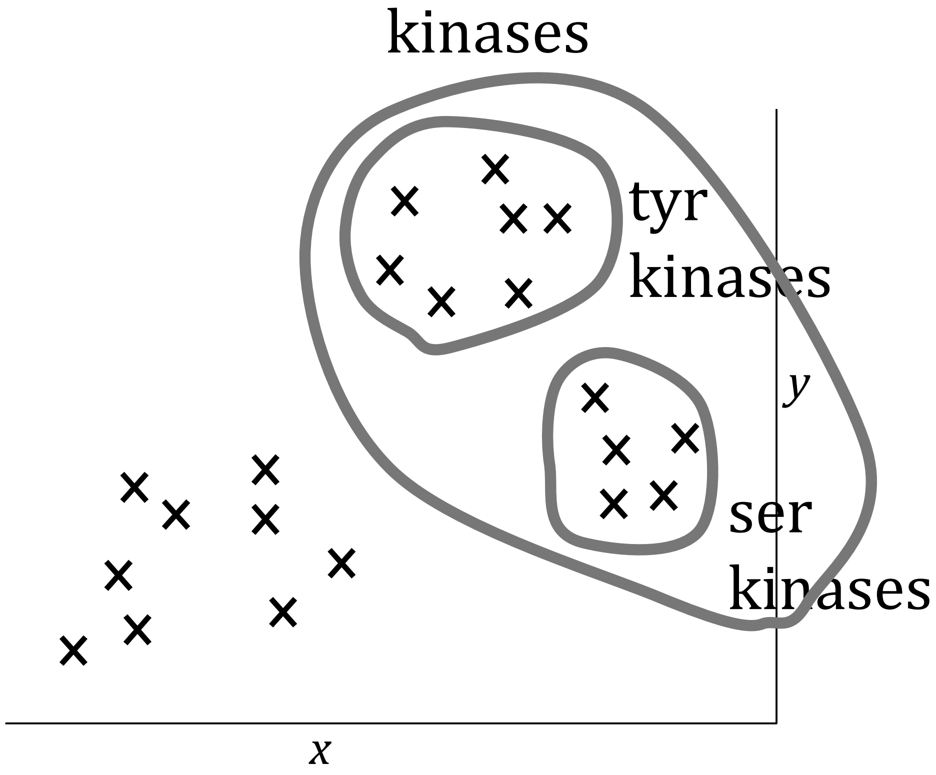
Protein spaces

Why ?

