How many protein folds are there ?

- in the protein data bank ?
- on earth ?
- possibly ?

What is a protein fold ? definition for today

- a common shape for proteins
- do not look at sequence similarity (changes much faster than structure)
- same order and size of secondary structure elements
- they evolved from a common parent protein
- allow for insertions, deletions and some large changes

Typical numbers

10^5 structures in protein data bank (PDB)

- much redundancy

1 ½ × 10^5 chains in PDB

- even more redundancy

Human-checked collections of structures

- 1962 "superfamilies" in SCOP (2009 out of date)
- 2626 "superfamilies" in CATH

Our classification:

- 2×10^4 different structures

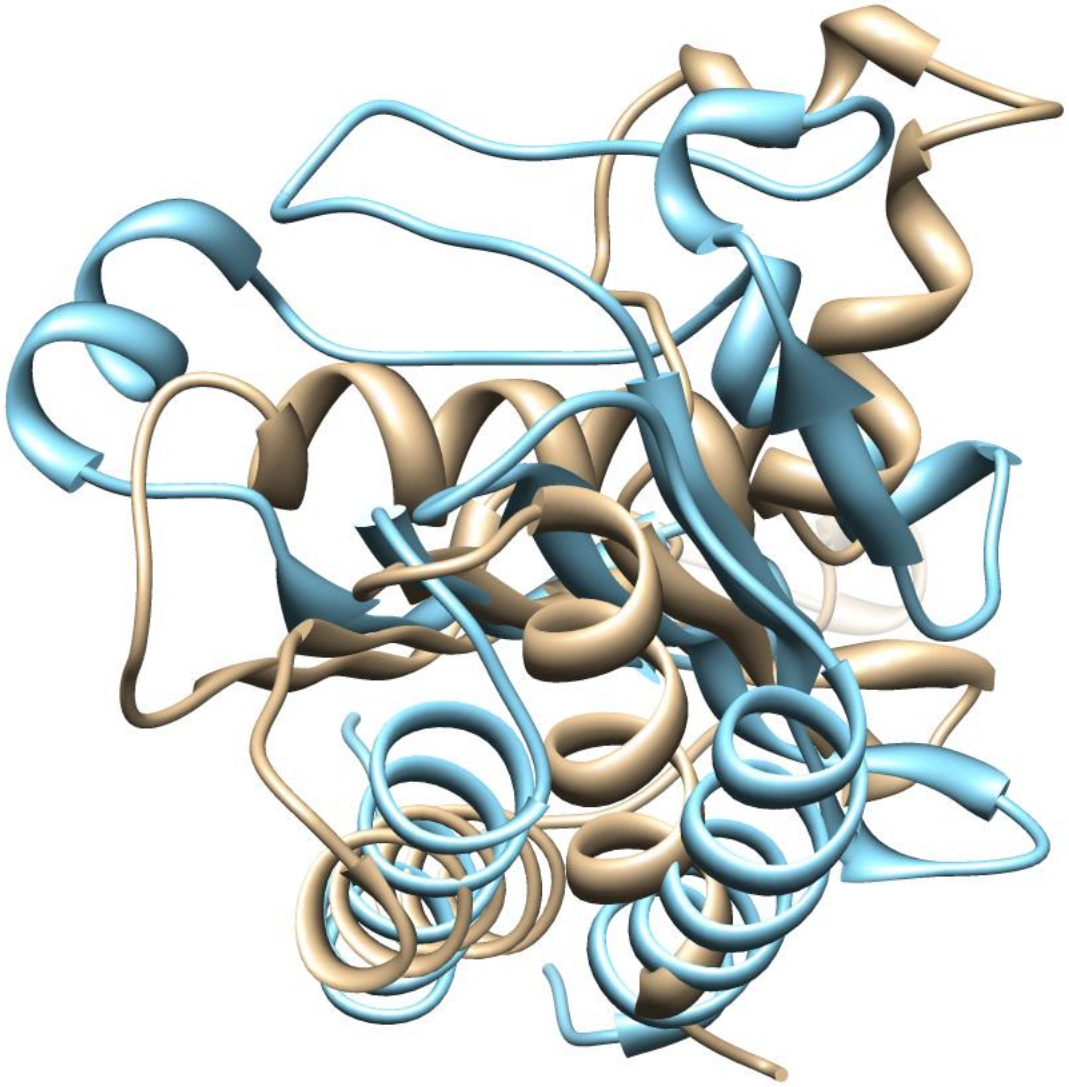
Sequences ?

- 3.5×10^7 sequences in "nr" sequence databank

What is a fold ?

Forget sequence identity

- are these the same fold ?

