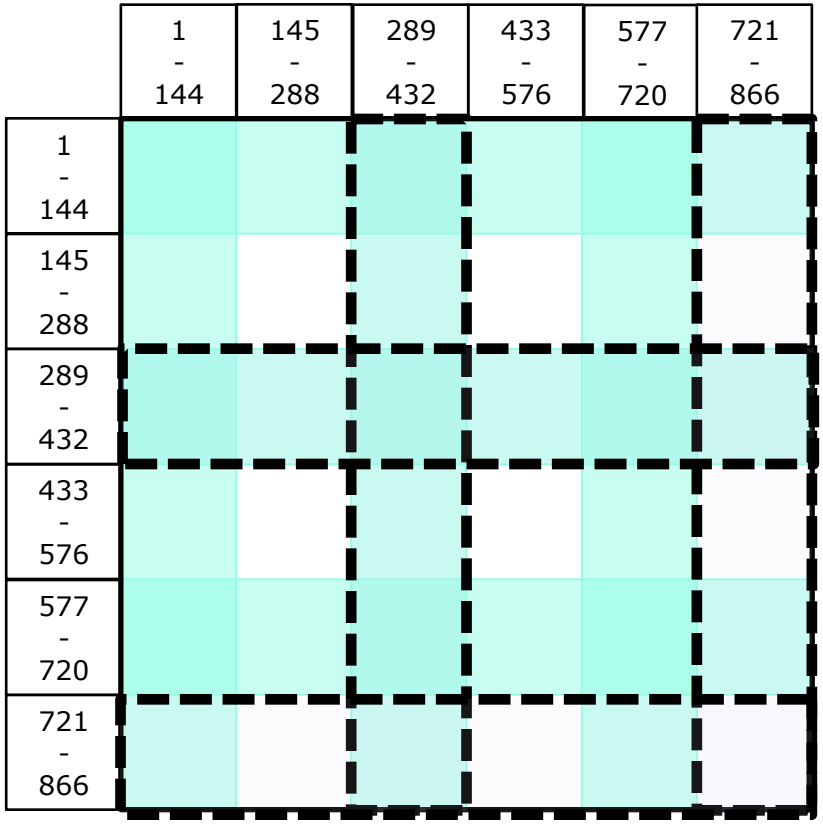
Common solutions

- Fill in the missing values (imputation);
- Ignore the missing data (marginalization).

