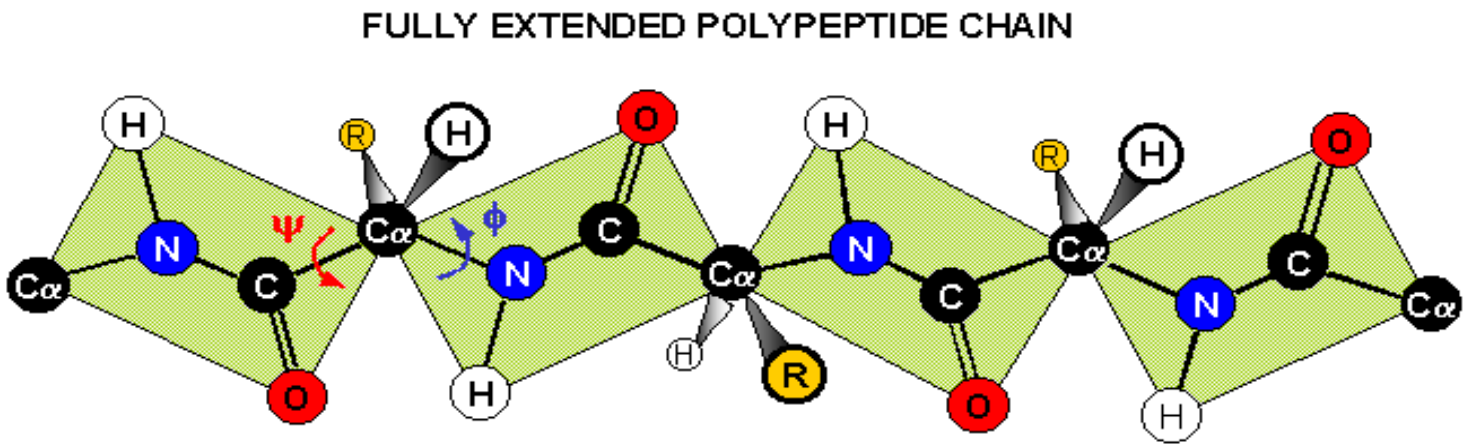
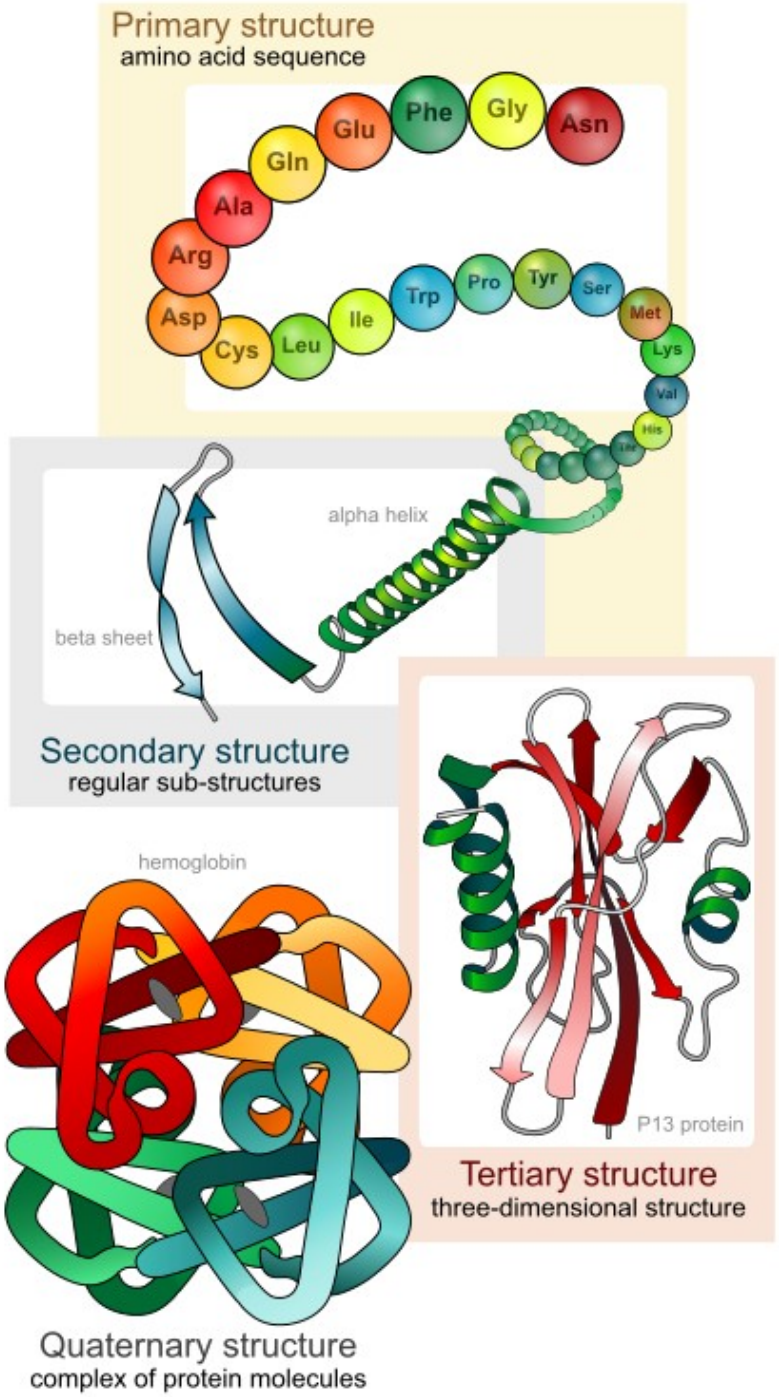
Protein Structures

