Dealing with Sequence redundancy

Morten Nielsen
Department of Health Technology,
DTU

Outline

- What is data redundancy?
- Why is it a problem?
- How can we deal with it?

Databases are redundant

- Biological reasons
  - Some protein functions, or sequence motifs are more common than others
- Laboratory artifacts
  - Some protein families have been heavily investigated, others not
  - Mutagenesis studies makes large and almost identical replica of data
  - This bias is non-biological

Why is it important

- If you have high redundant data
  - and the redundancy is artificial, you will learn something non-biological
  - A machine learning method could focus on the largest class, and might never learn the minority patterns
  - If you have redundancy between the data used for model development and evaluation, the “trained” model could become a look-up table with limited power to generalize

Date redundancy

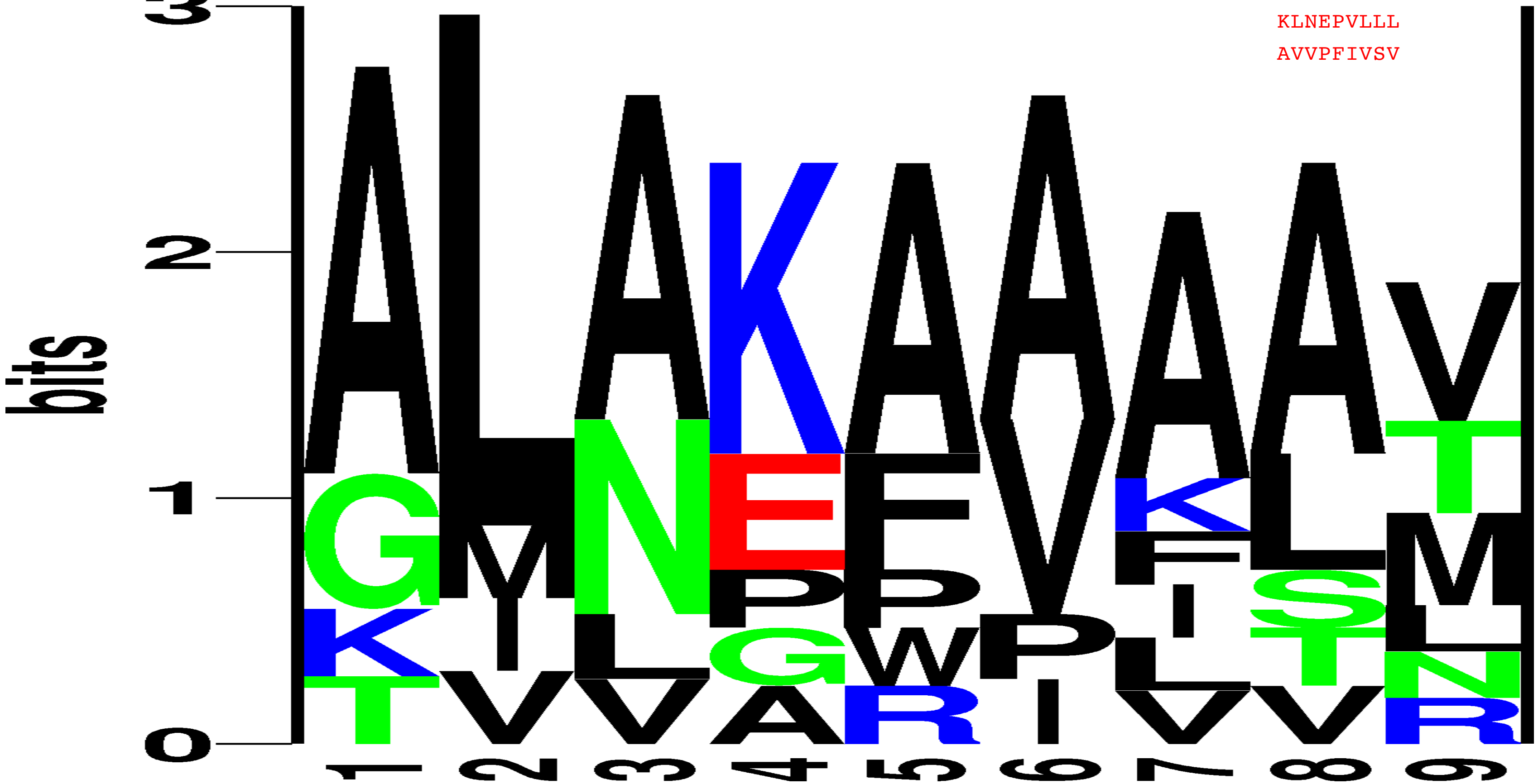
10 MHC restricted peptides

```
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV
```

What can we learn?

1. A at P1 favors binding?
2. I is not allowed at P9?
3. K at P4 favors binding?
4. Which positions are important for binding?