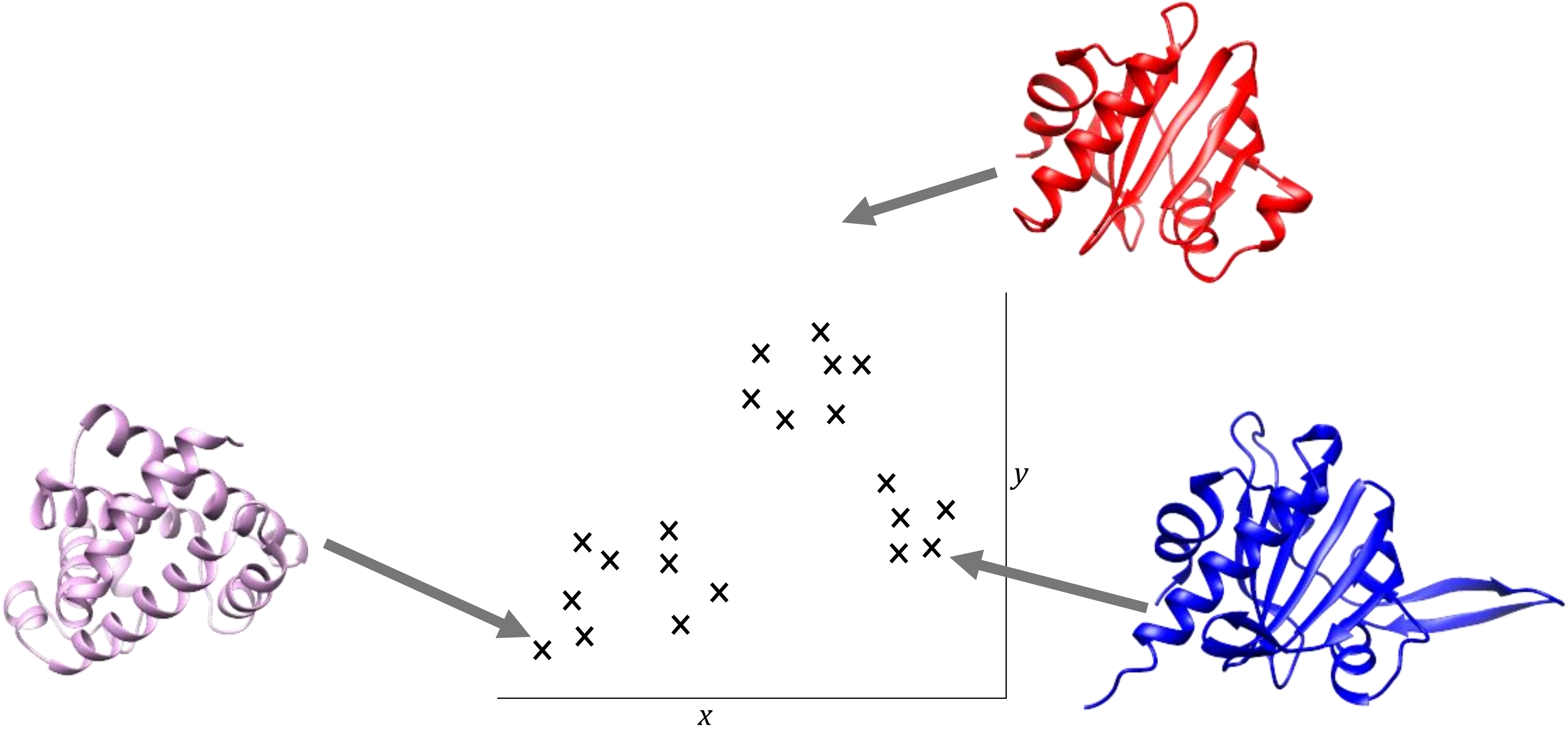
- We like to compare objects – prediction of properties
- groups of proteins – do they exist ?

Spaces – do they exist ?

- what is protein space ?
- who cares ?



A space of protein structures ?



who talks about spaces ?

Here

- sequence space (proteins)
- structure space (proteins)

Others – often not really spaces

- small molecule space / drug space
- tree space
- the set of solutions to a combinatorial problem
  - how many paths does the travelling salesman problem offer ?

What does a space mean to me ?

- usually a classic vector space / rarely a discrete space

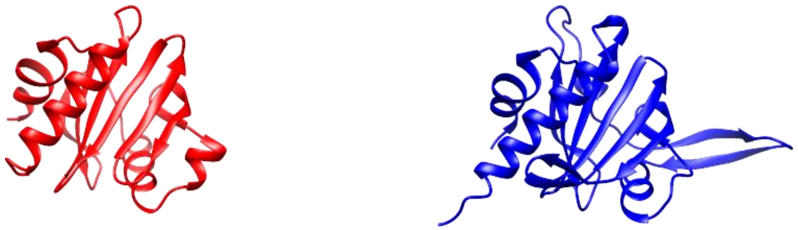
The questions

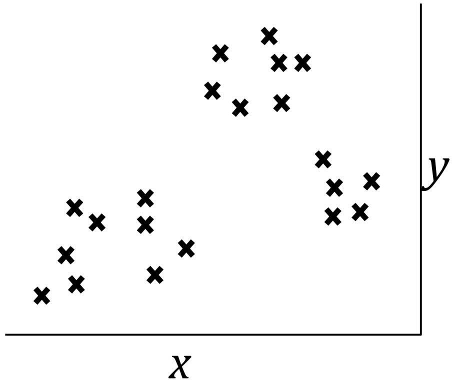
I want spaces that are

- objective
- reproducible
- tell me if A is similar to B

Proteins

- sequence space (discrete)
- structure space (continuous)
    - sequence and other spaces - continuous