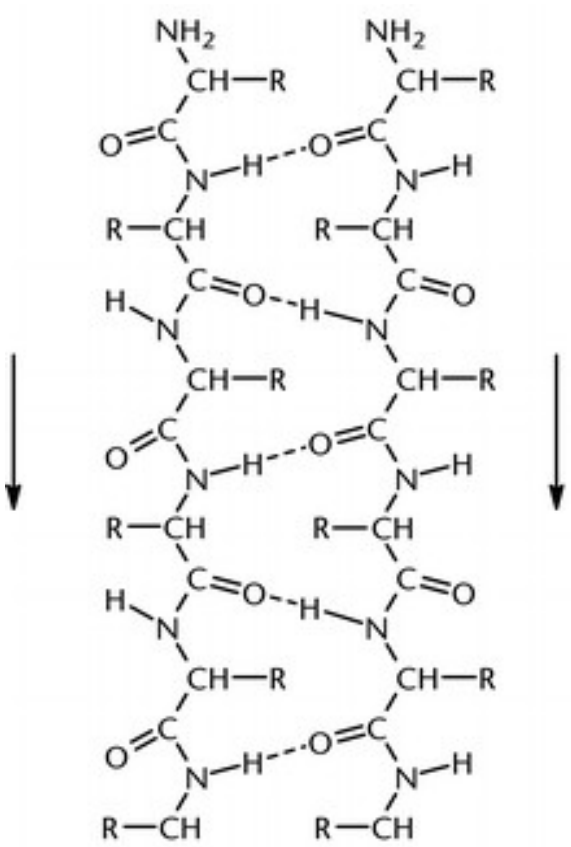
Protein as polygonal curve in 3D



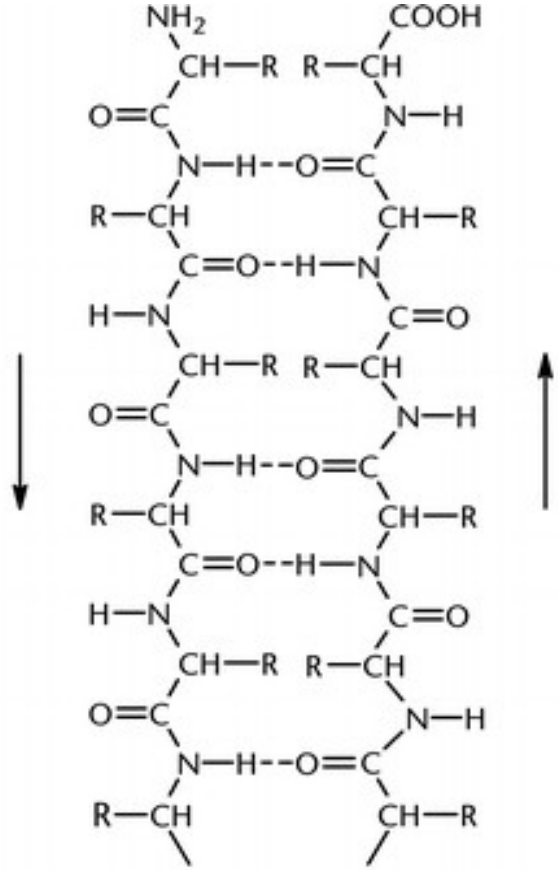
- Each residues is associated to two dihedral angles between two rigid planes.
- These two angles per site contain enough information to encode the curve completely.


Primary structure
amino acid sequence
Phe
Gly
Asn
Glu
Gln
Ala
Arg
Asp
Cys
Leu
Ile
Trp
Pro
Tyr
Ser
Met
Lys
Val
His
alpha helix
beta sheet
Secondary structure
regular sub-structures
P13 protein
Tertiary structure
three-dimensional structure
hemoglobin
Quaternary structure
complex of protein molecules

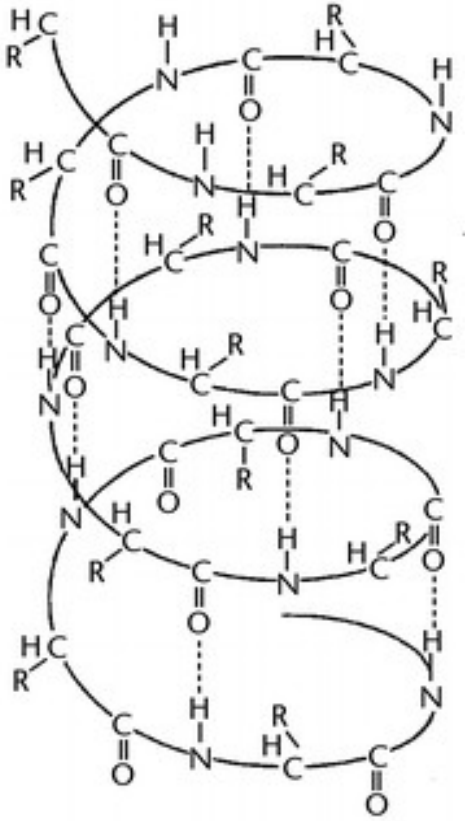
Secondary Structures



Parallel β pleated sheet

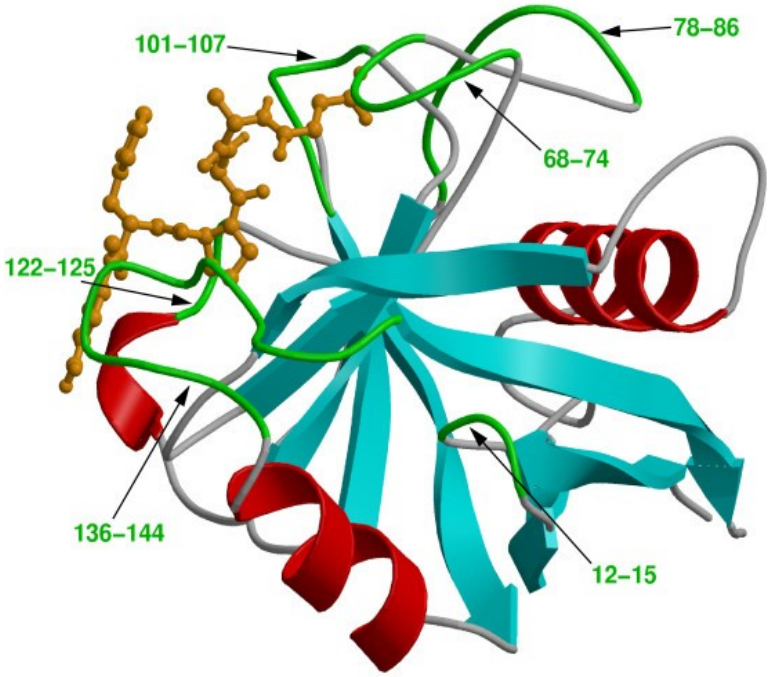


Antiparallel β pleated sheet



Right-handed α helix

Secondary Structures



- Spontaneous (form first)
- Rigid
- Cooperative (all or nothing)
- Scaffold to the tertiary structure.