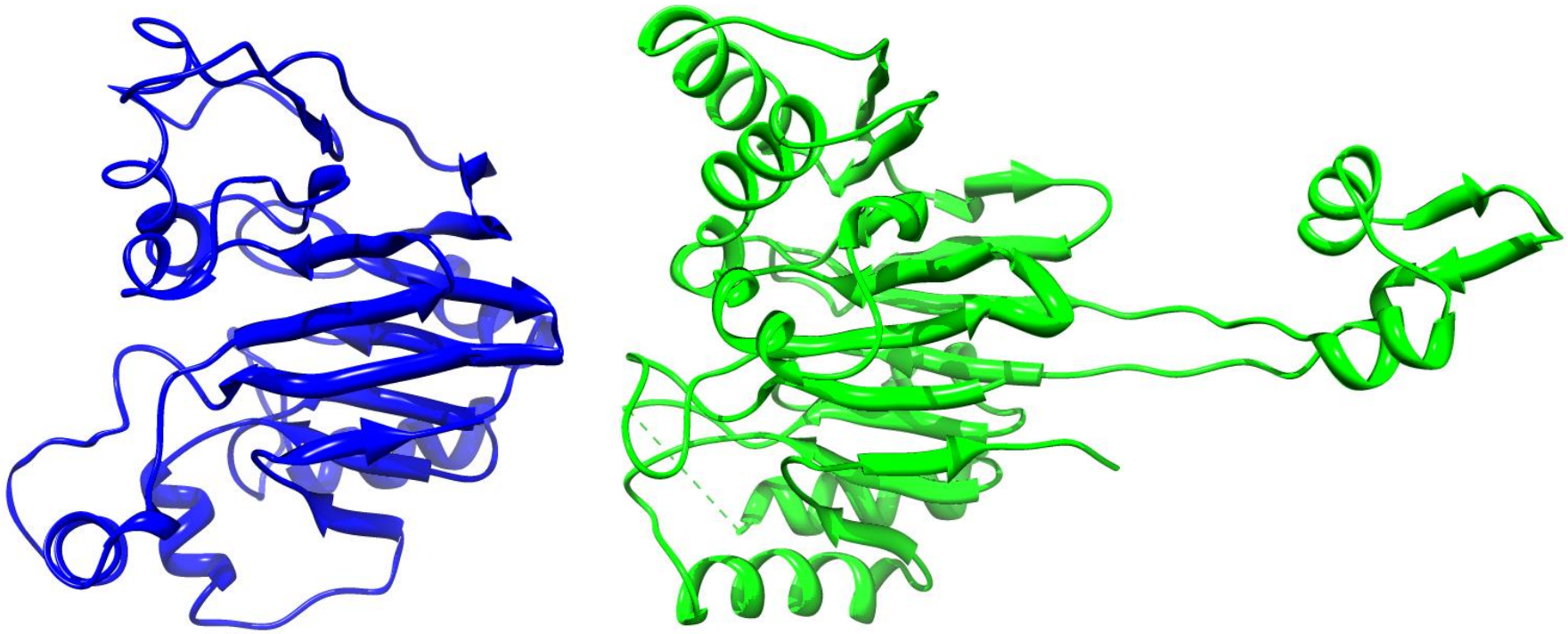


3fpv 2w3e, cannot be aligned by sequence methods

What is a fold ?

Forget sequence identity

- are these the same fold ?

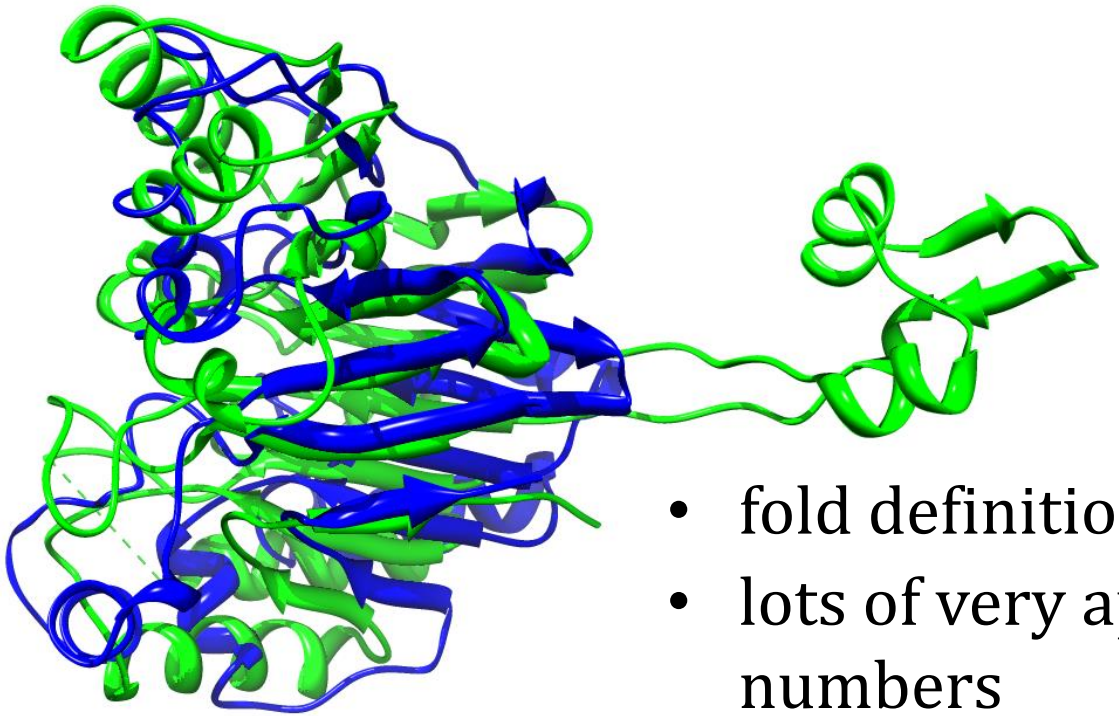


3g1p 1y44, cannot be aligned by sequence methods

What is a family ?