Spaces

Conventional spaces

- 1D (x), 2D (x, y), 3D (x, y, z)
  - 4D (x, y, z, w), …
- let us estimate how big a space or problem is – examples

Example - sequence alignments – picking penalties

1. gap opening
2. gap widening

The optimal parameters are a point in a 2D space (one point)

Discrete spaces

Discrete space

- how many variables do I have ? (a, b, c, …)
- how many values can each variable have ?
  - a 3 values, b 4 values, c 5
  - number of points in space = $3 \times 4 \times 5$

Representing a Sequence

Protein sequence and structural coordinates

	1	2	3	4	5	6	7	…	N_{res}
x	1.2	2.3	…						10.3
y	2.4	3.5	…						11.1
z	1.7	2.9	…						15.5
seq	W	A	C	A	A	…			D

A protein is a set of 3D points

A protein is a set of 4D points / descriptors if we add sequence

- 4th dimension is not continuous
- This is NOT sequence space

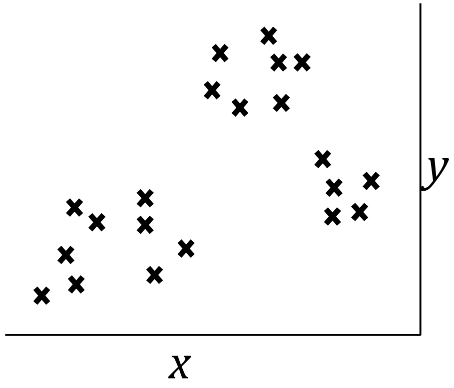
The sequence points

Usually, a protein is a set of points

I want one point = one protein

Consider proteins of length N_{res}

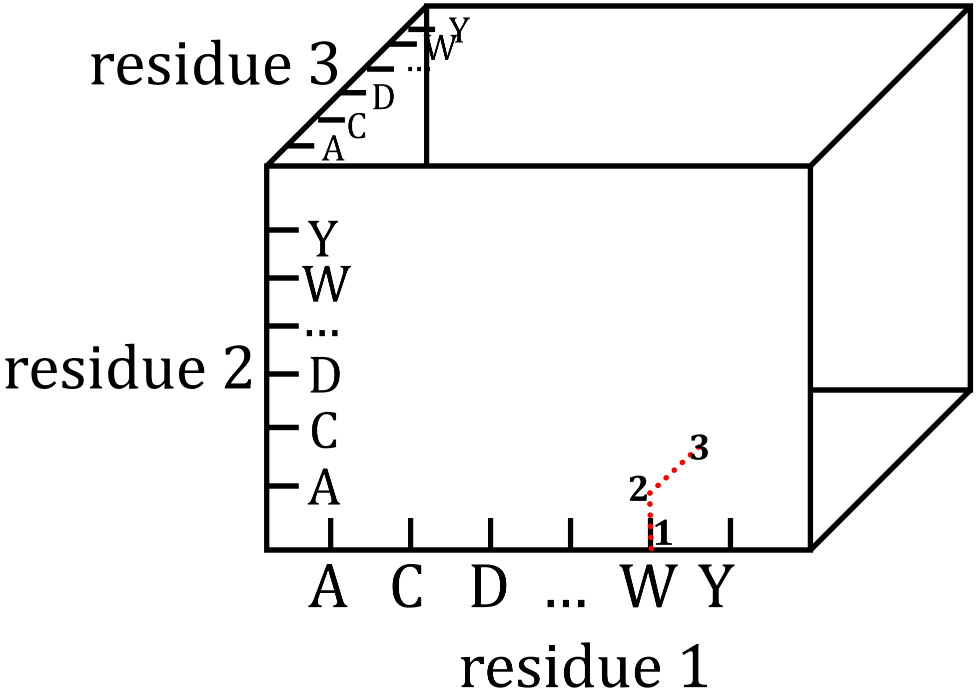
- look at the first few (3) points



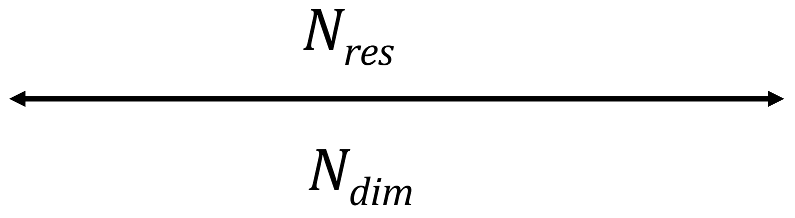