Conservation

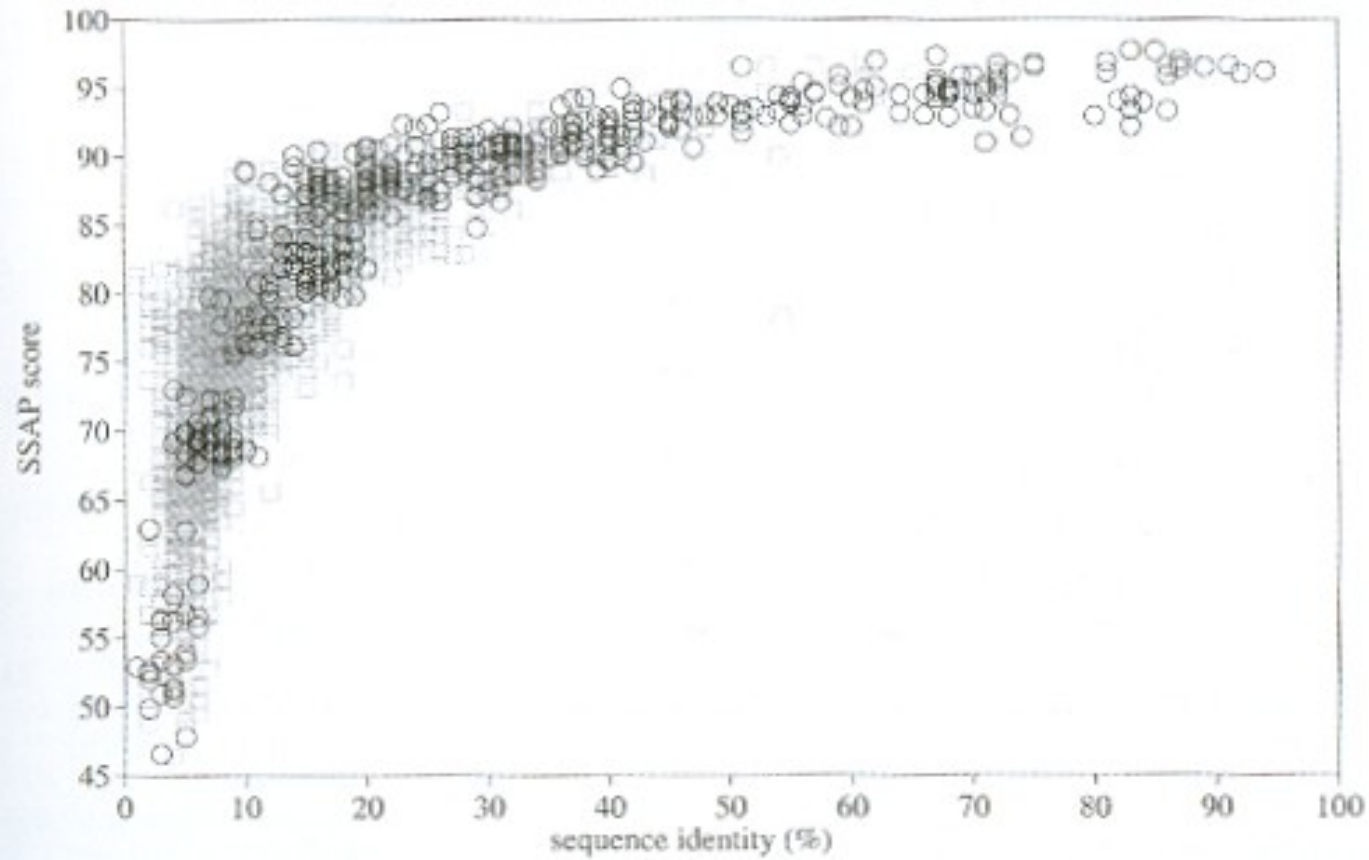
Nucleotide << Protein << Structure

Recent/closely related

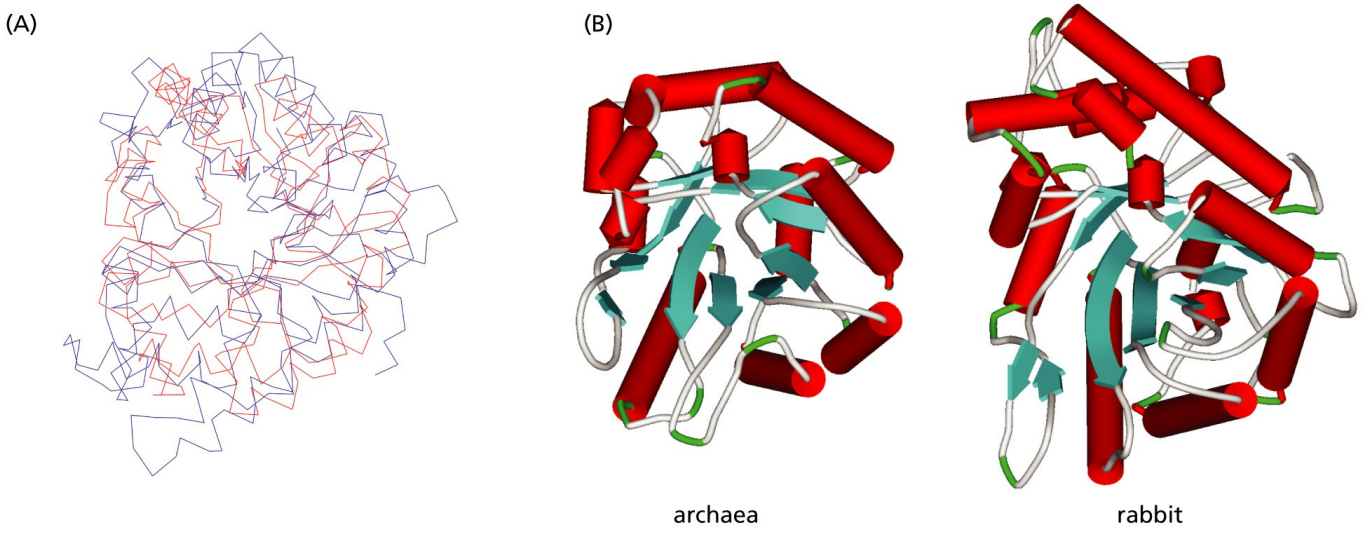
Ancient/distantly related

Comparing protein structures

Structures are evolving at a much slower rate!



Similar structure can infer same function (even with low seq. conservation)



(C)

```
Rabbit :     1 KELSDIAHRIVA-PGKGILAADESTGSIAKRLQSIGTENTEENRRFYRQLLLTADDRVNPCIGGVILFHE
               111111111111 1111111111111111111111           111111111 1111  111111111
Archaea:     1 NLTEKFLRIFARRGKSIILAYDHGIEHGPADFMDNPDS-------ADPEYILRLARDAG--FDGVVFQRG

Rabbit :    70 TLYQKADDGRPFPQVIKSKGGVVGIKVDKGVVPLAGTNGETTTQGLDGLSERCAQYKKDGADFAKWRCVL
               11111       11111      1111111111        1111111       1111111111111111111
Archaea:    62 IAEKY------YDGSV-----PLILKLNGKTTLYN---GEPVSVANC----SVEEAVSLGASAVGYTIYP

Rabbit :   140 KIGEHTPSALAIMENANVLARYASICQQNGIVPIVEPEILPDGDHDLKRCQYVTEKVLAAVYKALSDHHI
                     1 1111111111111111111111111111111     11  1111111111111111
Archaea:   114 -------GSGFEWKMFEELARIKRDAVKFDLPLVVWSYPRGGKVVNE-TAPEIVAYAARIALELGA----

Rabbit :   210 YLEGTLLKPNMVTPGHACTQKYSHEEIAMATVTALRRTVPPAVTGVTFLSGGQS--EEEASINLNAINKC
                  111111111               1111111111 11111  1111111111  11111111111111
Archaea:   172 ----DAMKIKYTG-------------DPKTFSWAVK-VAGKV--PVLMSGGPKTKTEEDFLKQVEGVLEA

Rabbit :   278 PLLKPWALTFSYGRALQASALKAWGGKKENLKAAQEEYVKRALANSLACQ
                    111111111111111                    11111111111111
Archaea:   222 -----GALGIAVGRNVWQRR----------------DALKFARALAELVY

Sequence id = 19.7%
```

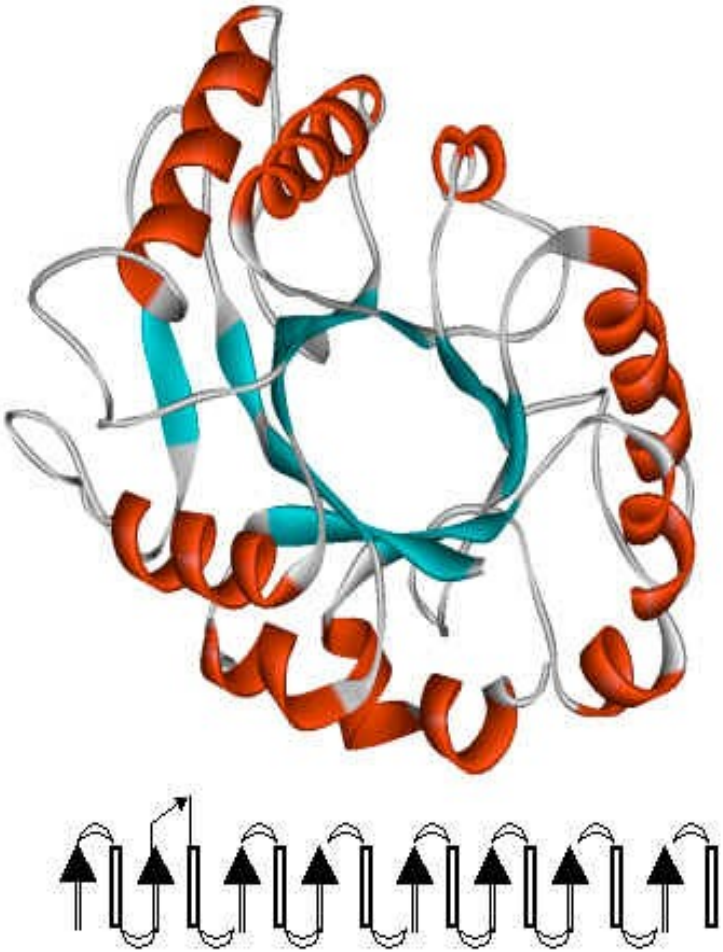
Similar structure but different function