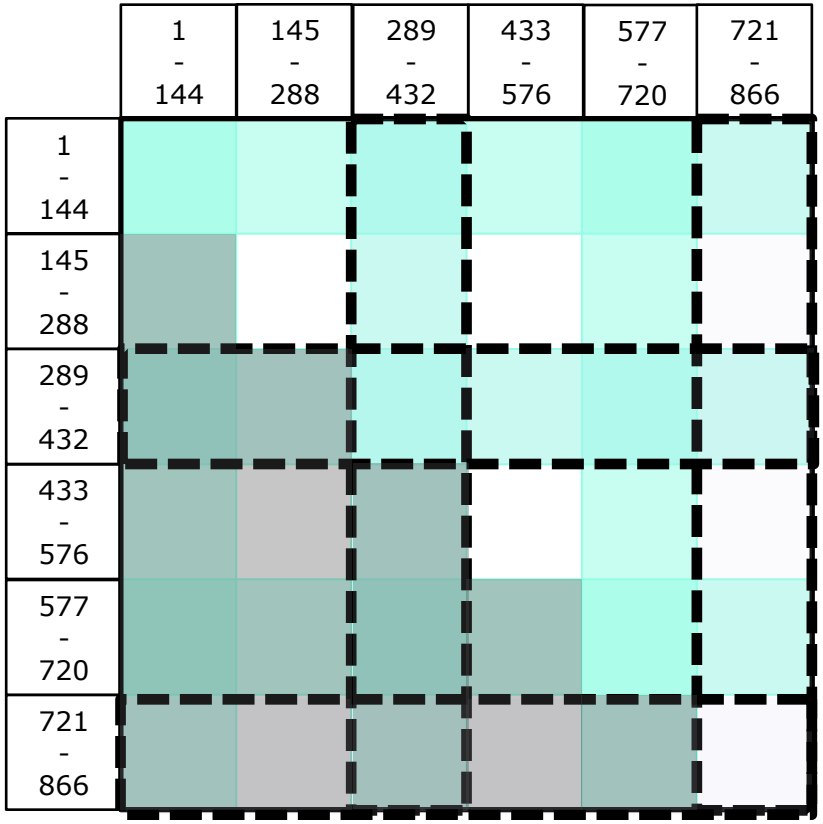
Clustering with missing values



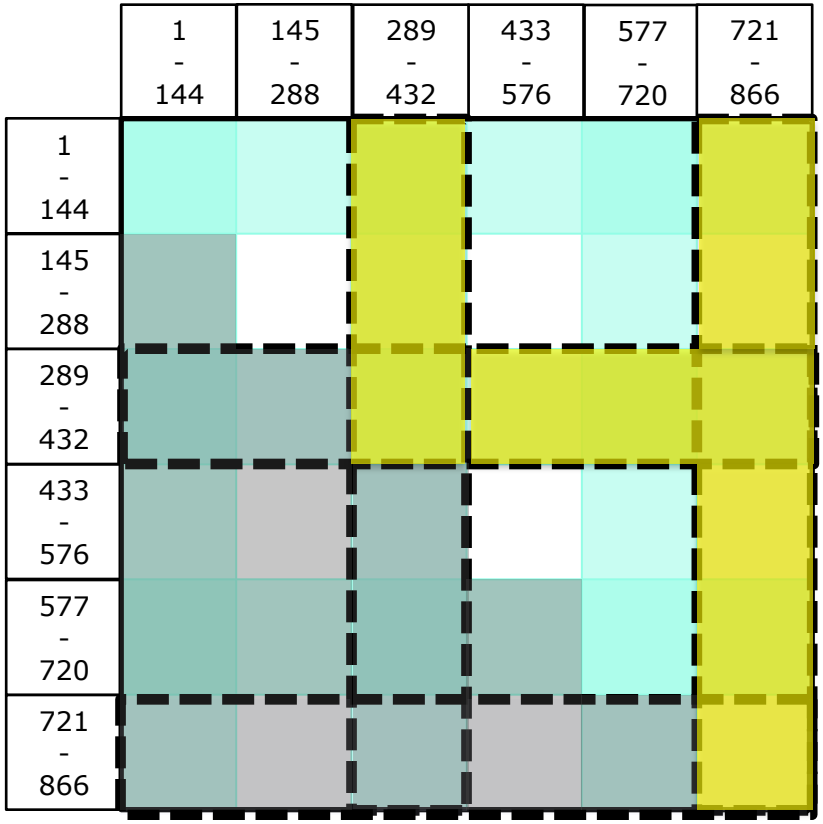
Clustering with missing values



Clustering with missing values



Clustering with missing values



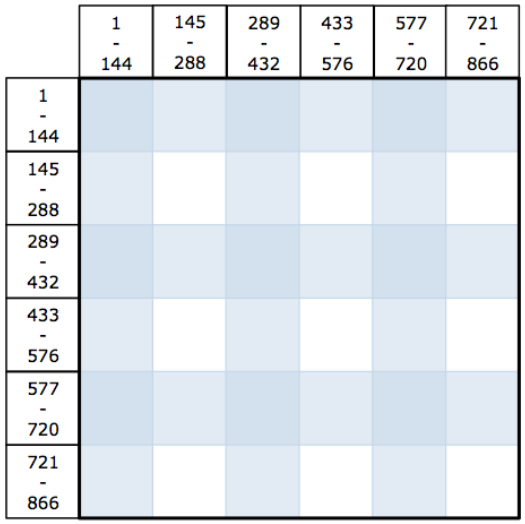
Clustering with missing values

	1 - 144	145 - 288	289 - 432	433 - 576	577 - 720	721 - 866
1 - 144	?	?	231 380 470	?	?	111 481 199
145 - 288		?	136 300 539	?	?	276 831 120
289 - 432			162 810 131	126 800 119	049 289 039	926 835 129
433 - 576				?	?	026 380 237
577 - 720					?	122 188 140
721 - 866						206 863 509