Finding a Sequence in This Space

Real diagram is a box of N_{res} dimensions

- this one 3 dimensions

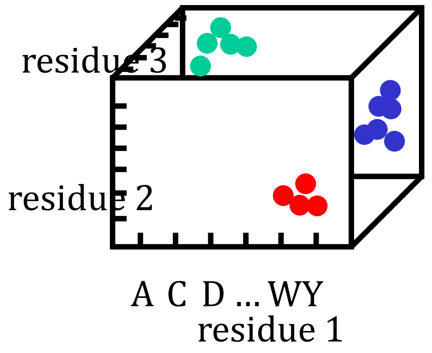


	1	2	3	4	5	6	7	...	Nres
x	1.2	2.3	...						10.3
y	2.4	3.5	...						11.1
z	1.7	2.9	...						15.5
seq	W	A	C	A	A	...			D



- looking for sequences…

Families in Sequence Space



Similar sequences are near each other

How realistic ?

- only works for $N_{seq1} = N_{seq2}$

Conceptual or practical

- important for discussions about protein families (conceptual)
- would you use it directly ? maybe with multiple sequence alignments

What is really ugly ?

- there is no natural ordering on axes

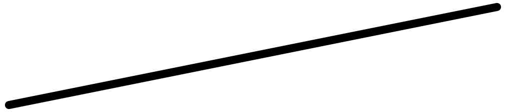
Summary

- we have a discrete space in which every protein is a point

General continuous spaces

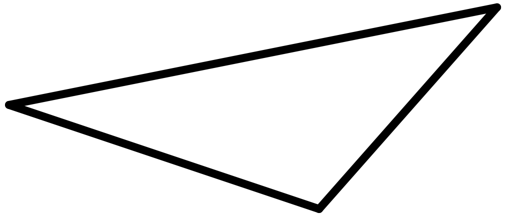
My sequence space

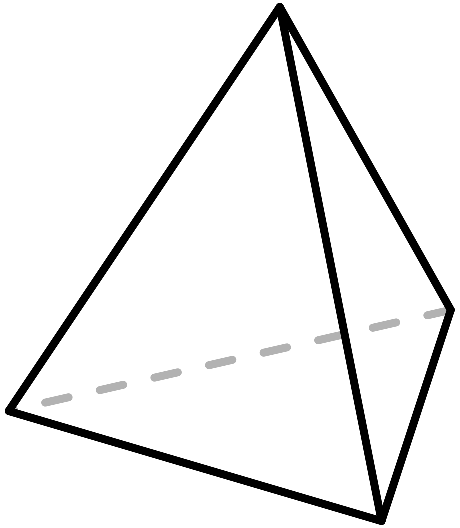
- conceptually useful   /   practically less so



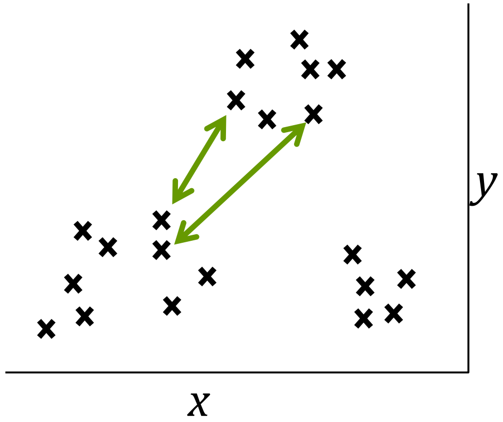
A generally useful approach

- 2 points fit in 1D (or less)
- 3 points fit in 2D (or less)
- ..
- N points can always fit into N-1 dimensions (maybe less)