TIM barrels

Most common structure

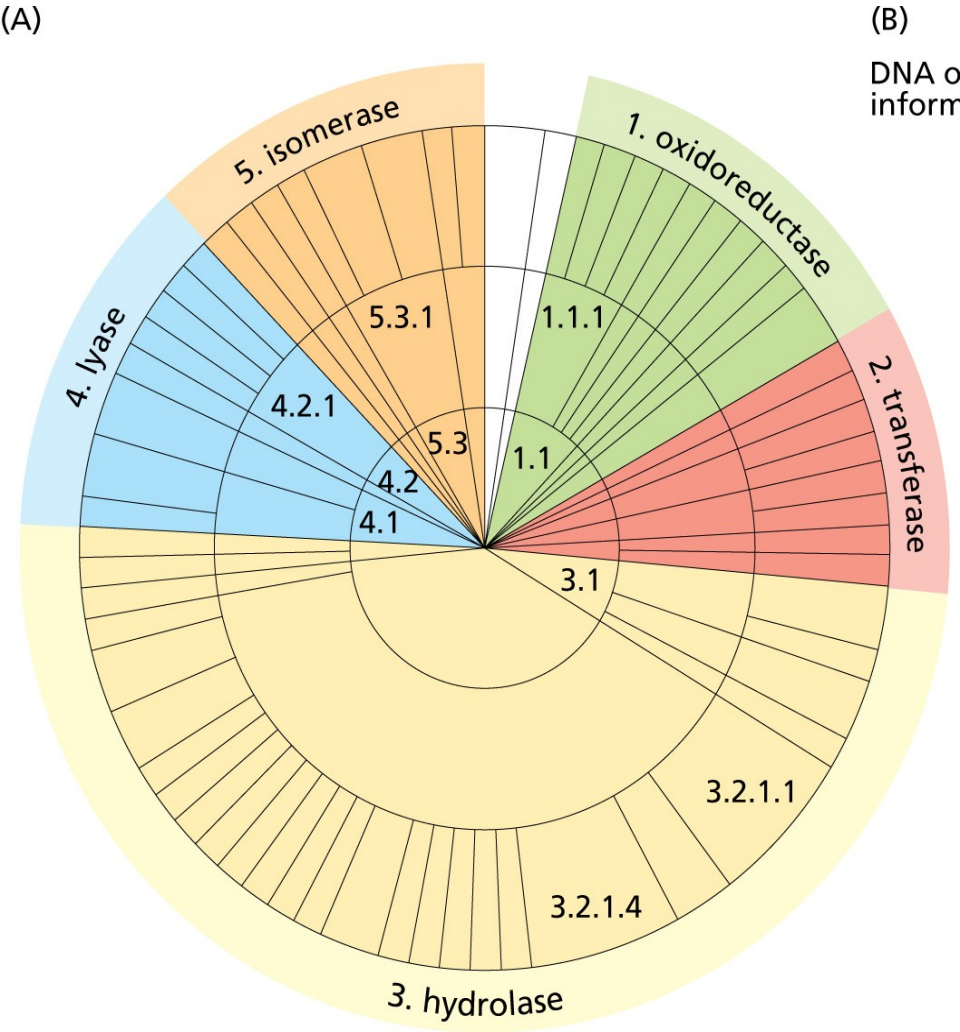
8 alpha helices/

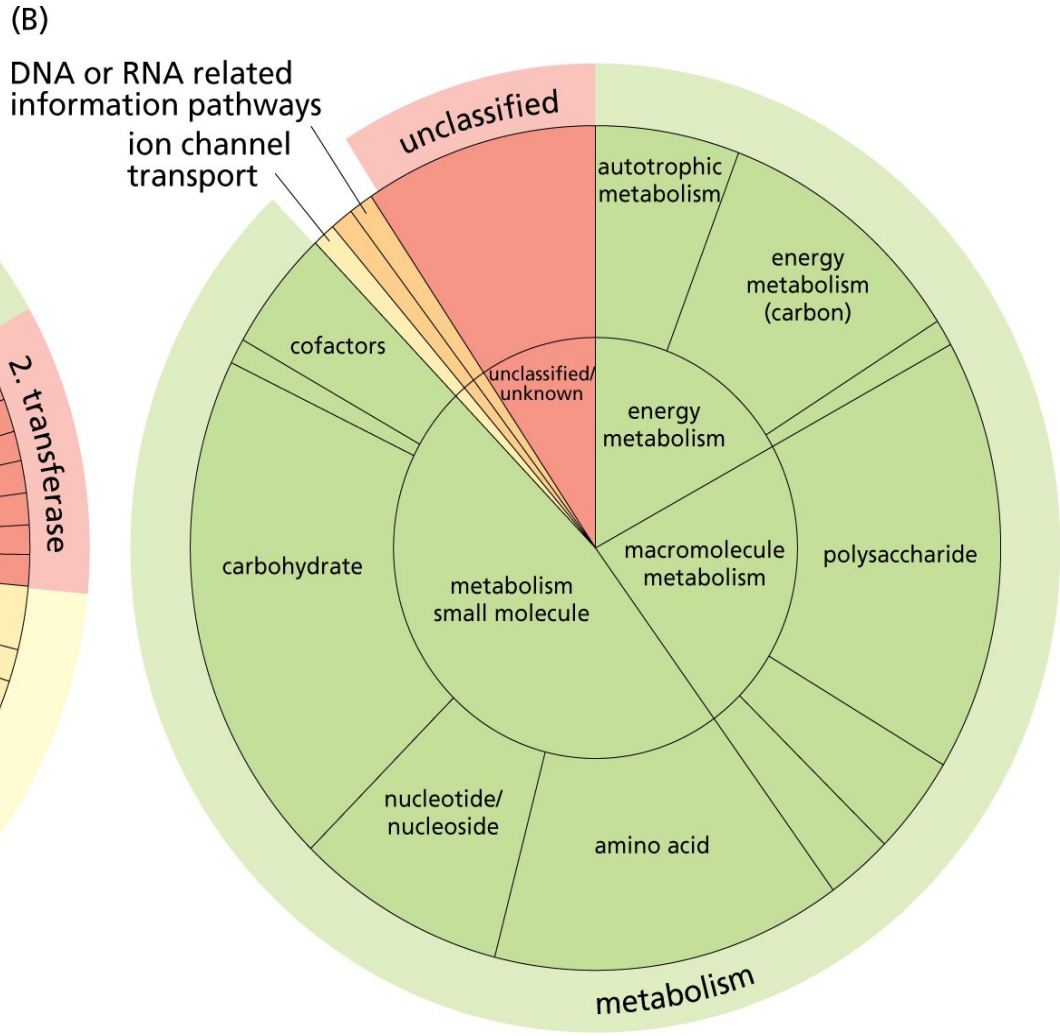
beta sheets



Topology diagram of Hevamine - one of the TIM barrel structures

Functional distribution of TIM fold

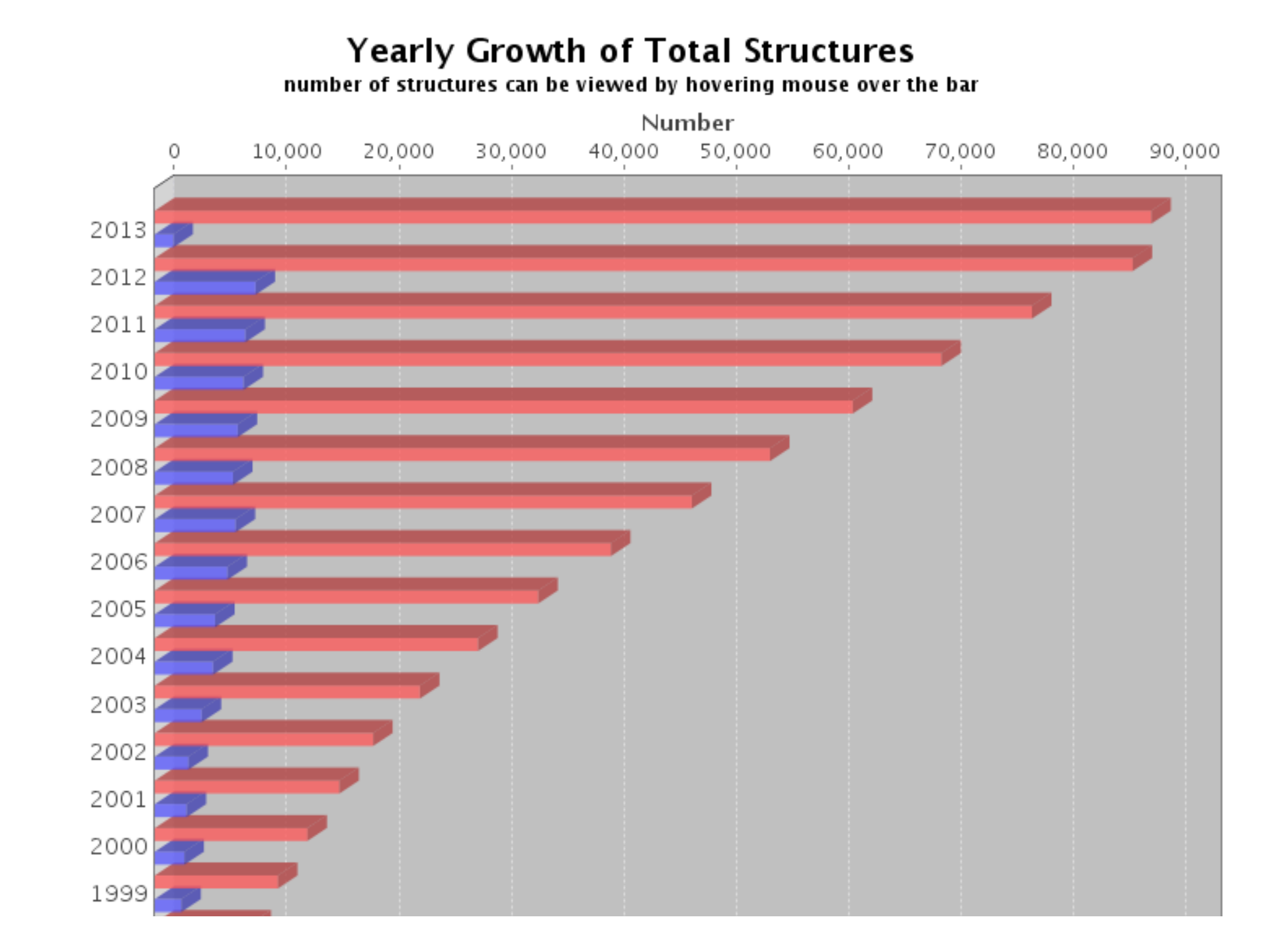




PDB

Protein **D**ata **B**ank

- The PDB is the primary repository of protein structure data.
- Contains only “real” protein structures
- Those solved using XRay Crystallography or Nuclear Magnetic Resonance (NMR)
- Continous growth, but slower pace when compared to sequencing


Yearly Growth of Total Structures
number of structures can be viewed by hovering mouse over the bar
Number
0
10,000
20,000
30,000
40,000
50,000
60,000
70,000
80,000
90,000
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999

The PDB format

- Flat file, column oriented
- Human readable
- **Human editable**
- Huge legacy problems

Flat File: A datafile without indexing structure or hierarchy. In contrast, to *relational database*, or *data grammar.*

Header

```
HEADER    IMMUNOGLOBULIN                          01-MAR-93   2IMM      2IMM   2
COMPND    IMMUNOGLOBULIN VL DOMAIN (VARIABLE DOMAIN OF KAPPA LIGHT      2IMM   3
COMPND   2 CHAIN) OF MCPC603                                            2IMM   4
SOURCE    HUMAN (HOMO $SAPIENS) RECOMBINANT SYNTHETIC M603 GENE         2IMM   5
AUTHOR    B.STEIPE,R.HUBER                                              2IMM   6
REVDAT   1   15-JUL-93 2IMM    0                                        2IMM   7
REMARK   1                                                             2IMM   8