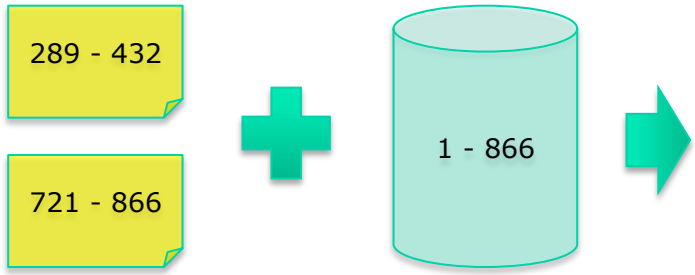
Clustering with missing values

	1 - 144	145 - 288	289 - 432	433 - 576	577 - 720	721 - 866
1 - 144	?	?	231 380 470	?	?	111 481 199
145 - 288		?	136 300 539	?	?	276 831 120
289 - 432			162 810 131	126 800 119	049 289 039	926 835 129
433 - 576				?	?	026 380 237
577 - 720					?	122 188 140
721 - 866						206 863 509

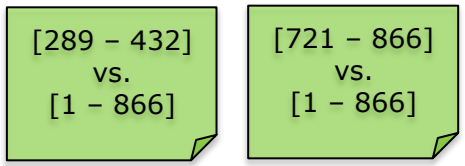
Split data set (here: six buckets)



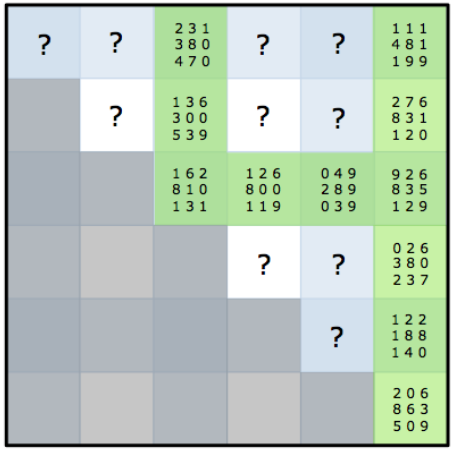
BLAST all against selected blocks



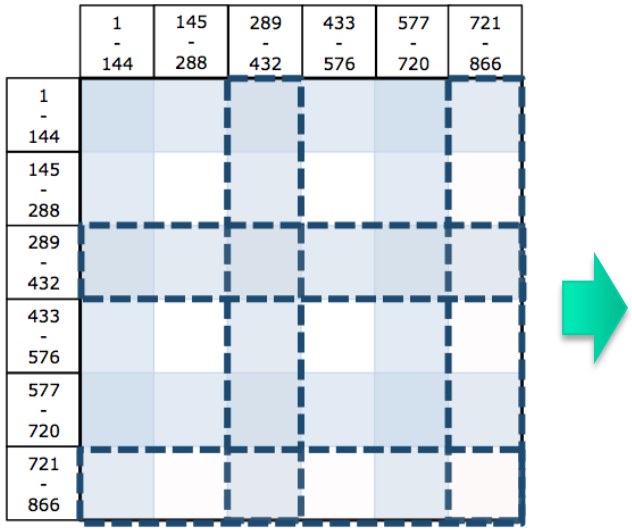
BLAST-based similarity blocks



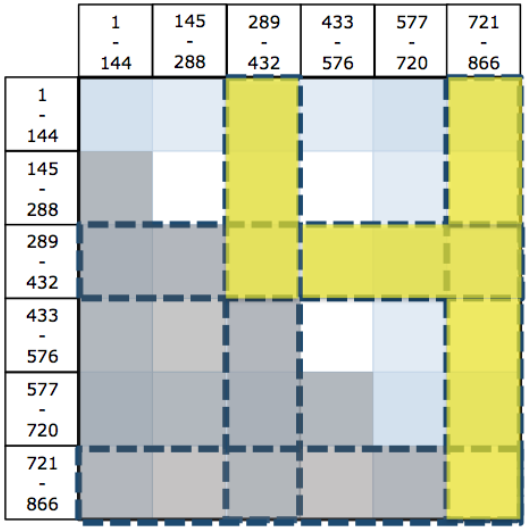
Block similarity matrix with "?" entries