Redundant data

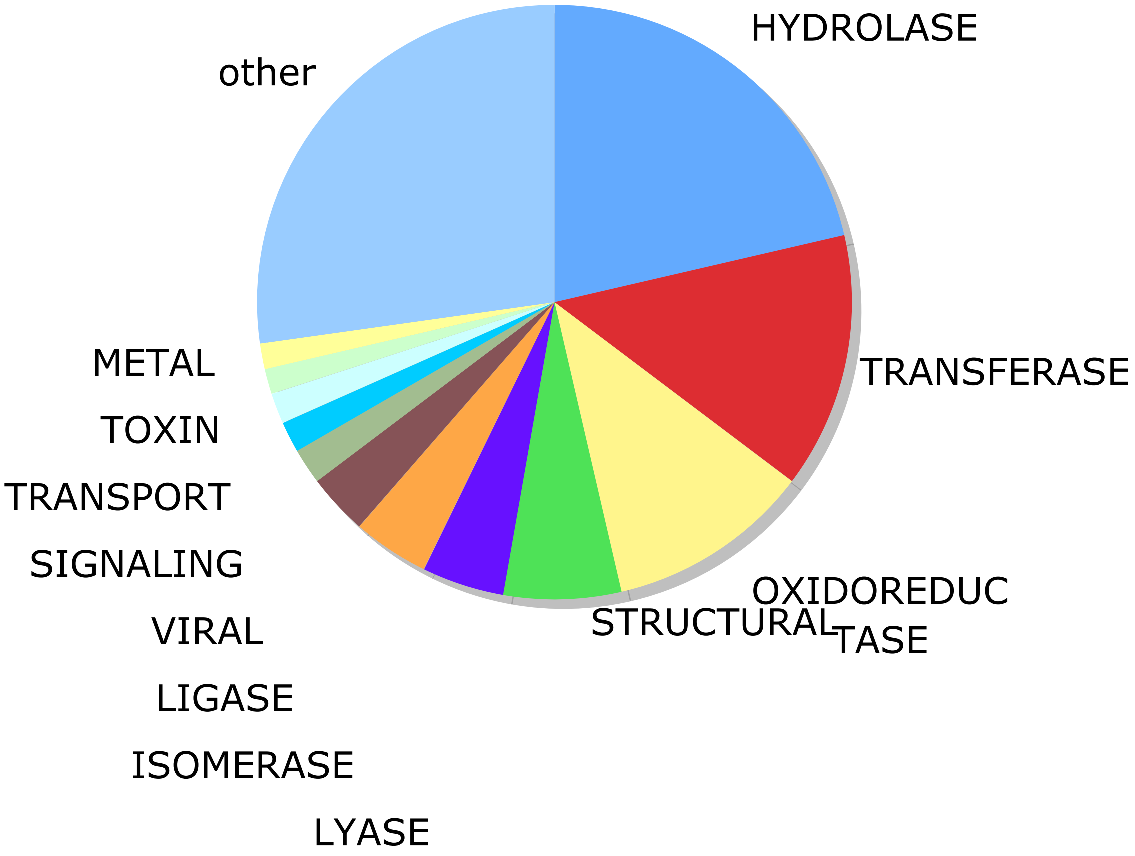


ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

PDB. Example

- 1055 protein sequence
- Len 50-2000
- 142 Function annotations
  - ACTIN-BINDING
  - ANTIGEN
  - COAGULATION
  - HYDROLASE/DNA
  - LYASE/OXIDOREDUCTASE
  - ENDOCYTOSIS/EXOCYTOSIS
  - …

PDB. Example


HYDROLASE
TRANSFERASE
OXIDOREDUCTASE
STRUCTURAL
LYASE
ISOMERASE
LIGASE
VIRAL
SIGNALING
TRANSPORT
TOXIN
METAL
other

What is similarity?

- Sequence identity?

```
ACDFG
ACEFG
```

80% ID versus 24% ID

```
DFLKKVPDDHLEFIPYLILGEVFPEWDERELGVGEKLLIKAVA------------MATGIDAKEIEESVKDTGDL-GE
DVLLGADDGSLAFVP---------- SEFSISPGEKIVFKNNAGFPHNIVFDEDSIPSGVDASKISMSEEDLLNAKGE
```

- Blast e-values
  - Often too conservative
- Other

Ole Lund et al.

(Protein engineering 1997)