REMARK   1 REFERENCE 1                                                 2IMM   9
REMARK   1  AUTH   B.STEIPE,A.PLUCKTHUN,R.HUBER                        2IMM  10
REMARK   1  TITL   REFINED CRYSTAL STRUCTURE OF A RECOMBINANT          2IMM  11
REMARK   1  TITL 2 IMMUNOGLOBULIN DOMAIN AND A                         2IMM  12
REMARK   1  TITL 3 COMPLEMENTARITY-DETERMINING REGION 1-GRAFTED MUTANT 2IMM  13
REMARK   1  REF    J.MOL.BIOL.                   V. 225   739 1992     2IMM  14
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070   2IMM  15

    [...]

REMARK   2                                                             2IMM  23
REMARK   2 RESOLUTION. 2.00  ANGSTROMS.                                2IMM  24
REMARK   3                                                             2IMM  25

   [...]
```

Seqres

```
[...]
SEQRES   1     114  ASP ILE VAL MET THR GLN SER PRO SER SER LEU SER VAL  2IMM  35
SEQRES   2     114  SER ALA GLY GLU ARG VAL THR MET SER CYS LYS SER SER  2IMM  36
SEQRES   3     114  GLN SER LEU LEU ASN SER GLY ASN GLN LYS ASN PHE LEU  2IMM  37
SEQRES   4     114  ALA TRP TYR GLN GLN LYS PRO GLY GLN PRO PRO LYS LEU  2IMM  38
SEQRES   5     114  LEU ILE TYR GLY ALA SER THR ARG GLU SER GLY VAL PRO  2IMM  39
SEQRES   6     114  ASP ARG PHE THR GLY SER GLY SER GLY THR ASP PHE THR  2IMM  40
SEQRES   7     114  LEU THR ILE SER SER VAL GLN ALA GLU ASP LEU ALA VAL  2IMM  41
SEQRES   8     114  TYR TYR CYS GLN ASN ASP HIS SER TYR PRO LEU THR PHE  2IMM  42
SEQRES   9     114  GLY ALA GLY THR LYS LEU GLU LEU LYS ARG              2IMM  43
[...]
```