Forget sequence identity

- are these the same family ?



- fold definition – very arbitrary
- lots of very approximate numbers

3g1p 1y44, cannot be aligned by sequence methods

Operational fold definitions

1. use definitions from literature (SCOP / CATH / ..)
   - often very hand-made, non-reproducible, out of date
2. second half – geometric definitions

How often does one see a new fold ?

Claim in 1990's

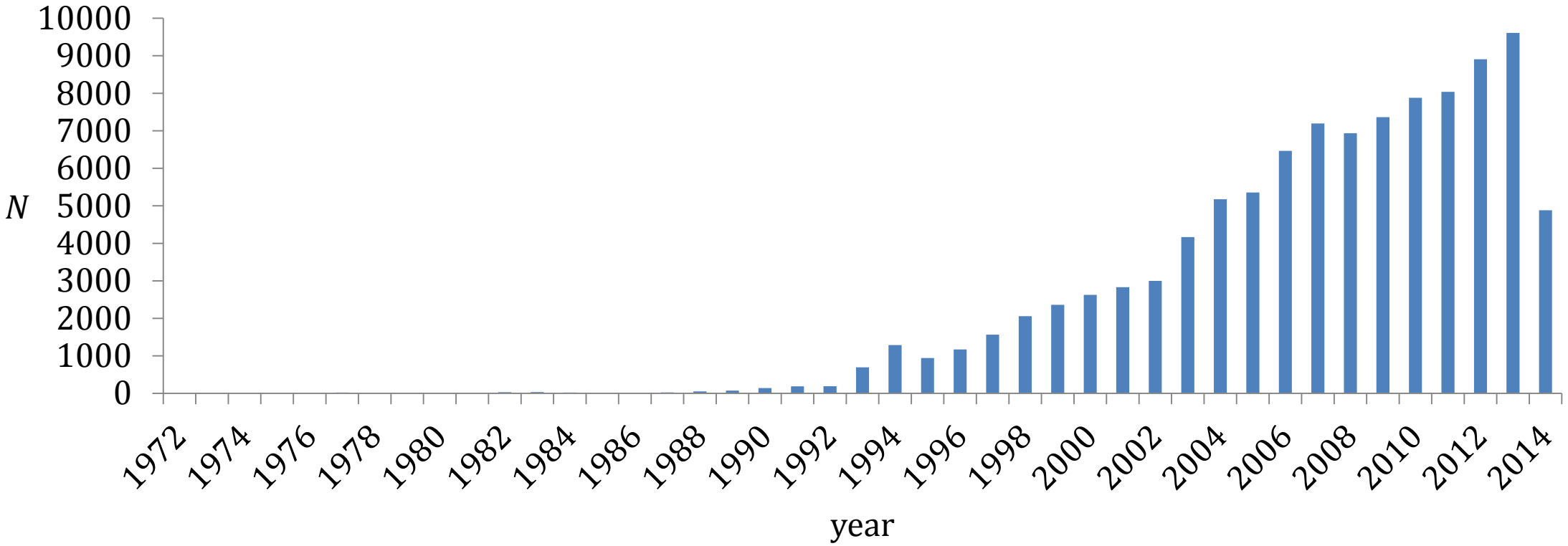
- mostly when a new structure is solved (80-90%)
  - looks like a structure which was already in databank

Important:

- even when you would not expect it from sequence similarity
- different sequences can still have the same fold

Not quite quantified ..

new structures per year



How many new folds ?

- max a few hundred each year (no really authoritative numbers)

Why is this interesting ?

Structure prediction

- do not have to predict structures *de novo*

  just find the fold for your structure and use alignment methods

Crystallography

- molecular replacement works for about ¾ of structures today
- requires a relatively closely solved structure

Why is this interesting ?

Structural genomics

- systematically solving structures
- How many are necessary for structure prediction and crystallography ?
- try to solve representative of every fold

Practical ?

- 10^3 or 10^4 folds might exist – not too many

For fun

- of the *n* possible protein structures, how many has nature tried ?