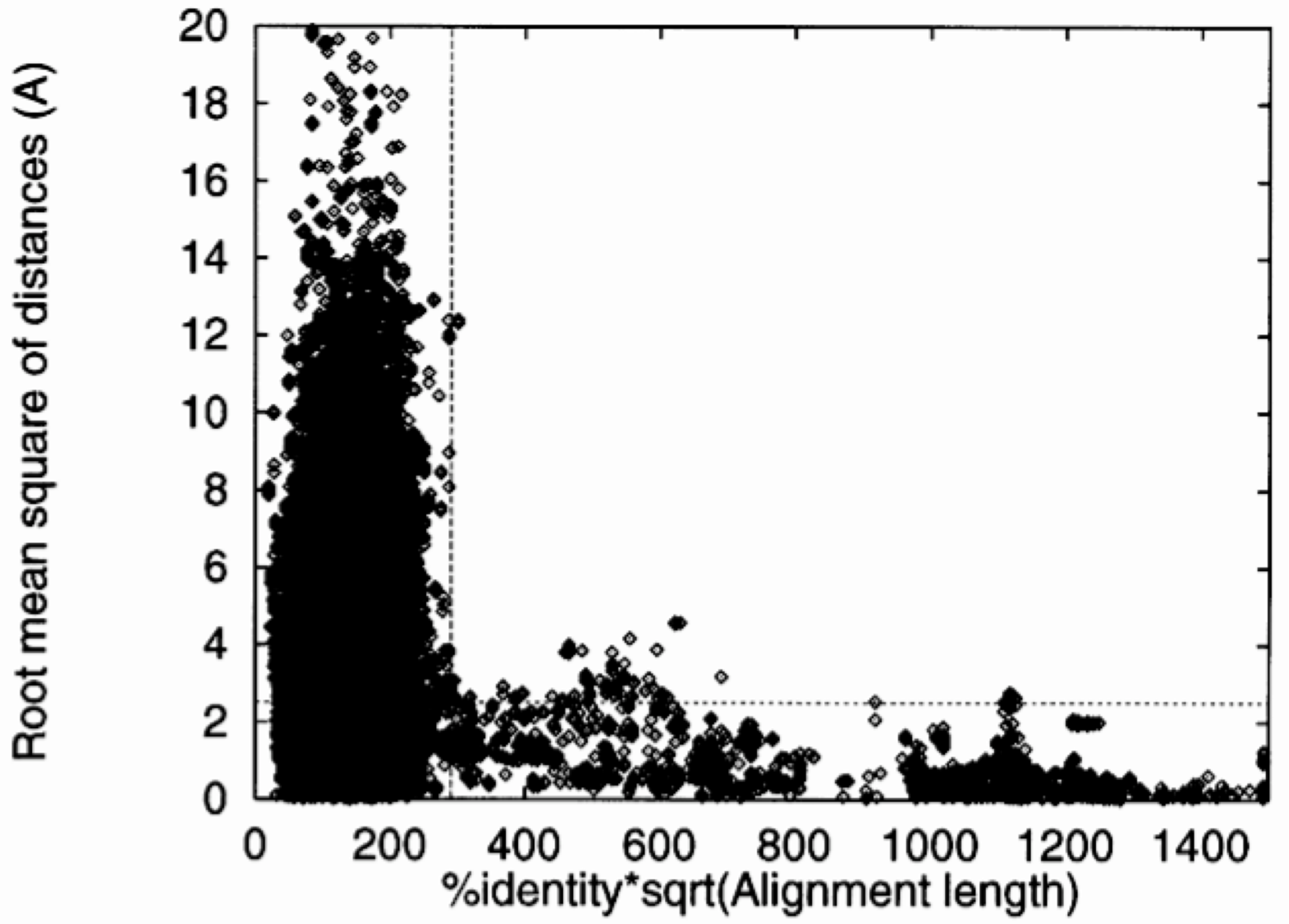


Fig. 2. Root mean square of distances of equivalent C$^{\alpha}$ atoms in the alignments of 942 sequences as a function of $\sqrt{LI_{seq}}$. The vertical line corresponds to the sequence-similarity-implies-structural-similarity threshold $\sqrt{LI_{seq}} = 290$ and the horizontal line is at 2.5 Å.

Ole’s formula

$$\%Id \cdot \sqrt{alen} > 290$$

$$\%Id > \frac{290}{\sqrt{alen}}$$

$$fid > \frac{2.9}{\sqrt{alen}}$$

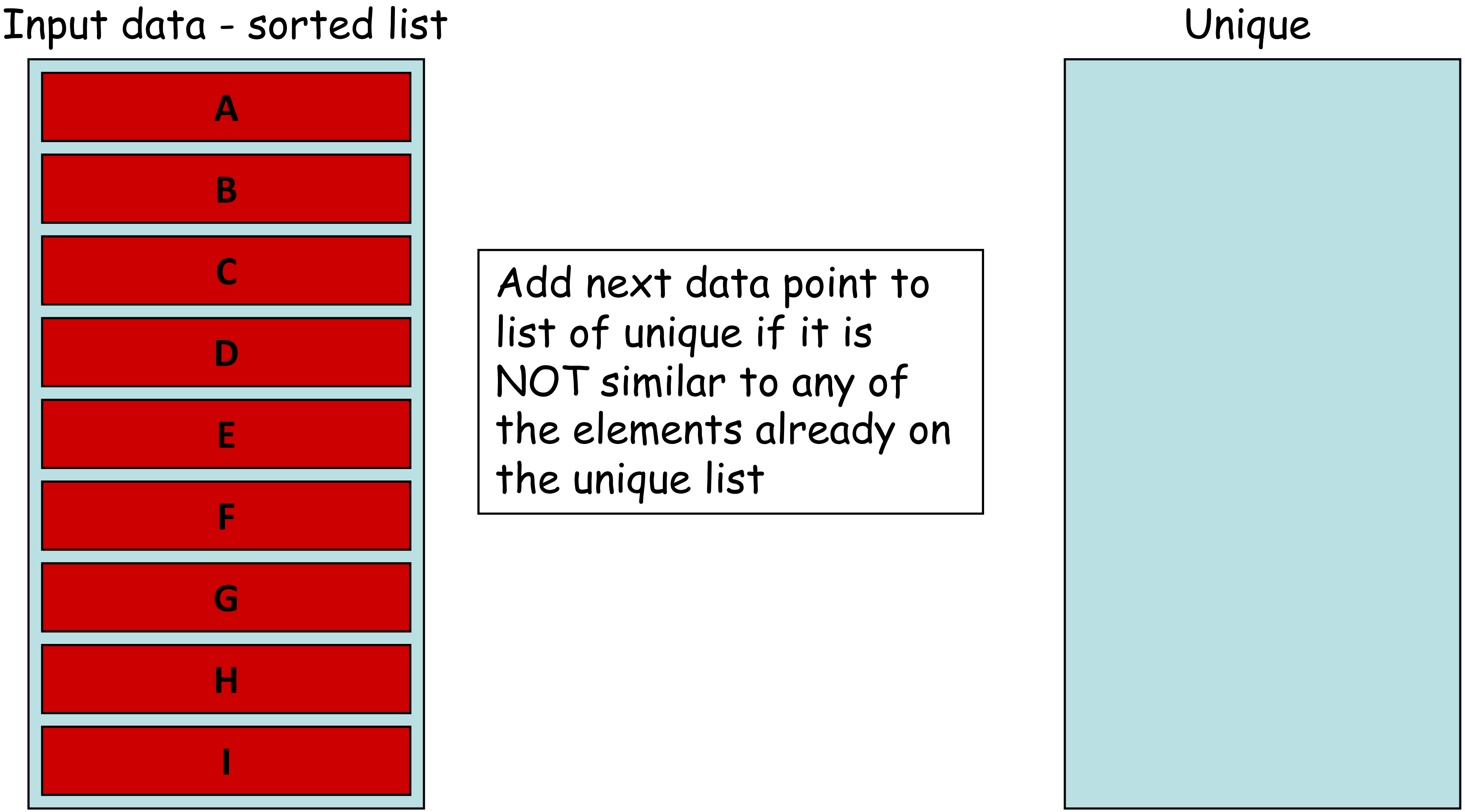
$$\boxed{Nid = fid \cdot alen > 2.9 \cdot \sqrt{alen}}$$

