# 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

- 8 × $10^4$ structures in protein data bank (PDB)
    - outrageous redundancy
- 1 ½ × $10^5$ chains in PDB
    - even more outrageous redundancy

human-checked collections of structures

- 1 962 "superfamilies" in SCOP (2009 out of date)
- 2 549 "superfamilies" in CATH

Bayerisch automatic:

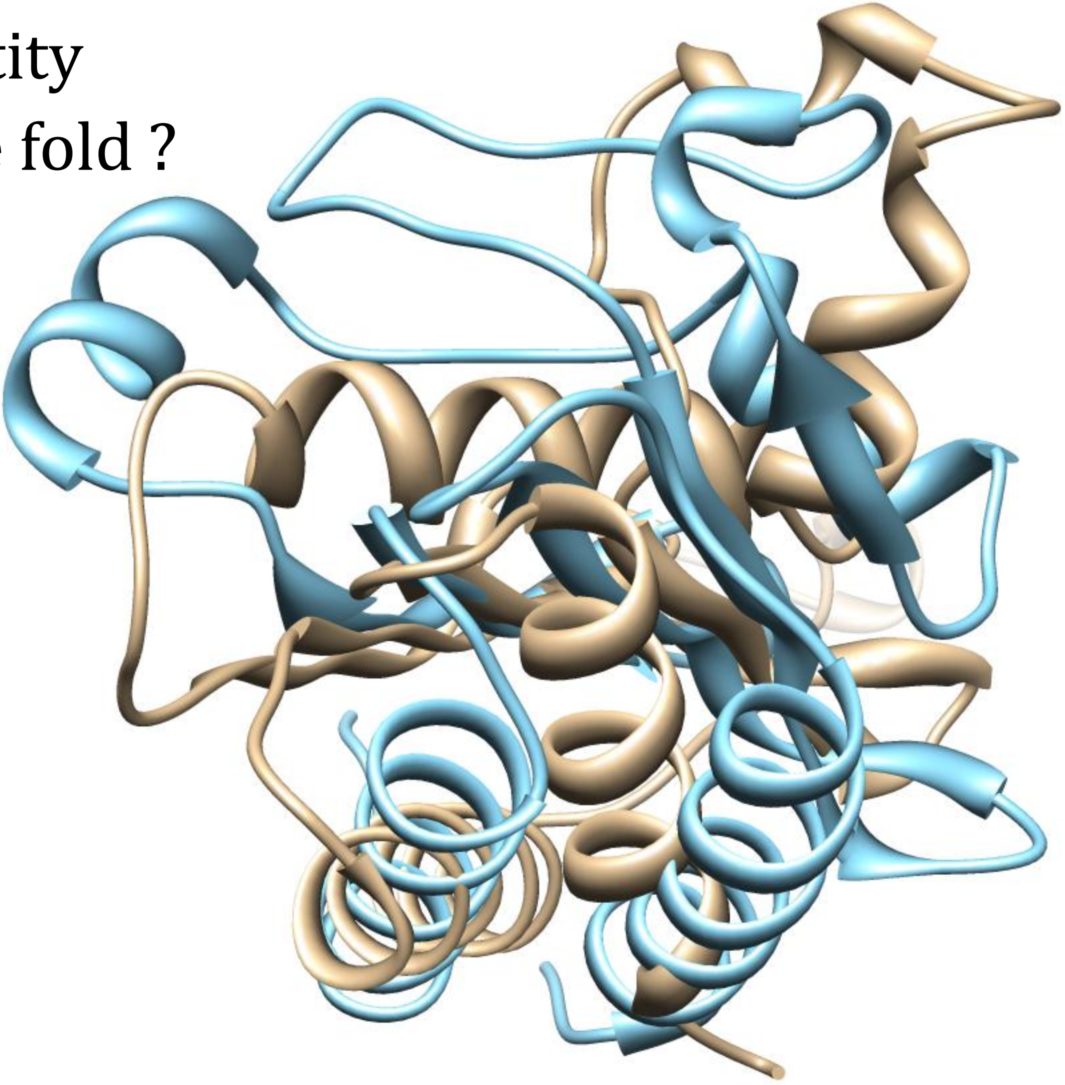
- 2 × $10^4$ different structures

Sequences ?