Problem

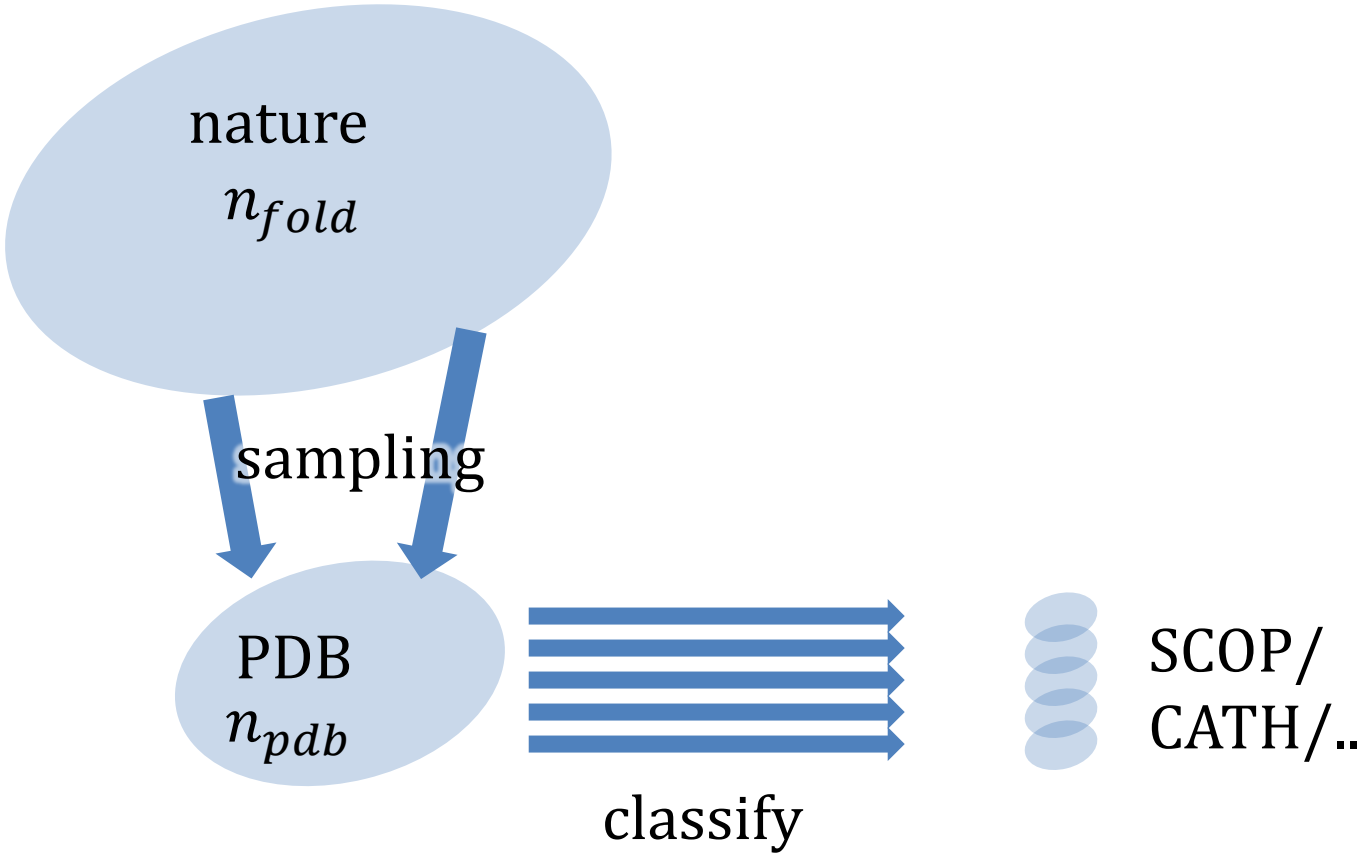
- How many folds are there ? n_{fold}
- How many proteins in PDB ? n_{pdb}

How would you approach the problem ? Examples

1. statistical – look at distribution of structures
   The PDB is a small sampling from n_{fold}

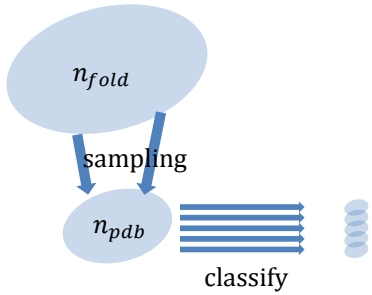
2. geometric – how many could there be ?

Statistical approach



statistical approach – very naïve

- say 10^4 classes in nature $n_{fold} = 10000$
- $n_{pdb} = 10^5$
  - would we seen every fold 10 times ?
    - some folds not seen, some seen 20 times



Look at set of numbers

- $n_{obs}(1), n_{obs}(2), \ldots$
- if $n_{fold} = \frac{1}{10}\, n_{pdb}$

$\langle n_{obs}(i) \rangle = 10$ (not so helpful)

variance will be big

$\langle x \rangle$ mean of x

Statistical approach

n_{fold} folds in nature

n_{pdb} number of samples (structures in PDB)

$n_{obs}(i)$ number of proteins seen in PDB with fold *i*

Classic problem

- bag with many coloured balls
- sampling of balls from bag

- consider simpler question – binomial distribution

binomial version

classic problem

- from 100 coin toss, probability of 10 heads or 50 heads..

p probability of outcome on some trial (½ for coins)

n trials, k success

$$p(k) = \binom{n_{trial}}{k} (p)^k (1-p)^{n-k}$$

what is the probability of seeing fold i $n_{obs}(i)$ times ?

$$p(n_{obs}(i)) = \binom{n_{pdb}}{n_{obs}(i)} (p_i)^{n_{obs}(i)} (1-p_i)^{n_{pdb}-n_{obs}(i)}$$

Where is the number of folds ?

binomial distribution

$$p(k) = \binom{n_{trial}}{k} (p)^k (1-p)^{n-k}$$

Say all folds are equally likely

(likelihood of a globin is the same as a β-sandwich)

$p = {}^{1}/_{n_{fold}}$

$$p(k) = \binom{n_{pdb}}{n_{obs}(i)} \left(\frac{1}{n_{fold}}\right)^k \left(1 - \frac{1}{n_{fold}}\right)^{n_{pdb}-k}$$

- first thoughts

Using the idea

- go to PDB get n_{pdb}
- go to your favourite classification
- see how many times each fold i occurs