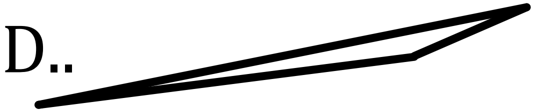




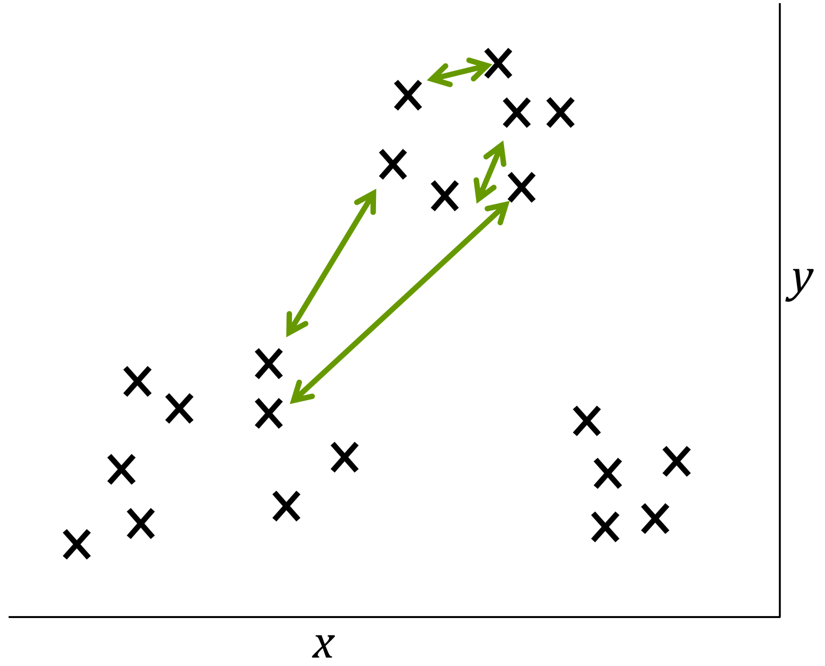
- my diagrams are usually 2D



- not all dimensions are equally important
  almost 1D..



Some protein spaces - sequences



Do I have a measure of similarity ? Many

Sequence-based

- % sequence similarity
- alignment scores
- *k*-mer similarity, ..

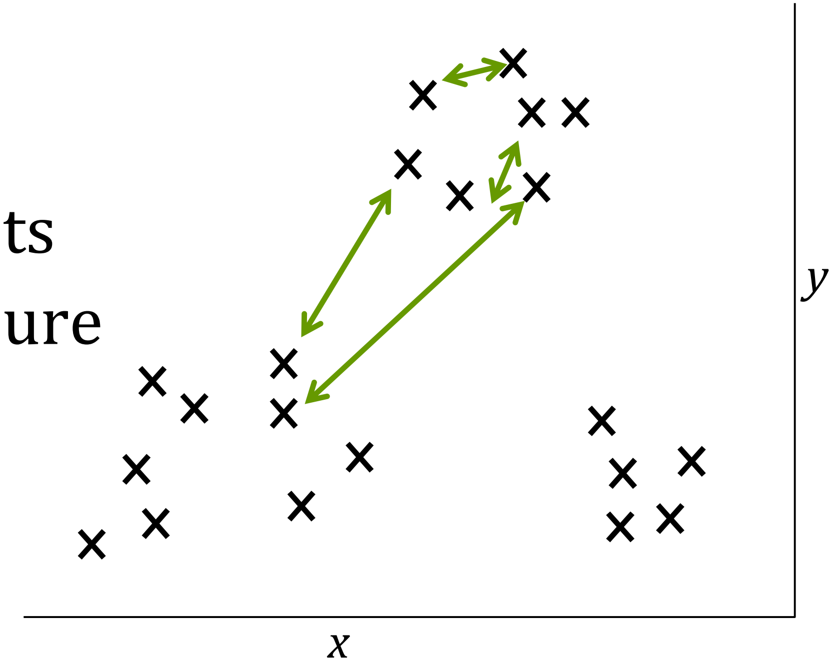
Whatever measure

- similar proteins will be close
- more distant relations will depend on the measure

Some protein spaces -structural

Structure-based

- superimpose and look at geometry
- count similarities in secondary structure elements
- break into fragments – use some similarity measure



General rule

- If I can define similarities there is an implied space

How big ?