- 2 × $10^7$ sequences in "nr" sequence databank

# What is a fold ?

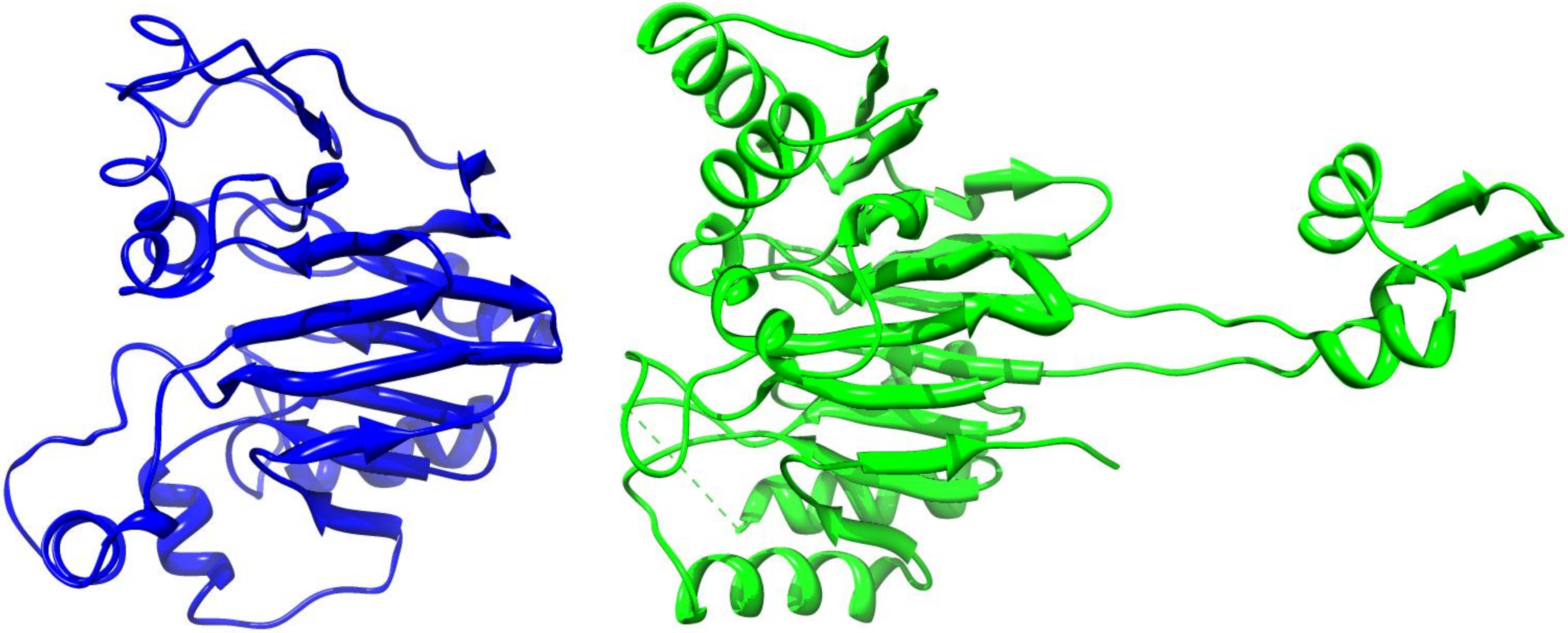
- forget sequence identity
  - are these the same fold ?



3fpv    2w3e, cannot be aligned by sequence methods

# What is a fold ?

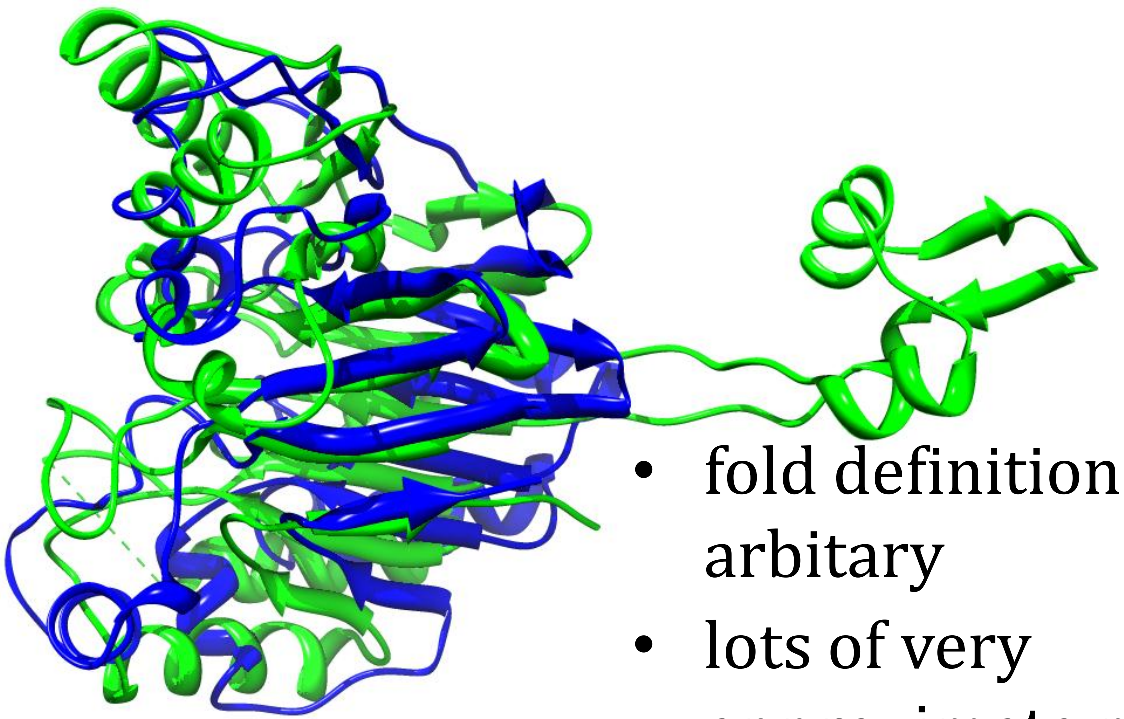
- forget sequence identity
  - are these the same fold ?



3g1p 1y44, cannot be aligned by sequence methods

# What is a family ?

- forget sequence identity
  - are these the same family ?



- fold definition – very arbitary
- 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    - often very hand-made, non-reproducible, out of date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 ?