- gives an answer

Results of naïve approach

450 classes in one estimate

Not good yet…

- n_{pdb} is really much much $< 10^5$ redundancy

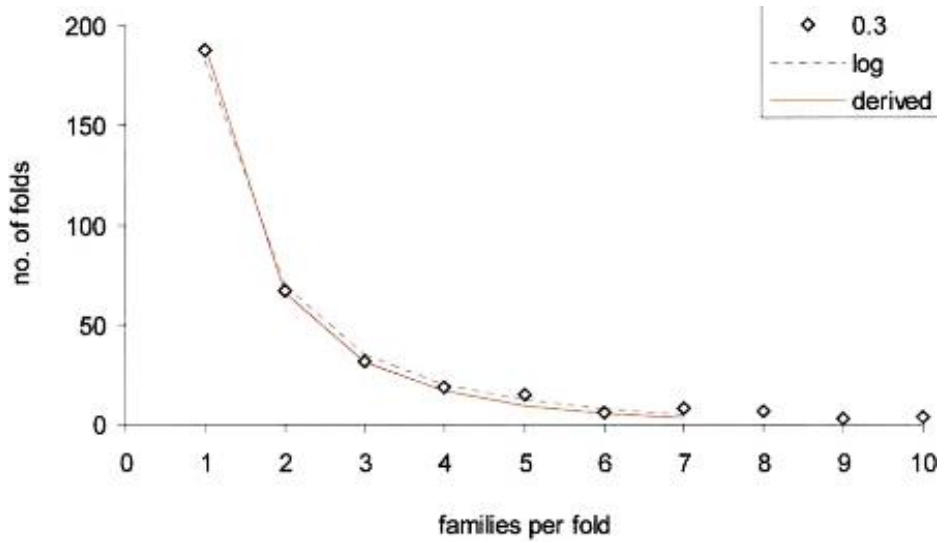
- some folds are common
- some are rare
  - Remember lectures on popular structures / unlikely structures

Wolf, Y.I., Grishin, N.V., Koonin, E.V. (2000) J. Mol. Biol. 299, 897-905

statistical approach - better

Use some functional form for distribution over protein folds

- stretched exponential $P(\lambda_i) = c \exp\left(-\alpha \lambda_i^{\beta}\right)$
  - λ_i relative probability of fold i
  - α, β constants to be fit



statistical approach - better

Do not assume probabilities are equal

- get the probability of each fold from classification (lots of different p_i)
- fit to stretched exponential – get a set of p_i

- do not use a binomial distribution (two outcomes)
  - use a multinomial (many outcomes)

- work back to get estimate of n_{fold}

statistical - summary

Estimates vary from 1000 to 4000 (and more)

- few estimates of 8000

Problems

- what is distribution of proteins over folds ?
  - leads to question .. why ? Relates back to "popular structures"

Is the PDB a fair sampling ? Lots of

- human proteins
- structural genomics proteins (easy to clone and crystallise)
- soluble proteins
- proteins related to diseases (in host or agent)
- proteins are easier if they are similar to a known one

geometric approach

Problems so far

- everything depends on definitions of similarity / what is a fold

- is there a system which is
  - simple
  - quantitative ?

Could you imagine a system

- which lets you generate random conformations
- must be protein like
- count the number which are possible for a threshold of simlarity

geometric approach

How many ways can a chain fold ?

- rules
  - compact
  - atoms do not hit each other
- less obvious
  - chain direction usually reverses
    - α-helix after 2 residues
    - β-strand after about 10 residues (typical)

Mission

- sample from possible chains fulfilling these conditions
  - can you sample from x, y, z ? Not easily
- work in a different space

cosine transform - diversion

Fourier transform – well known

- go from real (complex) space to frequency space
- or from frequency space to real (complex)

"cosine transform" similar

- work with real (not imaginary) parts

coordinate filtering example

filtering / transform example

real coordinates (x, y, z)

transform

frequency signals (h, k, l)

discard high frequency components (keep N)

frequency signals (h, k, l)

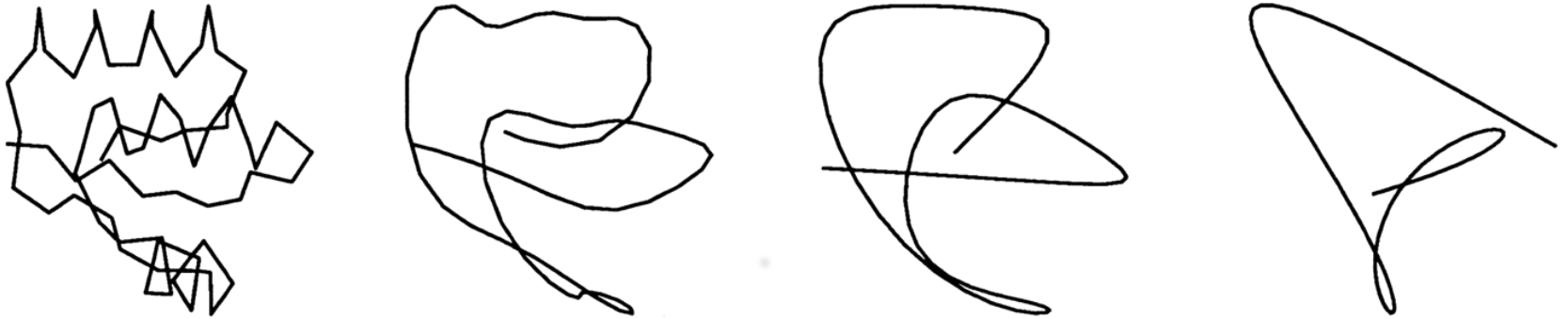
transform

smoothed real coordinates (x, y, z)

Example transform

example protein (1ctf)

- transform → frequencies
- keep only N =22, 11 and 6 points (frequency space)
- transform back to real space



original | 22 points | 11 points | 6 points

Crippen, G.M., & Maiorov, V. (1995) J. Mol. Biol. 252, 144-151

Why is N sampling important ?

real → frequency → real

- we can lose as much detail as we want

in this application

- we start in frequency domain and generate as much detail as we want

N (number of frequencies) controls flexibility

- $N = 2$ allows only straight lines, $N = 3$ allows one bend, ..