Actual structure

ATOM 1 N LYS A 3 22.090 44.427 -4.959 1.00 69.21 N

ATOM 2 CA LYS A 3 21.478 45.792 -5.038 1.00 70.52 C

ATOM 3 C LYS A 3 22.509 46.914 -4.869 1.00 70.66 C

ATOM 4 O LYS A 3 22.928 47.206 -3.746 1.00 71.84 O

ATOM 5 CB LYS A 3 20.396 45.941 -3.967 1.00 69.67 C

ATOM 6 CG LYS A 3 19.521 47.167 -4.143 1.00 67.14 C

ATOM 7 CD LYS A 3 18.180 46.992 -3.440 1.00 67.94 C

ATOM 8 CE LYS A 3 17.268 46.025 -4.181 1.00 64.21 C

ATOM 9 NZ LYS A 3 15.872 46.531 -4.271 1.00 61.28 N

ATOM 10 N LYS A 4 22.905 47.530 -5.988 1.00 67.76 N

ATOM 11 CA LYS A 4 23.886 48.628 -6.007 1.00 63.08 C

ATOM 12 C LYS A 4 23.138 49.953 -5.859 1.00 58.88 C

ATOM 13 O LYS A 4 22.747 50.573 -6.846 1.00 57.69 O

ATOM 14 CB LYS A 4 24.660 48.629 -7.332 1.00 62.71 C

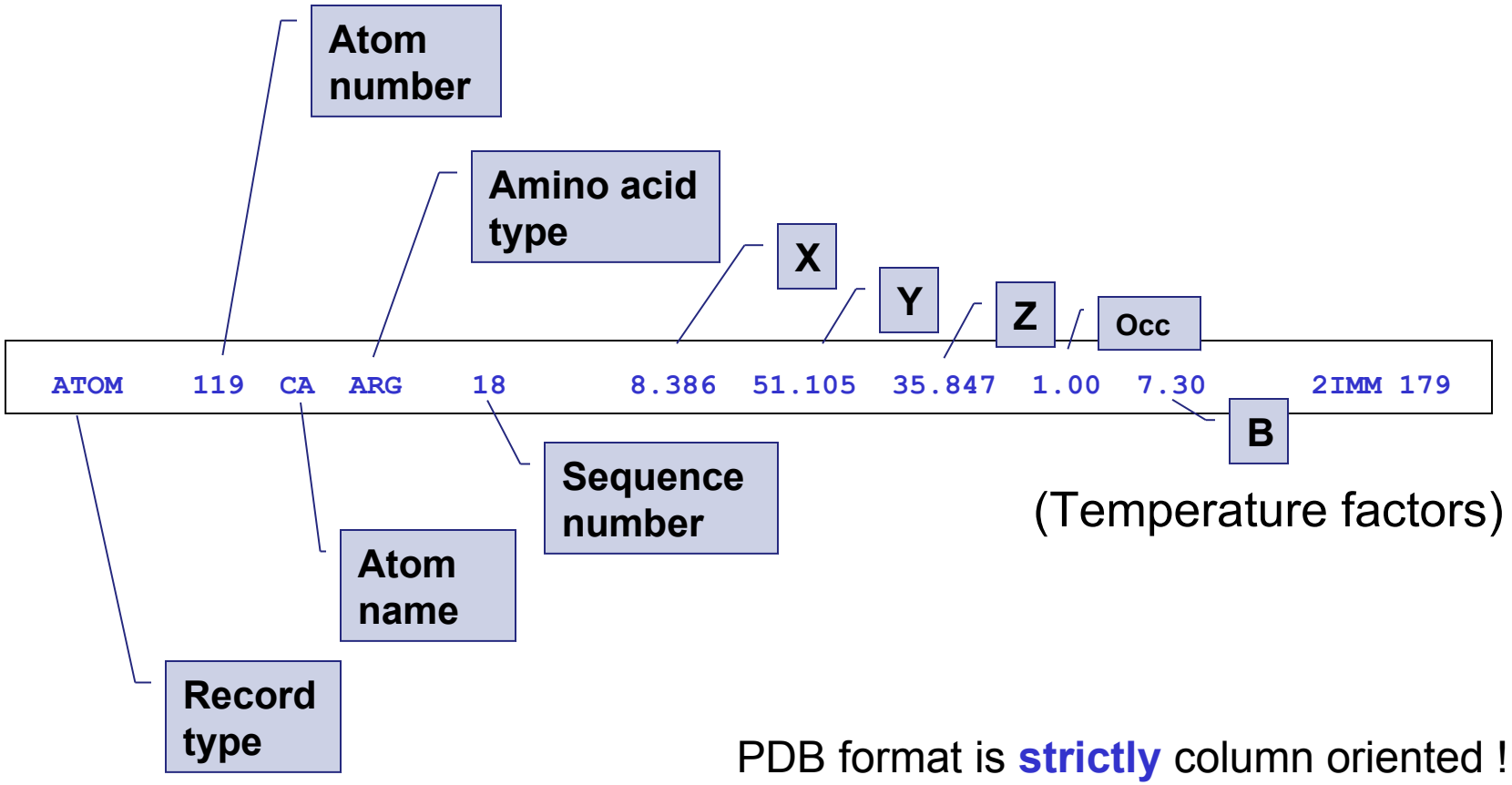
ATOM 15 CG LYS A 4 25.105 47.259 -7.828 1.00 60.48 C