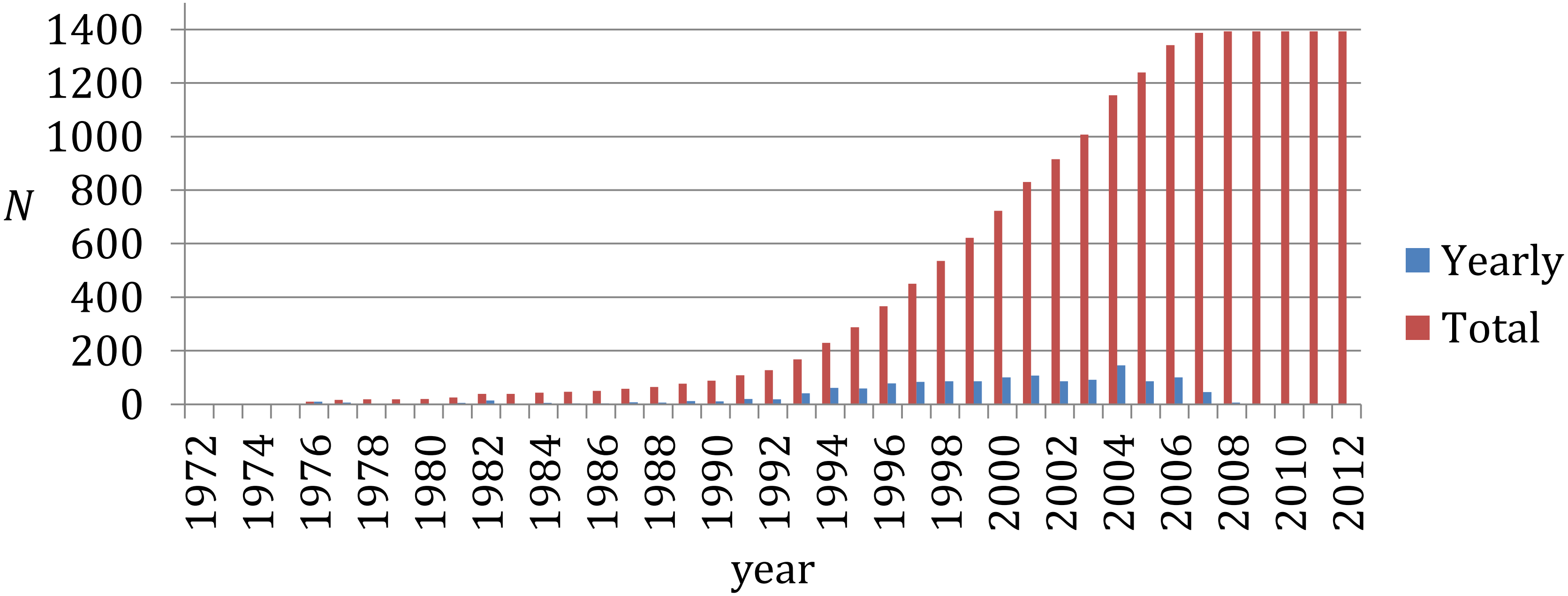
- Claim in 1990's
  - mostly when a new structure is solved (80-90%)
    - looks like a structure which was already in databank

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

- Quantified ..

# new folds per year

- How many new structures per year ?
  - source PDB web page / scop 1.75
  - count number of new "families" each year



- there are still new folds in 2012 – problem with fold definition

# Why is this interesting ?

- claim (1992) $10^3$ protein folds*
- if one has a representative for each fold
  1. should be able to model all sequences
     - solving structures is no longer necessary
       - find appropriate fold and build model
  2. if there is a known structure it is easier to solve a related structure (molecular replacement)
- common aim
  - try to solve representative of every fold