Block selection (coverage: 33%)

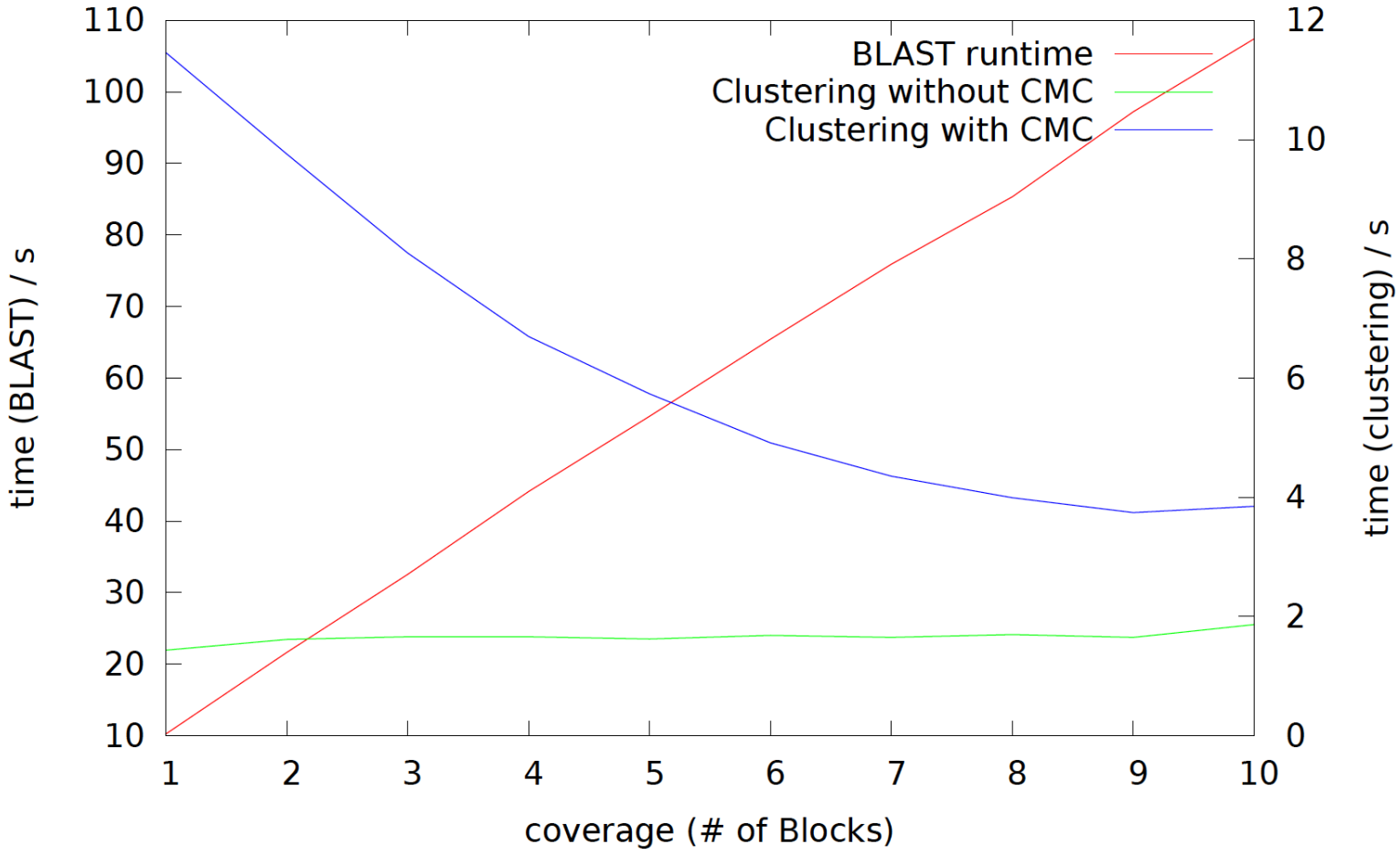


Blocks for similarity computation