Note, this formula is only relevant for protein sequences

How to deal with redundancy

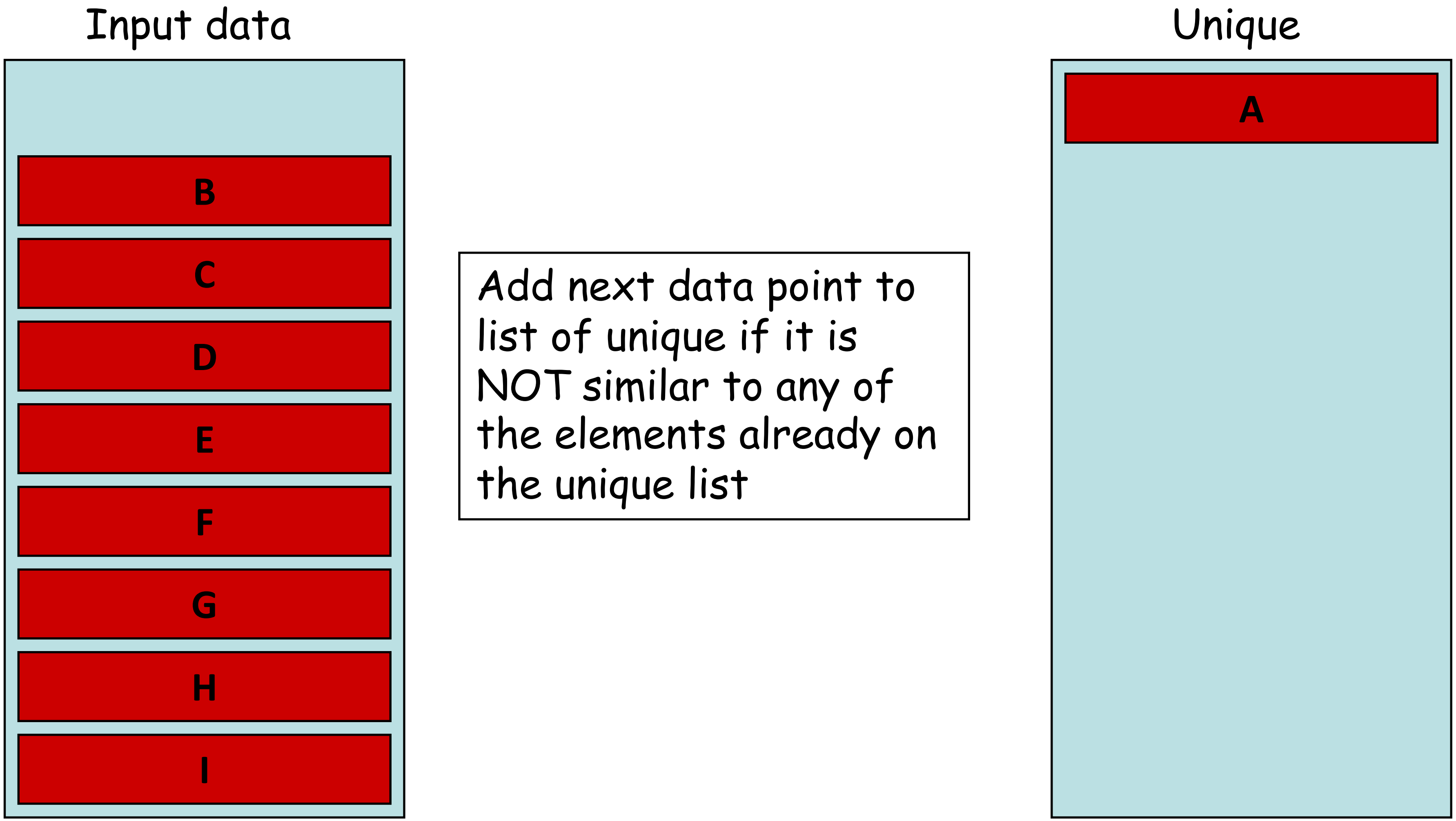
- Hobohm 1
  - Fast
  - Requires a prior sorting of data
- Hobohm 2
  - Slow
  - Gives unique answer always
  - No prior sorting