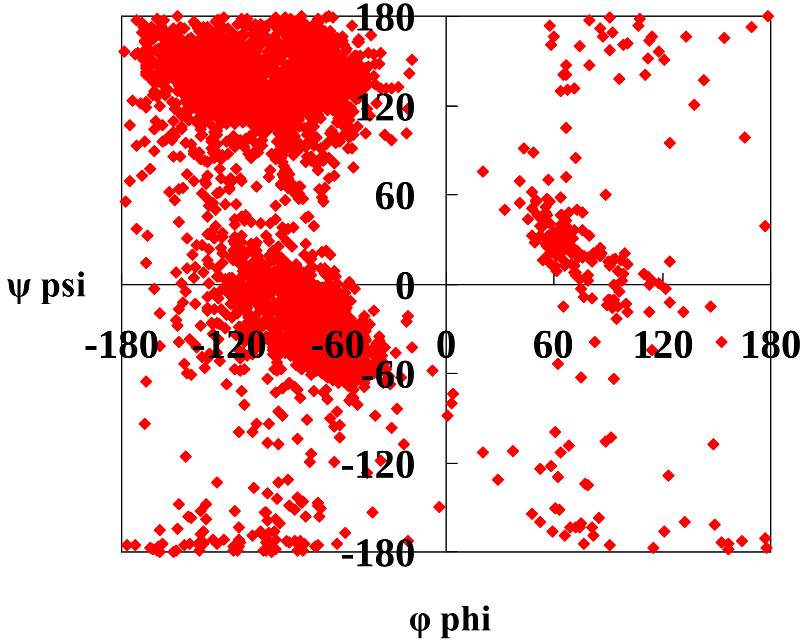
Sequence space ? (discrete)

- $20 \times 20 \times 20 \ldots = 20^N$

Conformational space – how to argue

- for each residue, there are at least 2 major groups (really more)
- maybe chop plot into 3 or 5 pieces
  - say there are c conformational possibilities
- c^N for some c
  - so 3^N or 5^N

- these spaces grow exponentially in the size of the protein



How general

- You can usually invent a space

- High dimensional spaces are not much fun (directly)
  - what do you do with 7-dimensional coordinates ?

What does one normally do ?