ATOM 16 CD LYS A 4 26.108 47.402 -8.964 1.00 62.38 C

ATOM 17 CE LYS A 4 26.398 46.077 -9.658 1.00 66.66 C

ATOM 18 NZ LYS A 4 26.344 44.911 -8.734 1.00 71.30 N

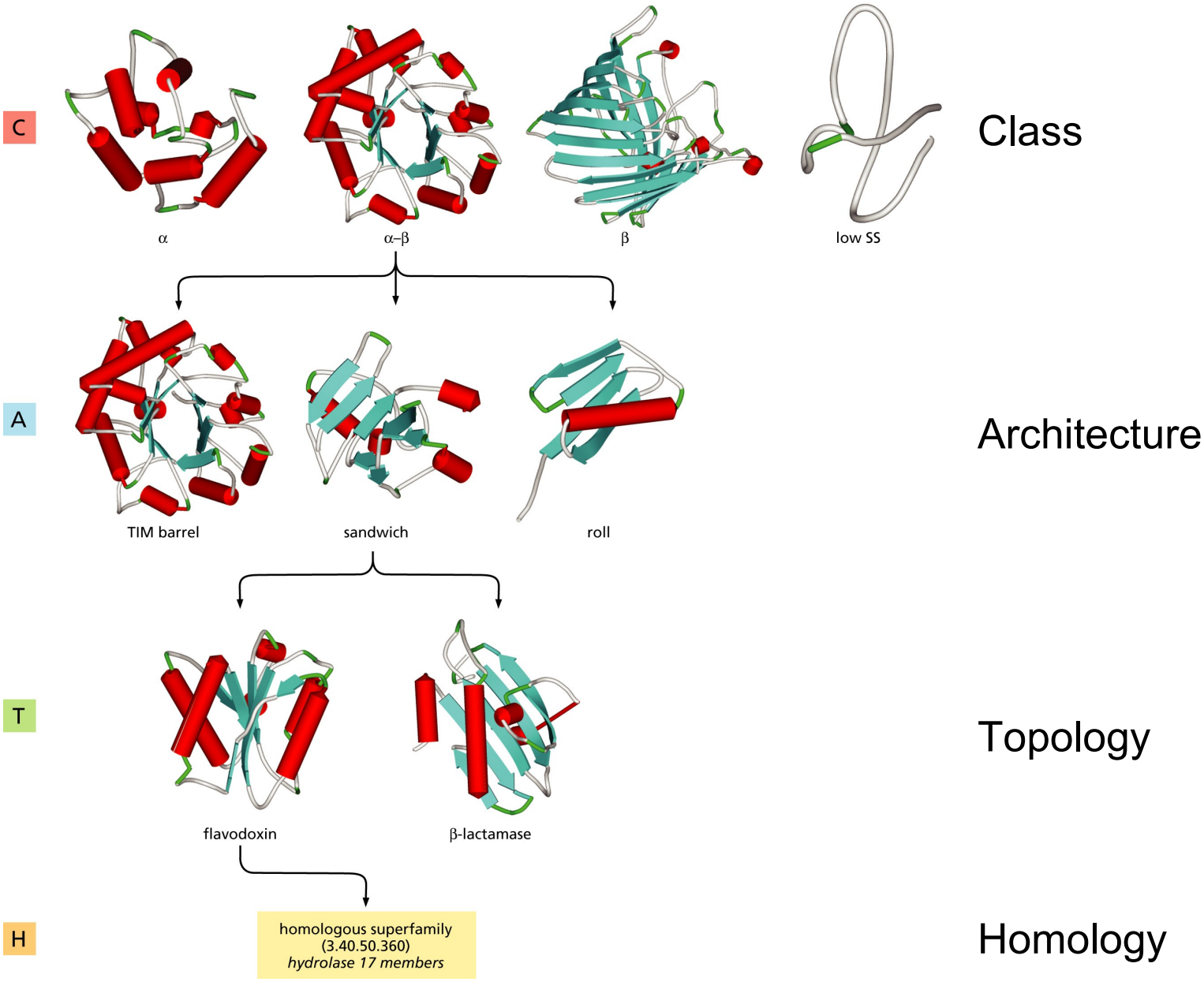
Atom

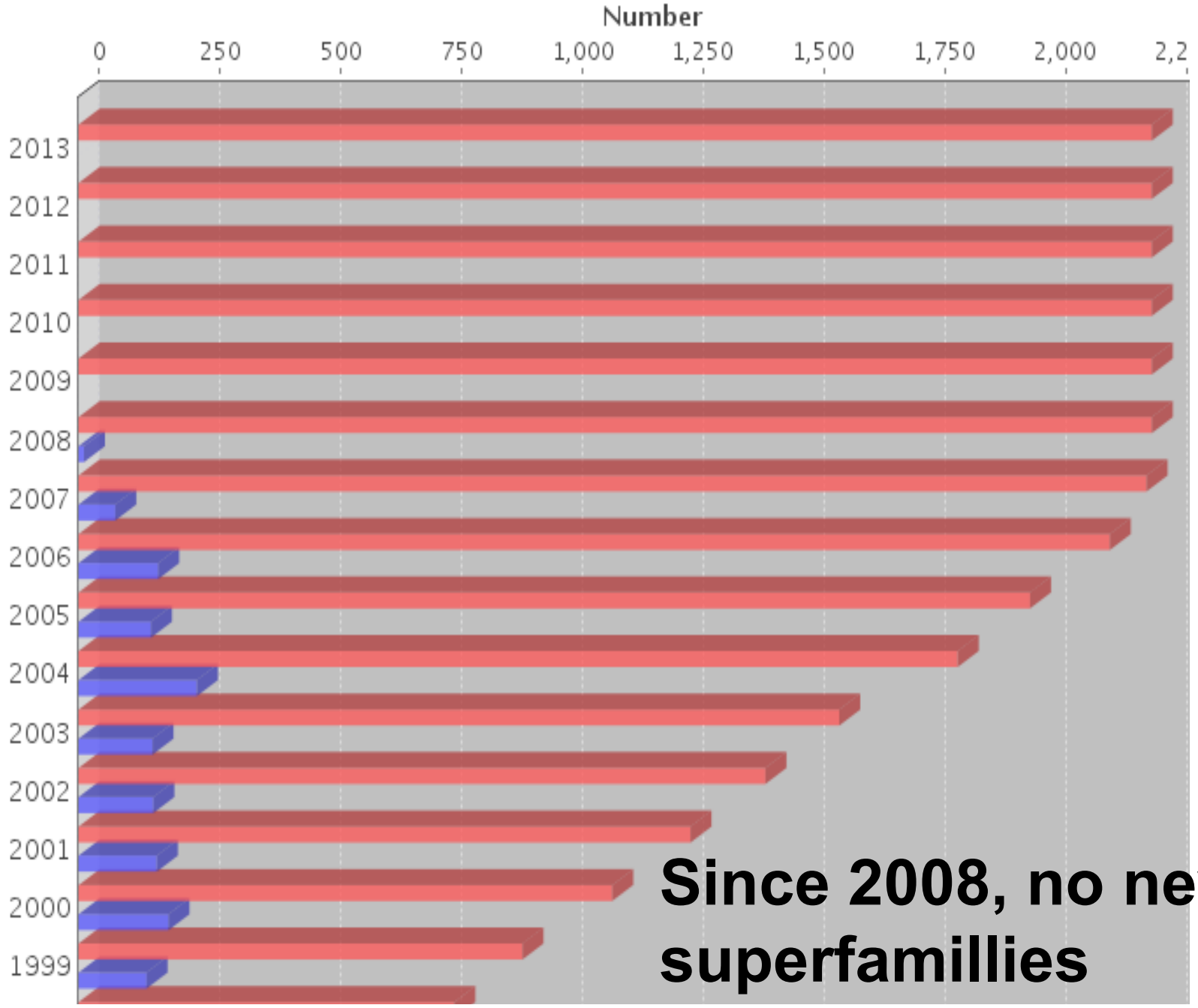


(Temperature factors)

PDB format is **strictly** column oriented !

CATH hierarchy





Since 2008, no new superfamillies were discovered.

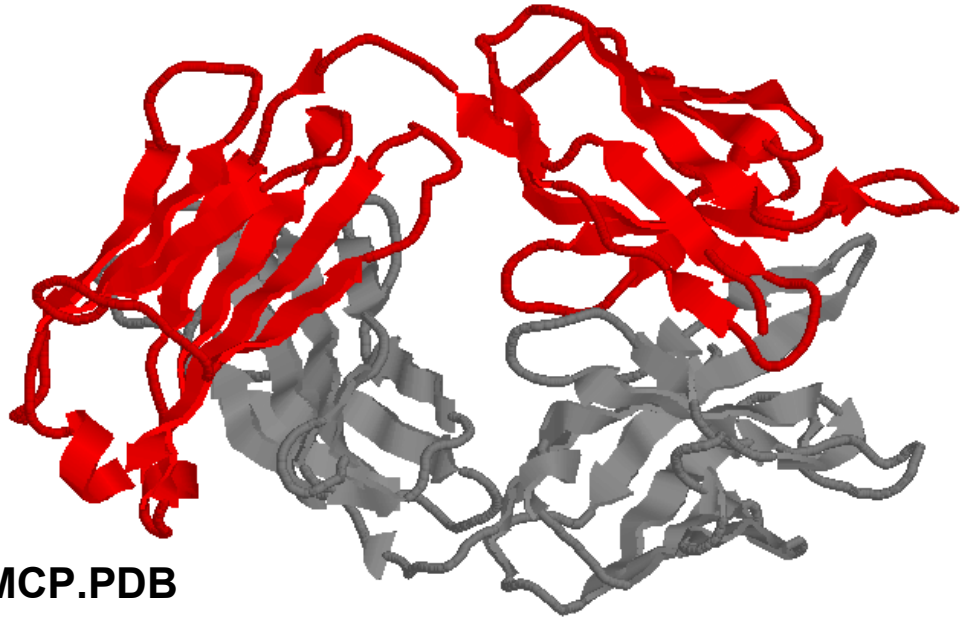
Domains

Domains are folding units, functional units, and units of inheritance.