Hobohm 1



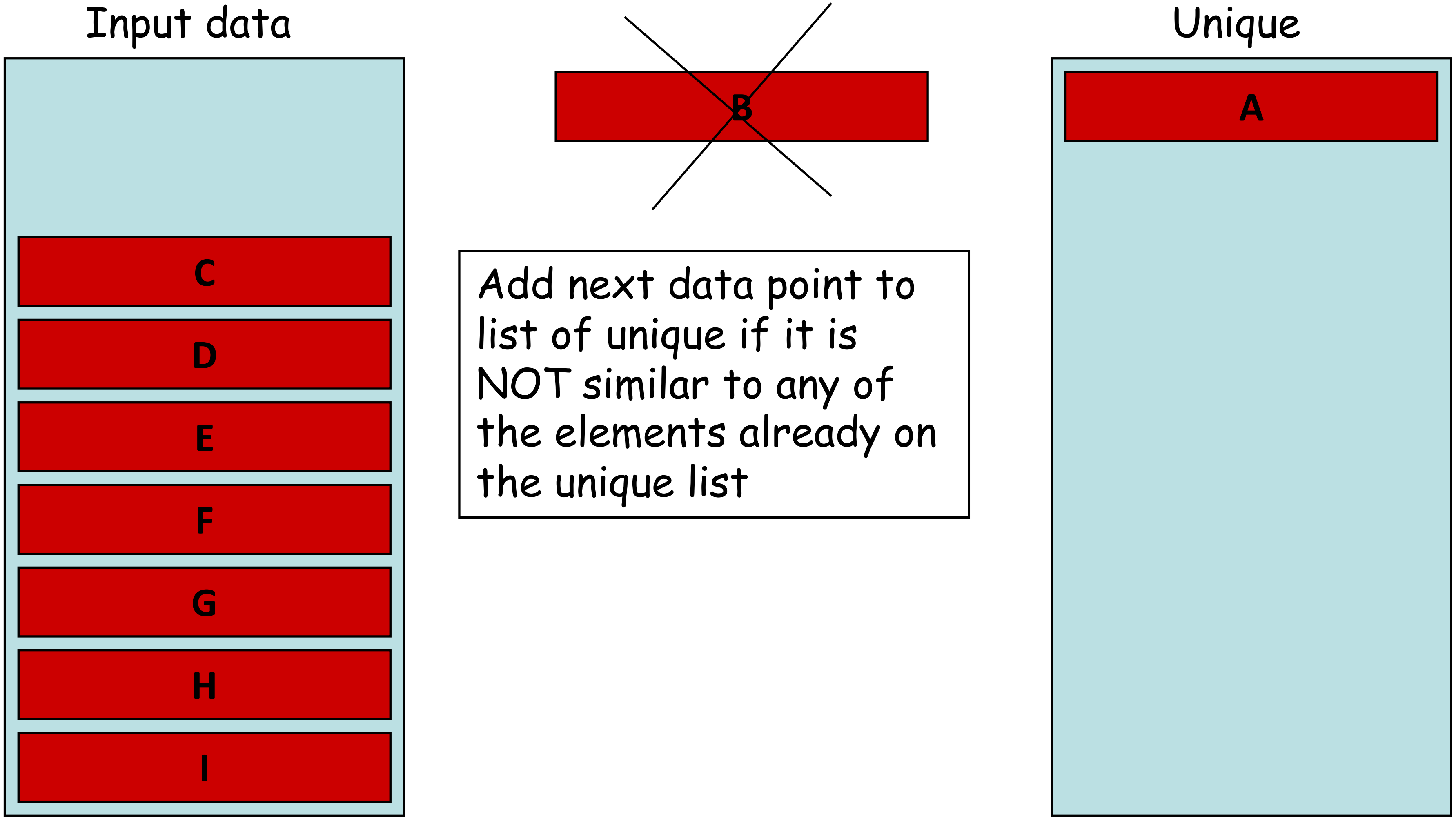
Add next data point to list of unique if it is NOT similar to any of the elements already on the unique list