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

*Chothia, C (1992), Nature 357, 543-544

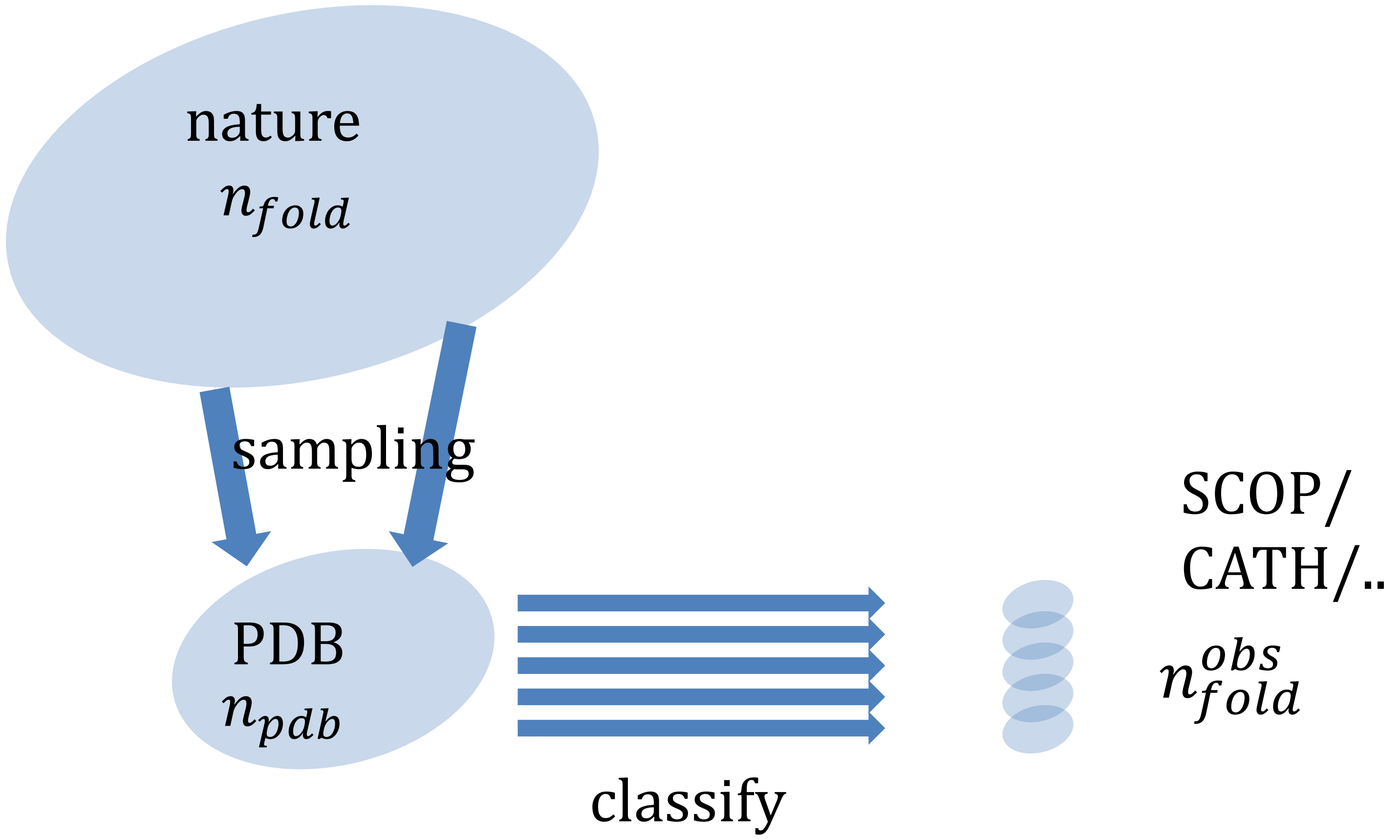
# Problem

- How many folds are there ? $n_{fold}$
- How many do we have in PDB ?
  - classify structures $n_{fold}^{obs}$

- How would you approach the problem ? Examples

  1. statistical – look at distribution of structures
  2. geometric – how many could there be

# Statistical approach



nature

$n_{fold}$

sampling

PDB

$n_{pdb}$

classify

SCOP/
CATH/..

$n_{fold}^{obs}$

# Statistical approach

- $n_{fold}$ folds in nature
- $n_{pdb}$ number of samples (structures in PDB)
- $n_{folds}^{pdb}$ number of different folds 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

# Statistical approach

1. from protein data bank (PDB)
   - survey all known structures and group them into "folds"
   - $n_{fold}^{obs}$ found PDB (of the $n_{fold}$ folds that exist)
2. step
   - visit each $i$ of $n_{fold}^{obs}$ folds and count the number of proteins with this fold
   - call this $n_{obs}(i)$ (how many proteins have fold $i$)
3. collect distribution data
   1. fold 1 has $n_1$ members, fold 2 has $n_2$ members… $n_{obs}(1), n_{obs}(2), \ldots$

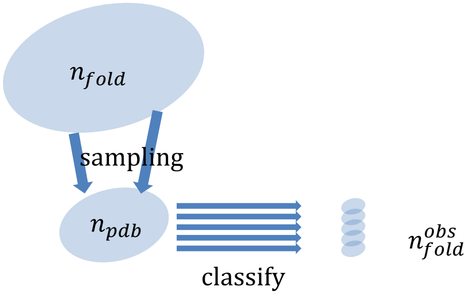
# statistical approach – very naïve

- say $10^3$ classes in nature $n_{fold} = 1\,000$
- we solve 1 000 structures $n_{pdb} = 1\,000$
  - would we seen every fold once ?
    - some folds not seen, some seen 10 times
- look at set of numbers
  - $n_{obs}(1), n_{obs}(2), \ldots$
  - if $n_{fold} = n_{pdb}$
    - $\langle n_{obs}(i) \rangle = 1$ (not so helpful)
    - variance will be big (numbers from 0 to 10)

$\langle x \rangle$ mean of $x$

# statistical approach – very naïve

- $10^6$ classes in nature $n_{fold} = 10^6$
- we have $10^3$ structures
- all structures should be different



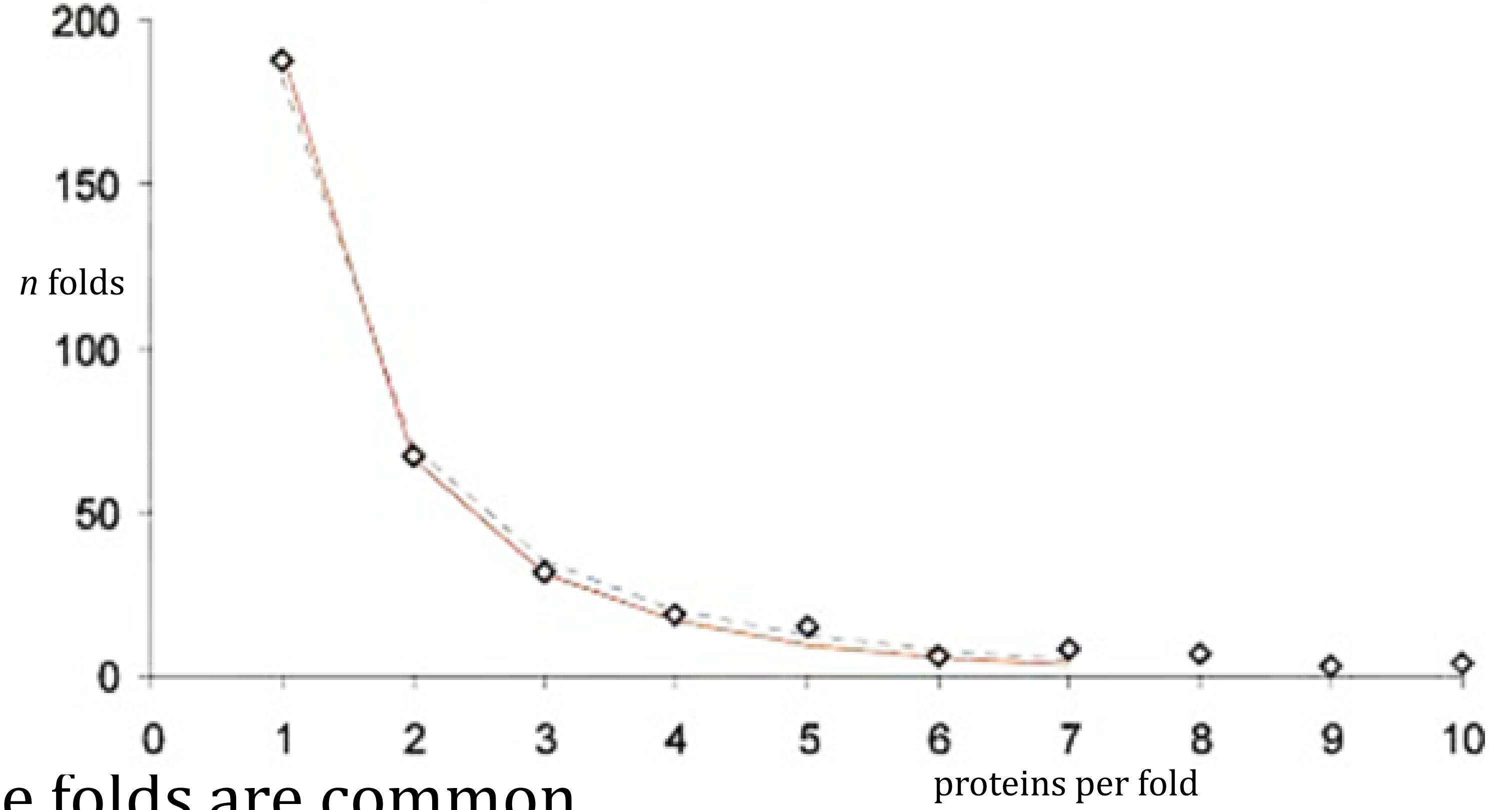
- multinomial / categorical distribution