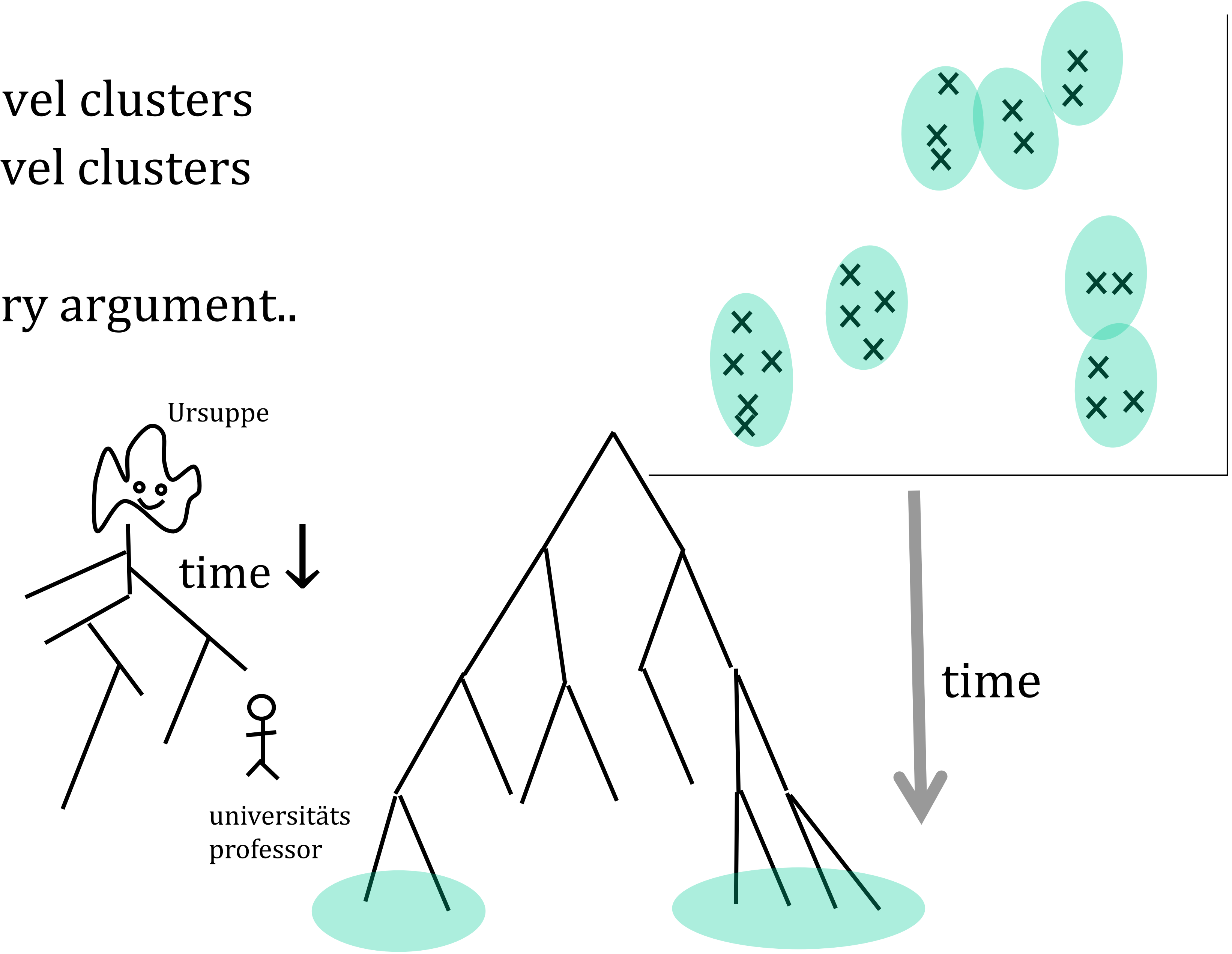
- reduce to fewer dimensions – find the best 2 or 3-dimensional representation of the data
  - distance geometry / principal components OR

- work with distances - coordinates are just something to think about

Should we expect a hierarchy ?

- 7 lowest level clusters
- 3 higher level clusters

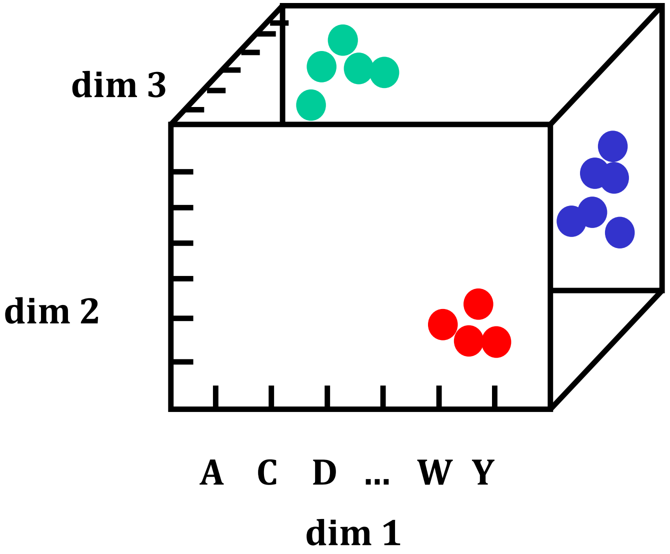
- evolutionary argument..

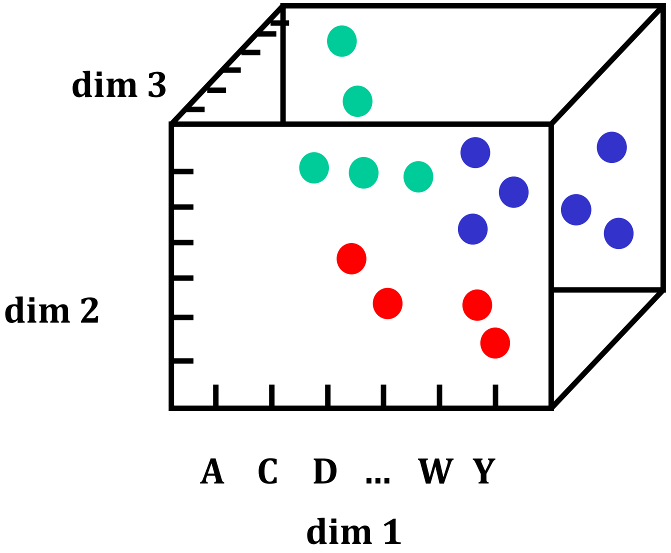


Do we expect protein families ?