Hobohm 1



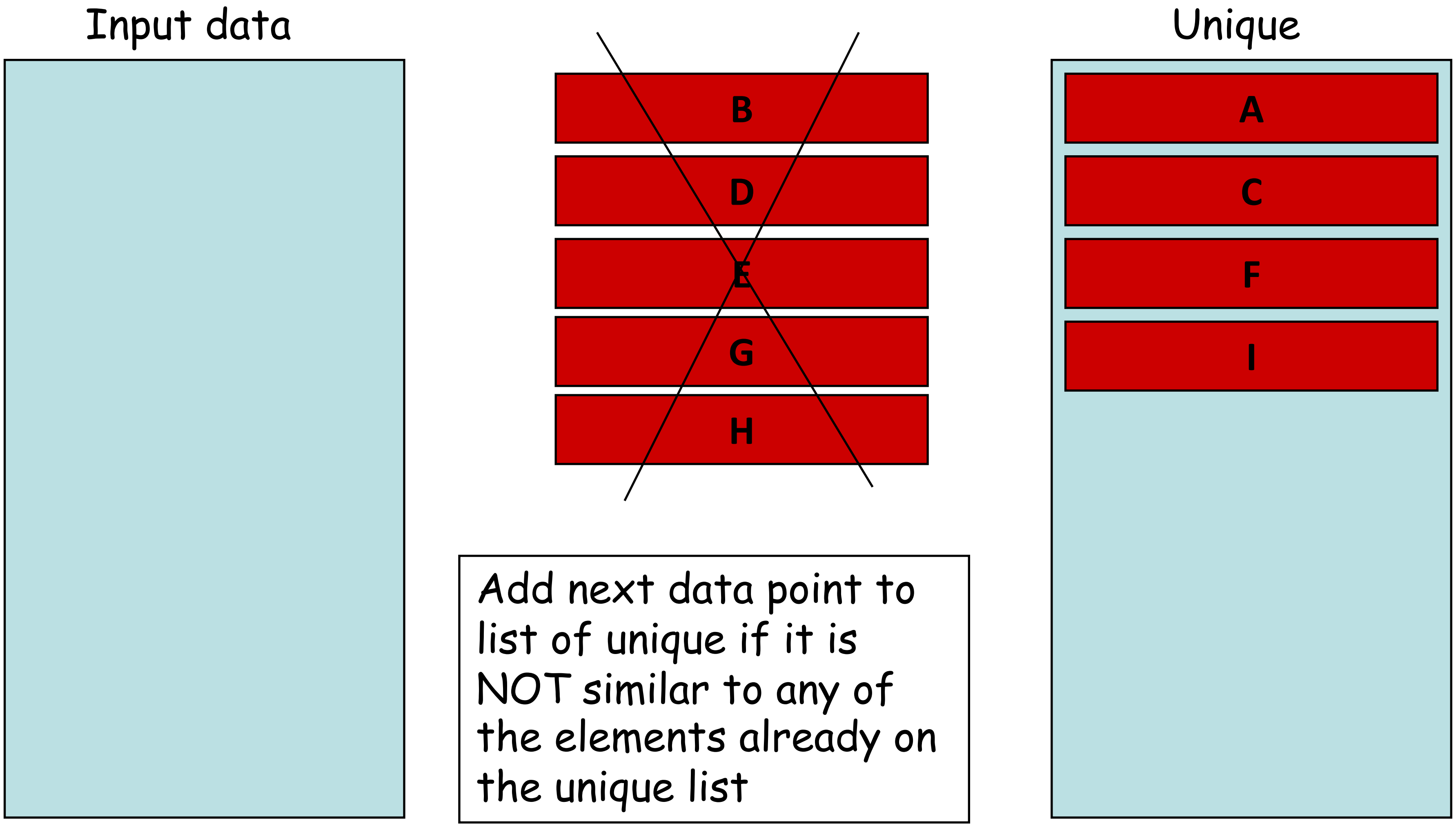
Add next data point to list of unique if it is NOT similar to any of the elements already on the unique list

Hobohm 1



Add next data point to list of unique if it is NOT similar to any of the elements already on the unique list

Hobohm 1



Add next data point to list of unique if it is NOT similar to any of the elements already on the unique list

Need only to align sequences against the Unique list!

Hobohm-2

- Align all against all
- Make similarity matrix D (N*N) with value 1 if is similar to j, otherwise 0
- While data points have more than one neighbor
  - Remove data point S with most nearest neighbors

Hobohm-2

D:

	A	B	C	D	E	F	G	H	I
A	1	1	1	0	0	0	0	0	0
B	1	1	1	0	0	0	0	1	1
C	1	1	1	0	0	0	0	0	0
D	0	0	0	1	1	1	1	1	1
E	0	0	0	1	1	1	1	1	1
F	0	0	0	1	1	1	0	0	1
G	0	0	0	1	1	0	1	1	1
H	0	1	0	1	1	0	1	1	1
I	0	1	0	1	1	1	1	1	1

Make similarity matrix N*N

Hobohm-2

D:

		A	B	C	D	E	F	G	H	I	N
	A	1	1	1	0	0	0	0	0	0	3
	B	1	1	1	0	0	0	0	1	1	5
	C	1	1	1	0	0	0	0	0	0	3
	D	0	0	0	1	1	1	1	1	1	6
	E	0	0	0	1	1	1	1	1	1	6
	F	0	0	0	1	1	1	0	0	1	4
	G	0	0	0	1	1	0	1	1	1	5
	H	0	1	0	1	1	0	1	1	1	6
S	I	0	1	0	1	1	1	1	1	1	7

Find point S with the largest number of similarities

Hobohm-2

D:

	A	B	C	D	E	F	G	H	I
A	1	1	1	0	0	0	0	0	0
B	1	1	1	0	0	0	0	1	1
C	1	1	1	0	0	0	0	0	0
D	0	0	0	1	1	1	1	1	1
E	0	0	0	1	1	1	1	1	1
F	0	0	0	1	1	1	0	0	1
G	0	0	0	1	1	0	1	1	1
H	0	1	0	1	1	0	1	1	1
I	0	1	0	1	1	1	1	1	1

N
3
5
3
6
6
4
5
6
7

D:

	A	B	C	D	E	F	G	H
A	1	1	1	0	0	0	0	0
B	1	1	1	0	0	0	0	1
C	1	1	1	0	0	0	0	0
D	0	0	0	1	1	1	1	1
E	0	0	0	1	1	1	1	1
F	0	0	0	1	1	1	0	0
G	0	0	0	1	1	0	1	1
H	0	1	0	1	1	0	1	1

N
3
4
3
5
5
3
4
5

Remove point S with the largest number of similarities, and update N counts

Hobohm-2 (repeat this)

D:

	A	B	C	D	E	F	G	H	N
A	1	1	1	0	0	0	0	0	3
B	1	1	1	0	0	0	0	1	4
C	1	1	1	0	0	0	0	0	3
D	0	0	0	1	1	1	1	1	5
E	0	0	0	1	1	1	1	1	5
F	0	0	0	1	1	1	0	0	3
G	0	0	0	1	1	0	1	1	4
H	0	1	0	1	1	0	1	1	5

D:

	A	B	C	E	F	G	H	N
A	1	1	1	0	0	0	0	3
B	1	1	1	0	0	0	1	4
C	1	1	1	0	0	0	0	3
E	0	0	0	1	1	1	1	4
F	0	0	0	1	1	0	0	2
G	0	0	0	1	0	1	1	3
H	0	1	0	1	0	1	1	4