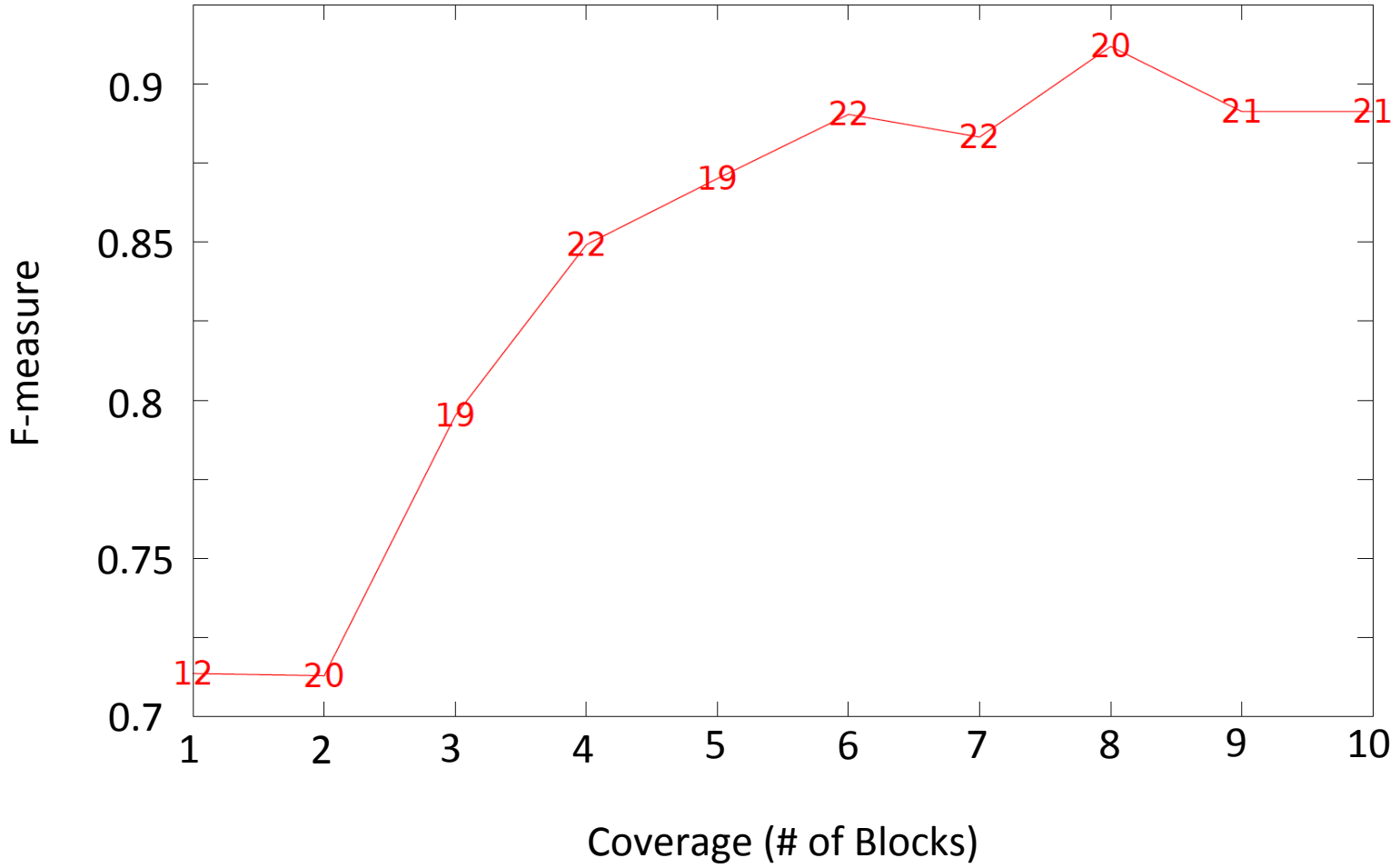


TransClust

Clustering with missing values



Clustering with missing values



Thank you!