No real answer

- we have an idea of spaces – sequence or structure based
- how are proteins distributed ?

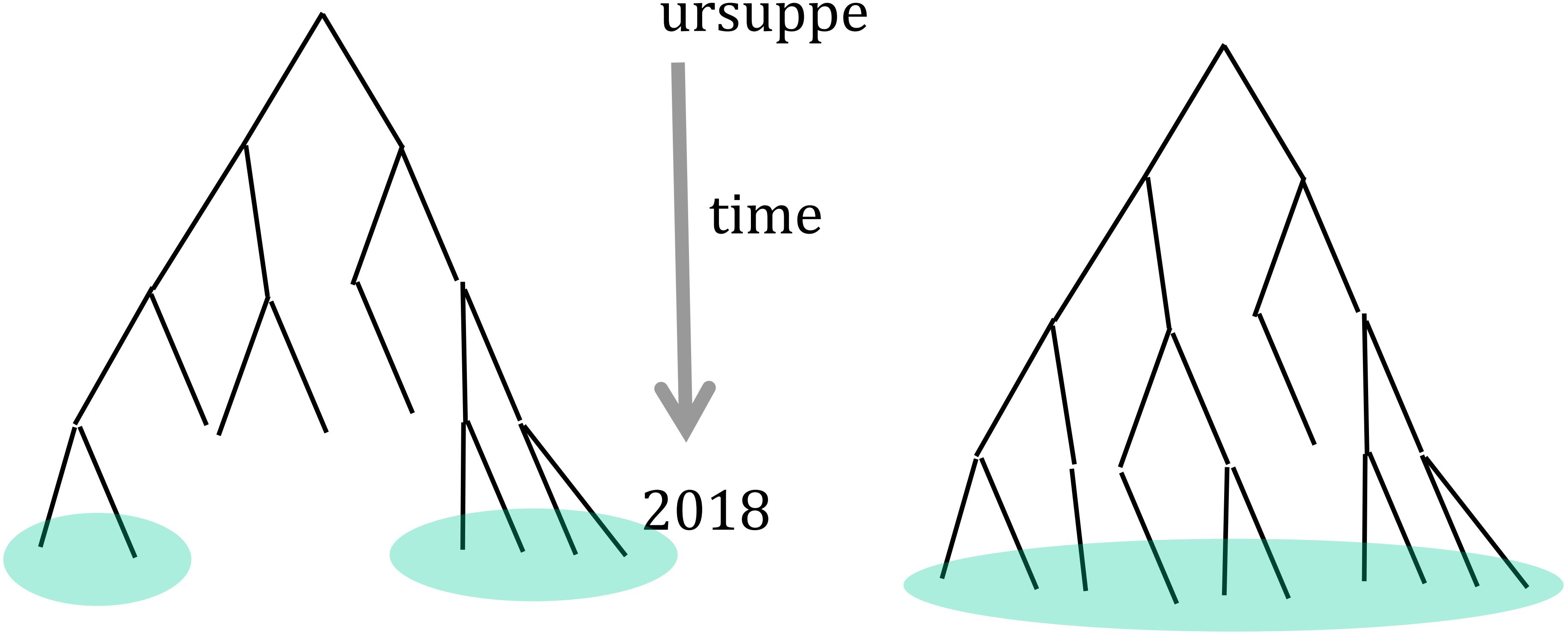




`Should you expect clusters ?

Evolution and phylogeny

Shape / density of tree of life



ursuppe

time

2018

clear
families

no families

Do not forget

- We can always define spaces
  - implicit in the word homology (proteins near in some space)
- Sequence and structure spaces are very different
  - lots of sequence families
  - fewer structural groups