Remove point S with the largest number of similarities

Hobohm-2 (until N=1 for all)

D:

	A	B	C	D	E	F	G	H	I
A	1	1	1	0	0	0	0	0	0
B	1	1	1	0	0	0	0	1	1
C	1	1	1	0	0	0	0	0	0
D	0	0	0	1	1	1	1	1	1
E	0	0	0	1	1	1	1	1	1
F	0	0	0	1	1	1	0	0	1
G	0	0	0	1	1	0	1	1	1
H	0	1	0	1	1	0	1	1	1
I	0	1	0	1	1	1	1	1	1

N
3
5
3
6
6
4
5
6
7

=>

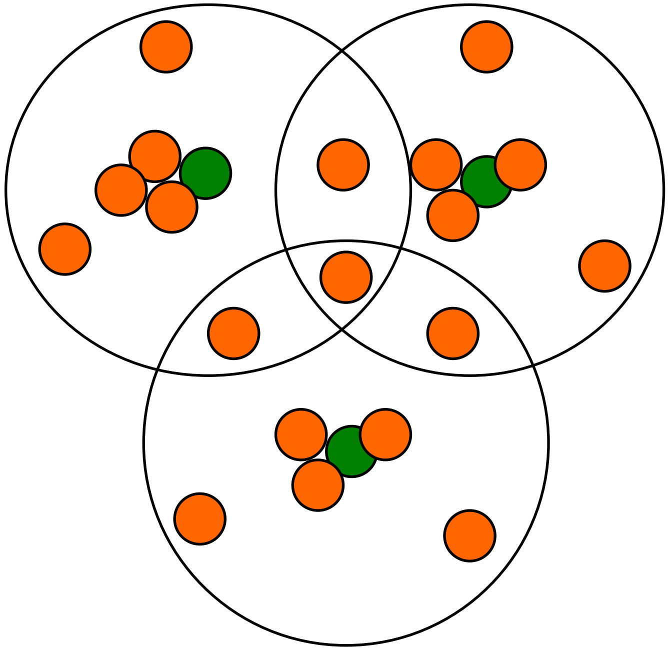
D':