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

- look at PDB structures
- put in classes
- look at distribution

# Results of naïve approach

- 450 classes in one estimate
  - silly

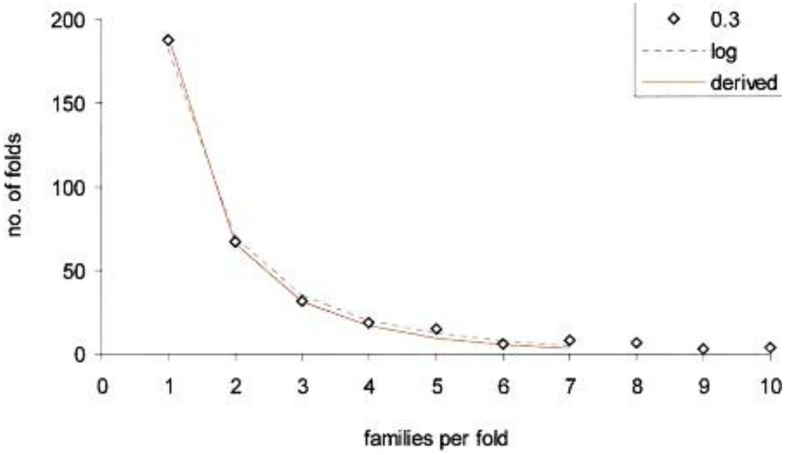


  - some folds are common
  - some are rare

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



# statistical approach - better

- general form of distribution
- $P(\lambda_i, n_{obs}) = \binom{n_{pdb}}{n_{obs}} (\lambda_i)^{n_{obs}} (1 - \lambda_i)^{n_{pdb} - n_{obs}}$
- $\lambda_i$
  - probability of fold
    (how many balls of a colour were in my bag at start)
  - values are not known
  - we just see a set of relative $\lambda_i$
- sort the list of populations of classes and fit parameters

# statistical version – results

- 3 756 folds
  - used folds defined by a literature classification
  - tried other statistical models
  - other definitions lead to different numbers
- 1 000 folds
  - different definitions, similar method
  - about 300 known (data from 2 000)

Govindarajan, S, Recabarren, R, Goldstein, (1999) R.A. Proteins, 35, 408-414
Wolf, Y.I., Grishin, N.V., Koonin, E.V. (2000) J. Mol. Biol. 299, 897-905

# statistical - summary

- Estimates vary from 1 000 to 4 000 (and more)
    - few estimates of 8 000

Problems

- what is distribution of proteins over folds ?
    - leads to question .. why ?
- is the PDB a fair sampling ? Lots of
    - human proteins
    - structural genomics proteins
    - soluble proteins
    - proteins related to diseases (in host or agent)
    - proteins are easier if they are similar to a known one