Some formulae

directions of the transform

from coordinates (real proteins) to frequencies

$$\hat{x}_k = \frac{2c_k}{N}\sum_{j=0}^{N-1} x_j \cos\left[\frac{(2j+1)k\pi}{2N}\right] \qquad k = 0,..N-1$$

from frequencies to coordinates

$$x_n = \sum_{k=0}^{N-1} c_k \hat{x}_k \cos\left[\frac{(2j+1)k\pi}{2N}\right] \qquad j = 0,\dots,N-1$$

do not remember for Klausur

so N controls the detail you want

details of formulae in two Folien

Sampling conformations

How can you sample wobbly lines (3 dimensions) ?

- not easy in real space

Method

- sample in frequency space
- convert to real space (one dimension x)

$$x_n = \sum_{k=0}^{N-1} c_k \hat{x}_k \cos\left[\frac{(2j+1)k\pi}{2N}\right]$$

in more detail

$$x_j = \sum_{k=0}^{N-1} c_k \hat{x}_k \cos\left[\frac{(2j+1)k\pi}{2N}\right]$$

- x_n the nth coordinate (what we want in real space)
- c_k usually 1 (not interesting)
- $\hat{x}_k$ coefficient for the k th frequency
- N how many samples (amount of detail / resolution)

Sampling from real coordinates

$$x_j = \sum_{k=0}^{N-1} c_k \hat{x}_k \cos\left[\frac{(2j+1)k\pi}{2N}\right]$$

decide on N (level of detail) and n_r number residues

while (step < max_step)

pick random $\hat{x}_k, \hat{y}_k, \hat{z}_k$ (uniform [-1,1])

(for lower frequencies, others set to zero)

convert to real coordinates, scale for n_r

check for overlap, repair / discard

check for similarity to stored structure, repair/discard

save coordinates

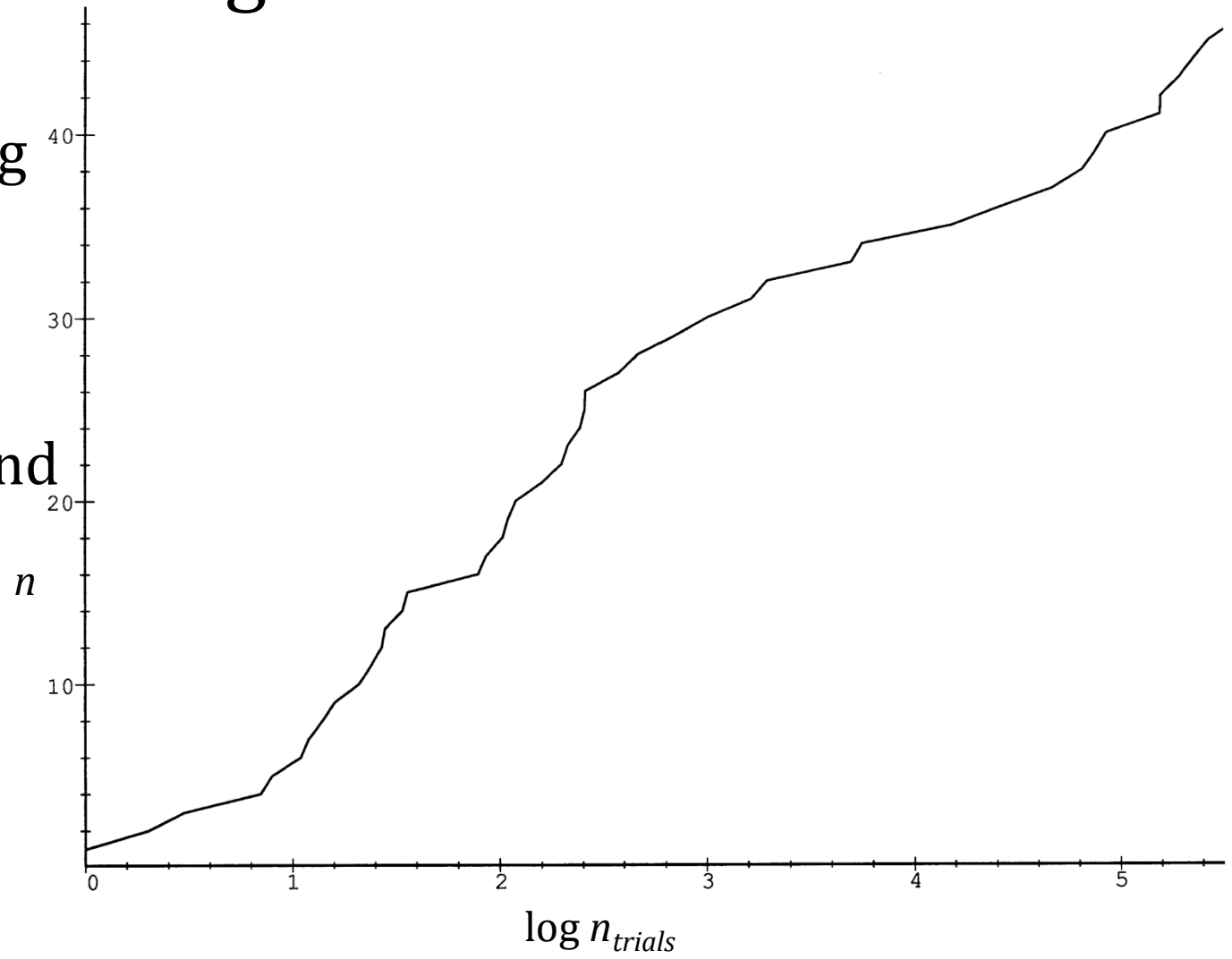
try random perturbations of each structure