	C	F	H
C	1	0	0
F	0	1	0
H	0	0	1

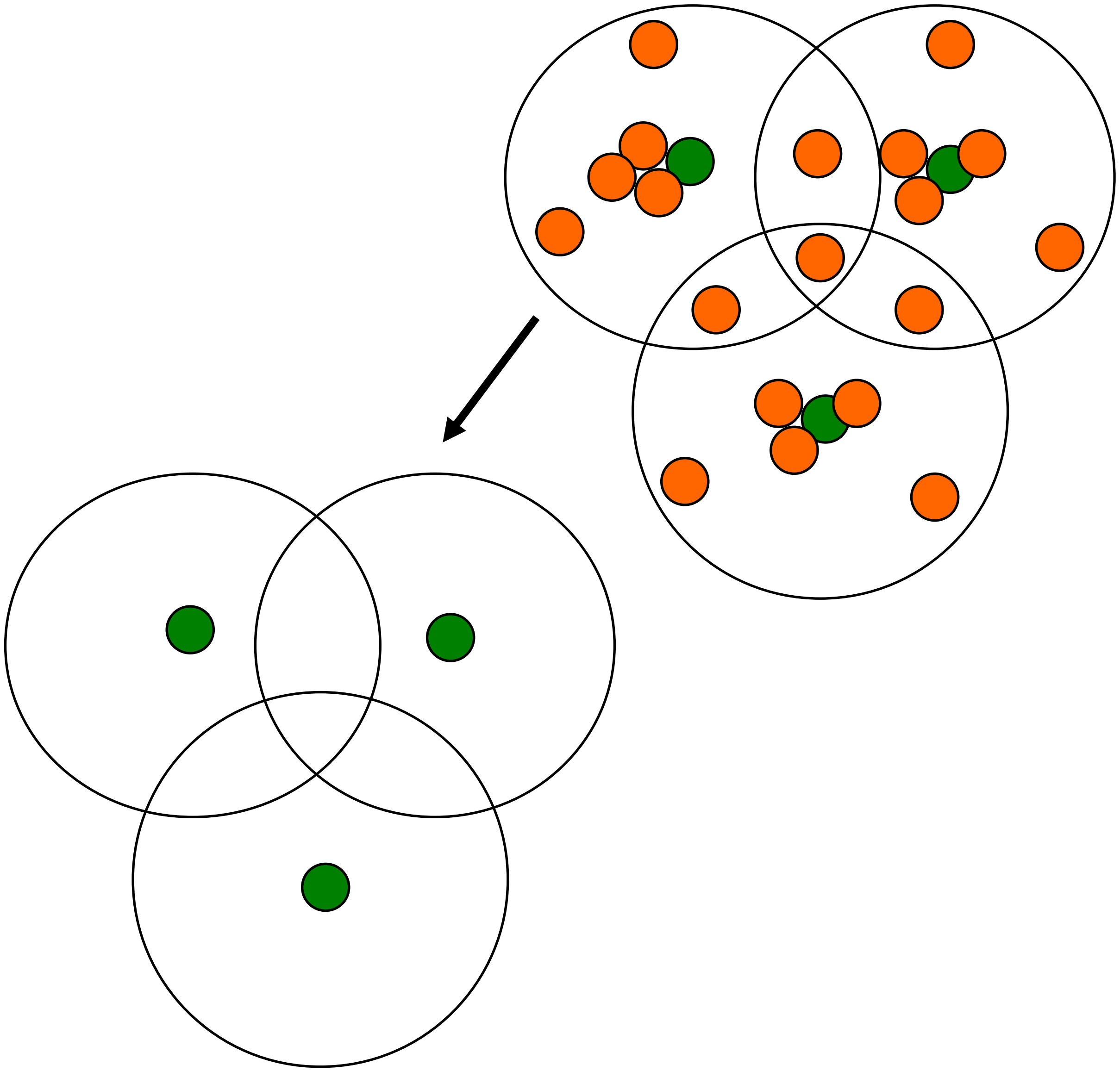
N
1
1
1

Unique list is C, F, H

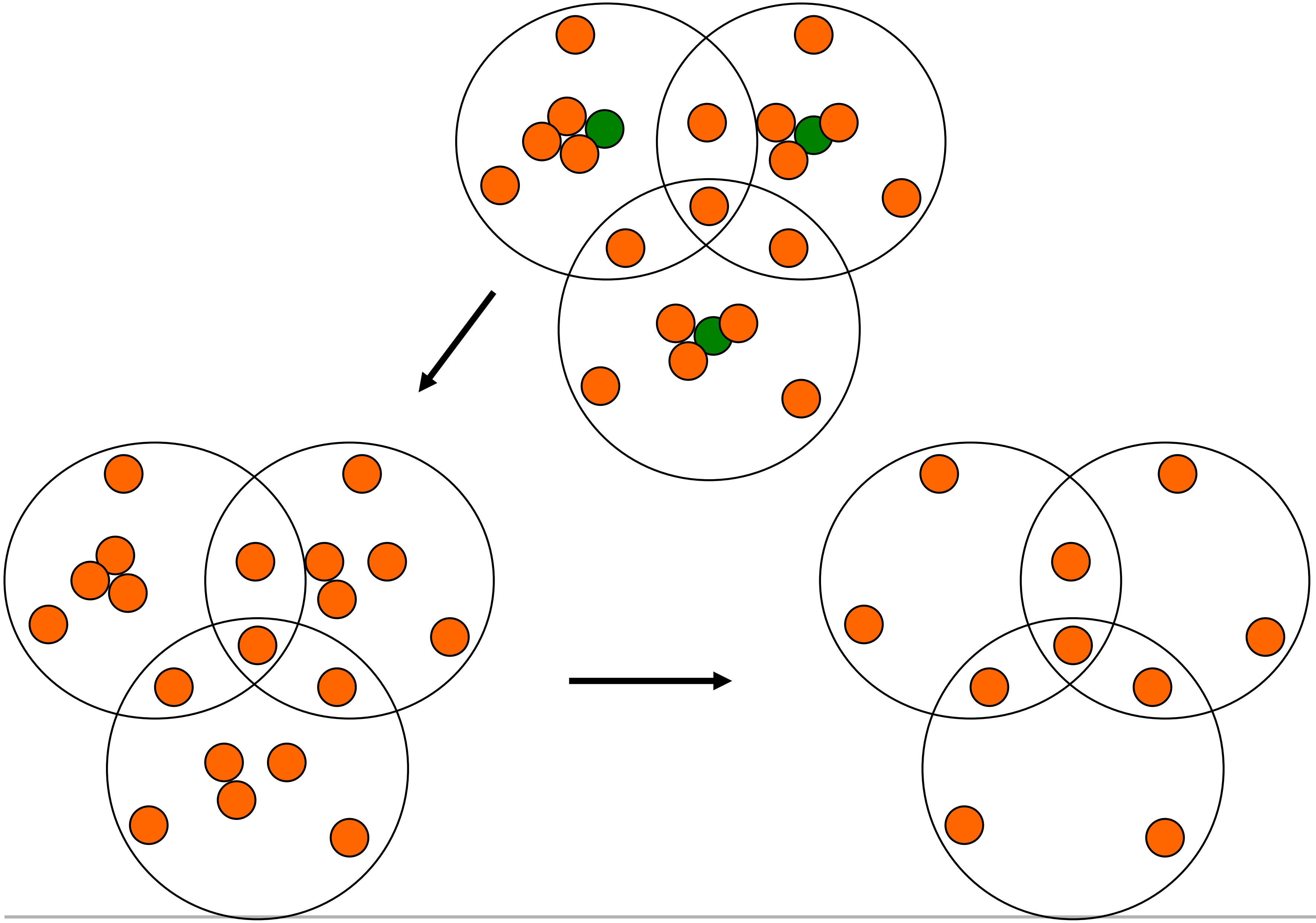
Hobohm



Hobohm-1



Hobohm-2



Why two algorithms?

- Hobohm-2
  - Unbiased
  - Slow (O2)
  - Focuses on lonely sequences
  - Example from exercise
    - 1000 Sequences alignment 2 hours
    - Hobohm-2: 22 seconds
- Hobohm-1
  - Biased. Prioritized list
  - Fast (0)
  - Focuses on populated sequence areas
  - Example from exercise
    - 1000 Sequences
    - Hobohm-1: 12 seconds
- Hobohm2 in general gives more sequences than Hobohm1

Hobohm-1 versus Hobohm-2

- Prioritized lists
  - PDB structures. Not all structures are equally good
    - Low resolution, NMR, old?
  - Peptide binding data
    - Strong binding more important than weak binding
- Quantitative data (yes no data)
  - All data are equally important