# 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

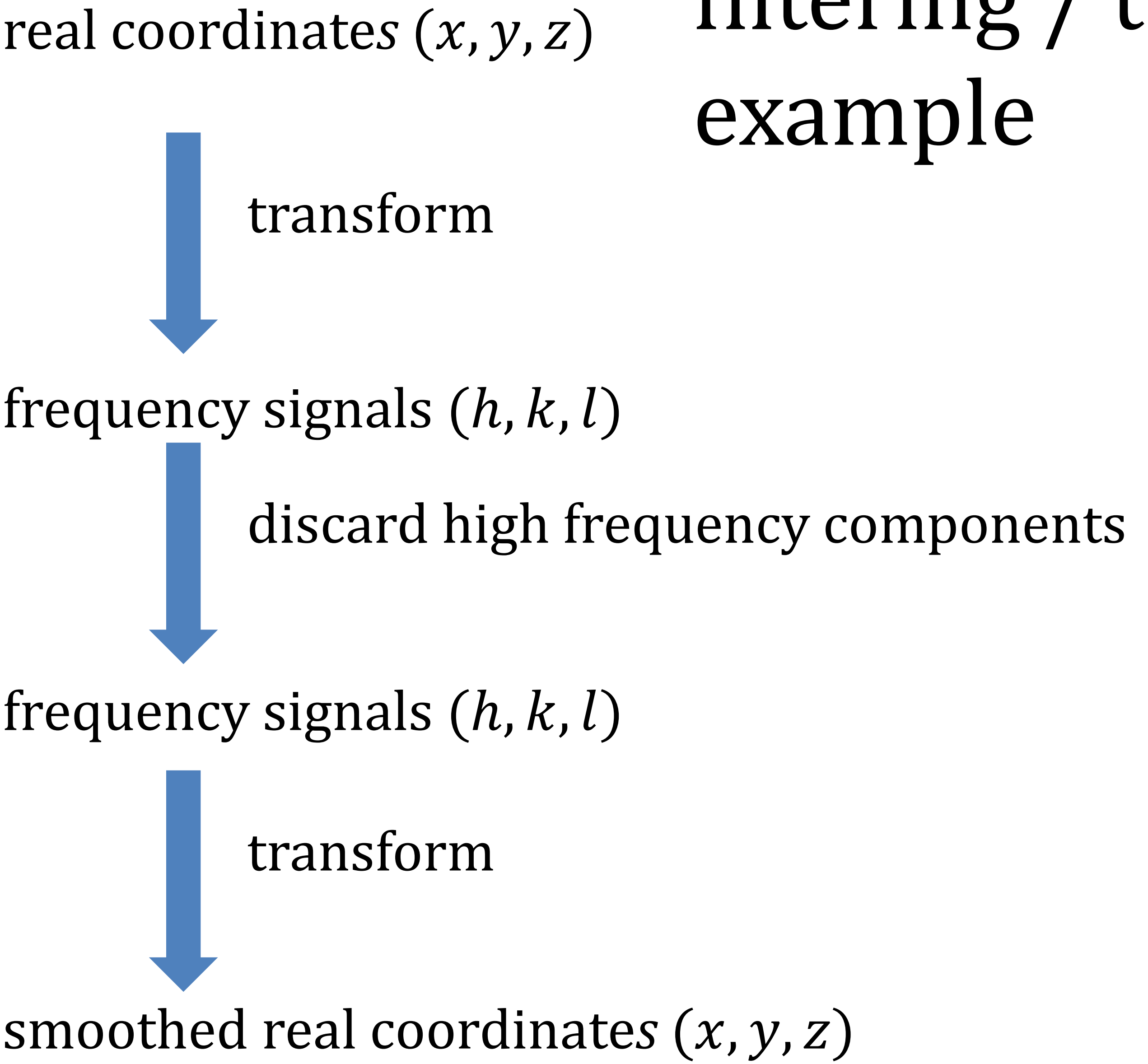
- sample from possible chains fulfilling these conditions
  - can you sample from $x, y, z$ ? Not easily
- work in a different space

# cosine transform - diversion

- Fourier transform – well known
  - go from real space to frequency space
  - or from frequency space to real
- "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

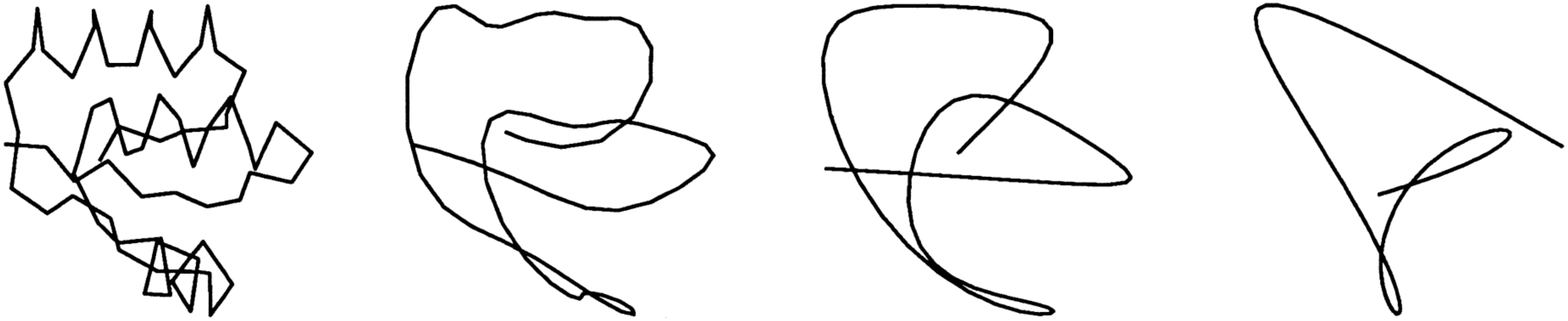
frequency signals $(h, k, l)$

transform

smoothed real coordinates $(x, y, z)$

# Example transform

- 1ctf ribosomal protein
  - transform → frequencies
  - keep only 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