add to set if sufficiently different structure is found

Finding new structures

time course of searching for structures

how many representatives are found with n_{trials} ?



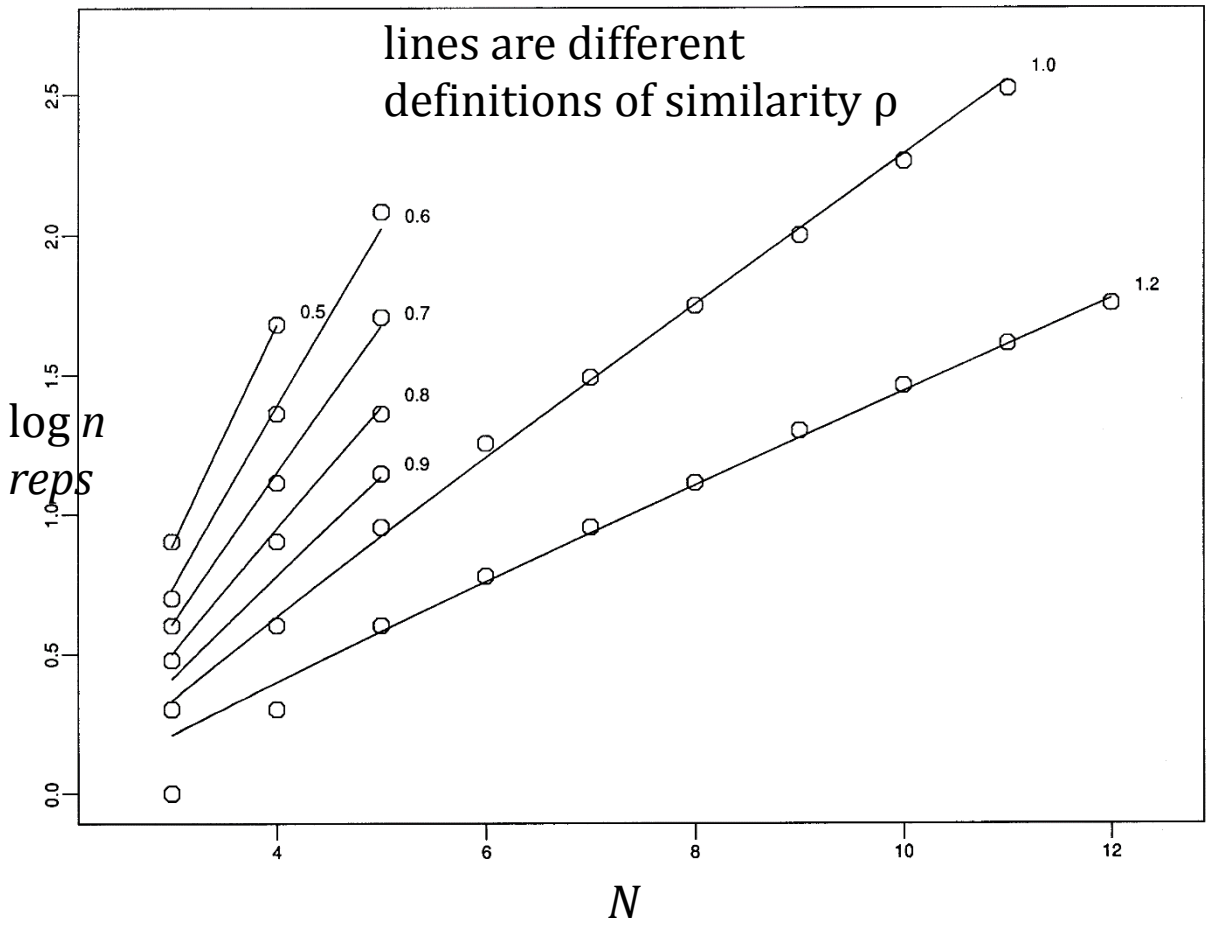
Crippen, G.M., & Maiorov, V. (1995) J. Mol. Biol. 252, 144-151

Estimating number of folds

Parameters

- definition of similarity ρ
- number of points in transform *N*

- Estimates are lower bounds



Crippen, G.M., & Maiorov, V. (1995) J. Mol. Biol. 252, 144-151

How many folds ?

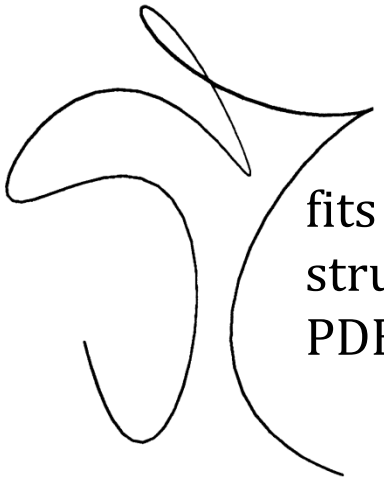
As many as you want

- 10^3 smaller structures (50 residues)
- very big numbers for larger structures

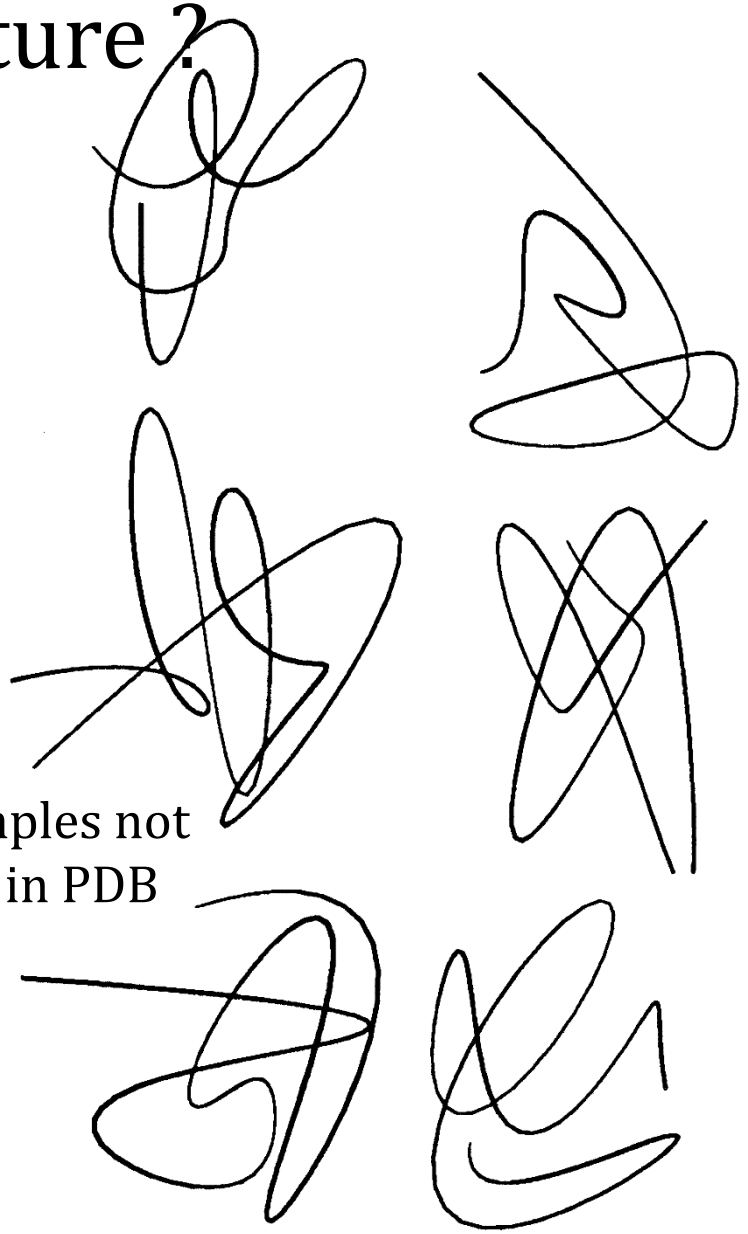
- many structures generated are similar to natural ones
- many may not be possible
  - representation a bit crude, does not capture enough detail
- may have found some structures that have not yet been discovered

agree with nature ?

- some look like real proteins



fits to > 100 structures in PDB



examples not seen in PDB

Crippen, G.M., & Maiorov, V. (1995) J. Mol. Biol. 252, 144-151

agree with nature ?

Would you expect to find the artificial structures in PDB ?

- many more structures since 1995
- PDB is a sample of structures from nature

Would you expect to find the structures in nature ?

- evolution:
  - mutate
    - sequence changes – maybe protein functions
    - sequence + structure change
      - almost certainly does not work (you die)
- very hard to visit all possible structures

Change original question

Now three questions

1. how many folds in PDB ?
   - we have the structures – mainly a question of definition
2. how many folds in nature ?
   - biology / chemistry /evolution question
3. how many folds could there be ?

summarise 1

- How many folds – why does it matter ?
  - modelling / structure / function prediction
  - finding evolutionary history
- Folds are not well defined
- Similar folds are not easy to recognise

- Statistical methods – many variations – one here
  - all use an arbitrary definition of fold
  - survey observed folds + distribution of proteins over these folds
  - more information not discussed here
    - many sequences in databanks
    - how are they distributed over folds ?

summarise 2

geometric approach

- pure sampling (not conclusive)
- avoids problem with sampling in real space
- has suggested new folds – chemically plausible

Is it likely that nature has visited all reasonable conformations ?

- difficulty in making a new stable protein shape
- sequence mutations explore sequences compatible with functioning protein
- structural changes usually deadly