# 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 $n$th 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)

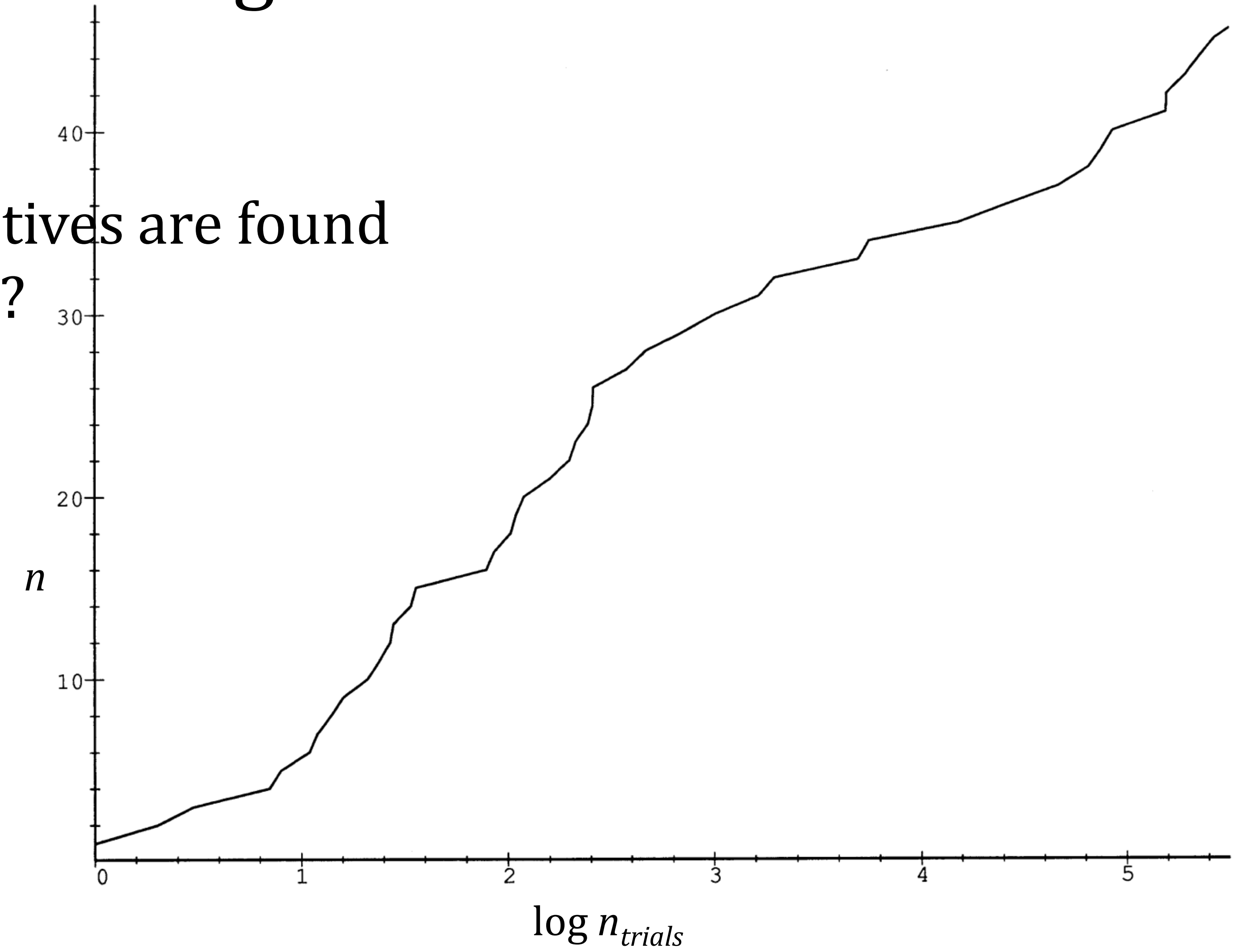
# Sampling from real coordinates

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

```
decide on $N$ (level of detail) and $n_r$ number residues
while (step < max_step)
        pick random $\hat{x}_k, \hat{y}_k, \hat{z}_k$
                (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
```

# Finding new structures

- how many representatives are found with $n_{trials}$ ?

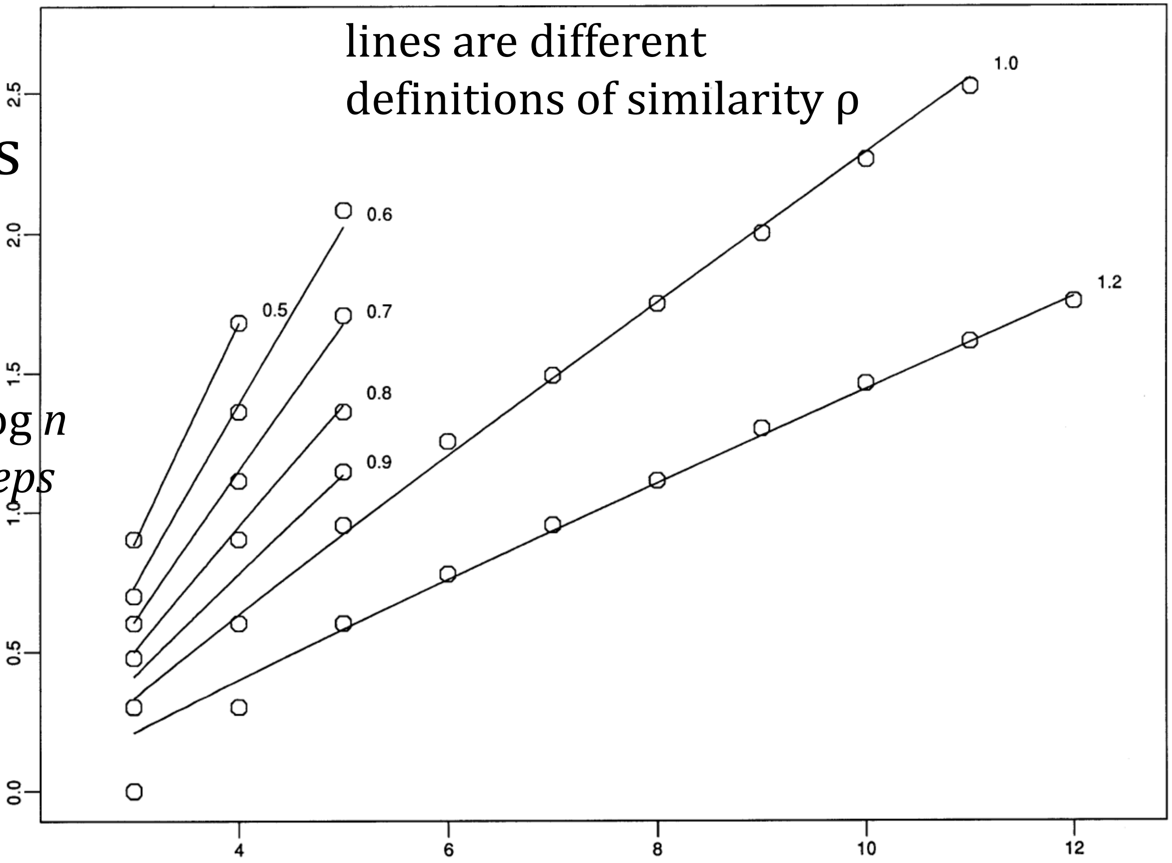


Crippen, G.M., & Maiorov, V. (1995) J. Mol. Biol. 252, 144-151

# Estimating number of folds

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

- fit to slightly arbitrary form



lines are different definitions of similarity ρ

log *n reps*

*N*

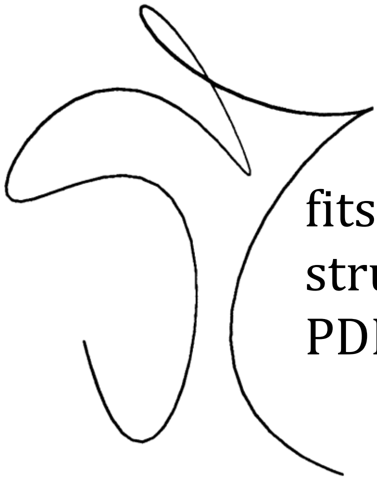
Crippen, G.M., & Maiorov, V. (1995) J. Mol. Biol. 252, 144-151

# How many folds ?

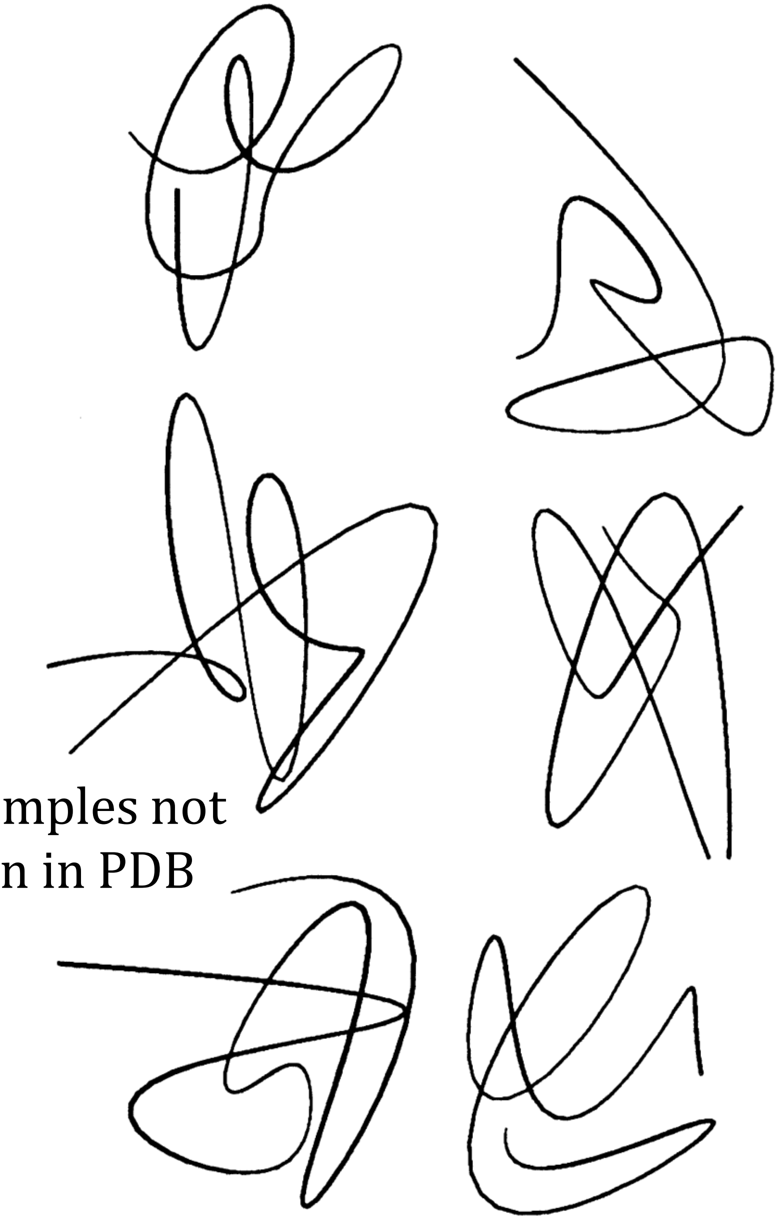
- as many as you want
  - $10^3$ smaller structures (50 residues)
  - very big numbers for larger structures
- many structures generated are similar to natural ones
- many may not be possible
  - representation a bit crude, does not capture enough detail
- may have found some structures that have not 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