Domains are ubiquitous in proteins

Large proteins are composed of compact, semi-independent units - domains.

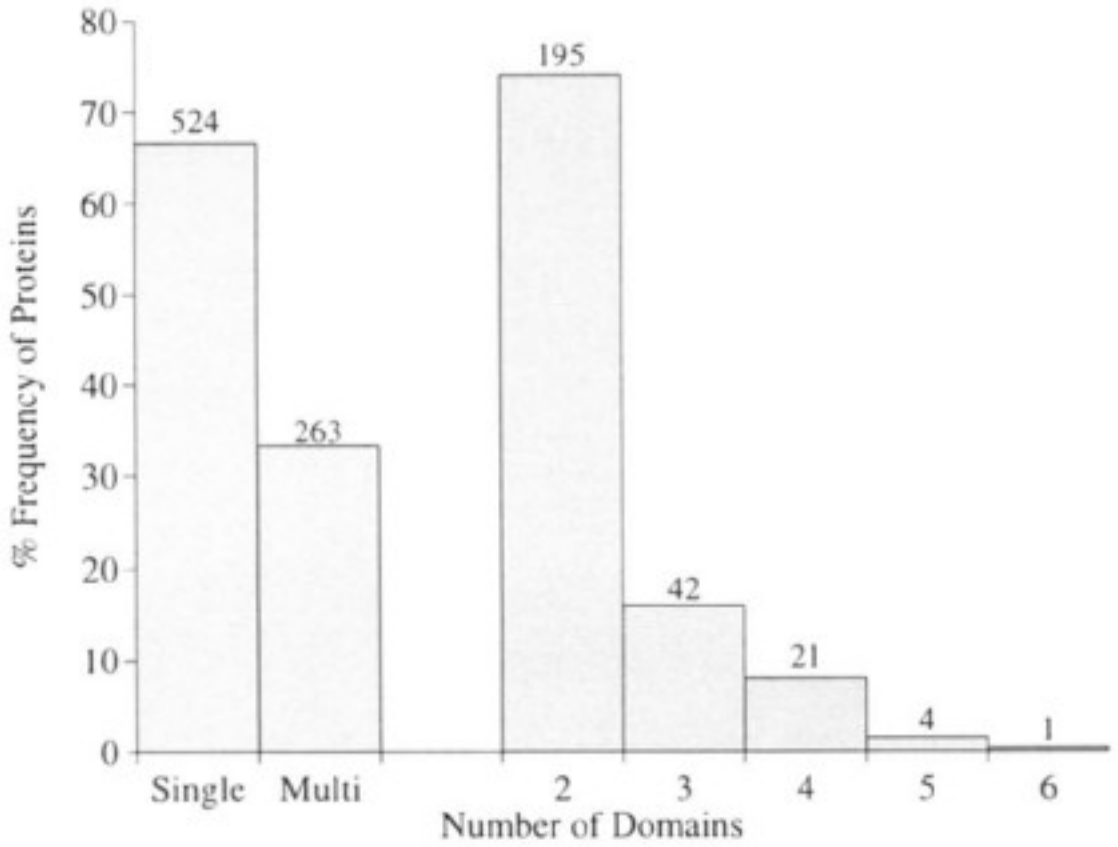
Reason:
Modularity
Folding efficiency



2MCP.PDB

Domains in proteins:

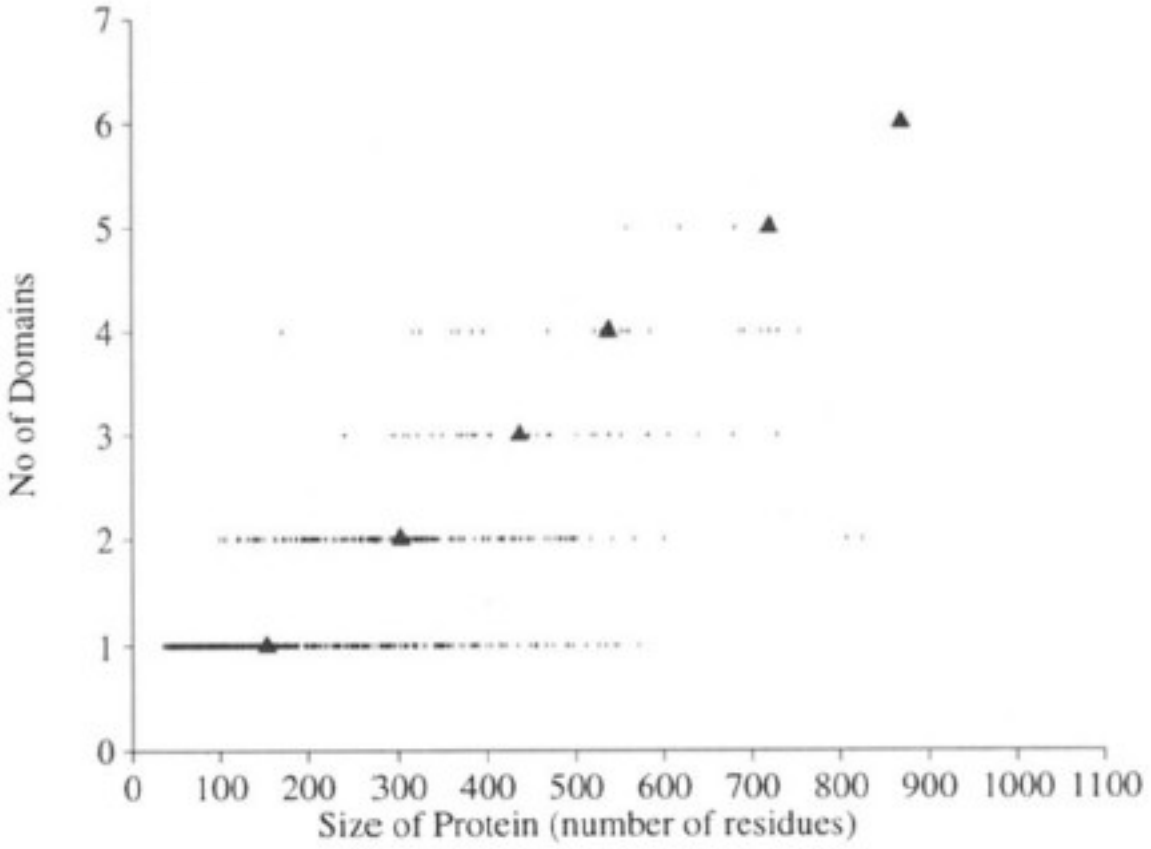
Number of domains in 787 representative proteins used as the basis for the CATH database



Jones S et al. (1998)
Protein Science **7**:233

Domains in proteins:

Non-random relationship between domain number and chain length in the 787 representative proteins used as the basis for the CATH database



Jones S et al. (1998)
Protein Science **7**:233

Predicting Structure

Three Paths to Protein Structure Prediction

Homology Modeling – requires homology to protein with known structure

Threading (Fold recognition) – uses known folds to predict structure

***Ab initio* prediction** – hardest case, not using any prior information

Threading Database Search

Premise is that most sequences match some 3-D structure that is already known

Given a database of known 3-D protein folds:

align the test sequence to each known protein

in real 3-D coordinate space (slow but exact)

in parameterized 1-D space (fast but approximate)

optimize some scoring function

sort out best sequence-structure alignment

assess alignments

Ab initio Prediction

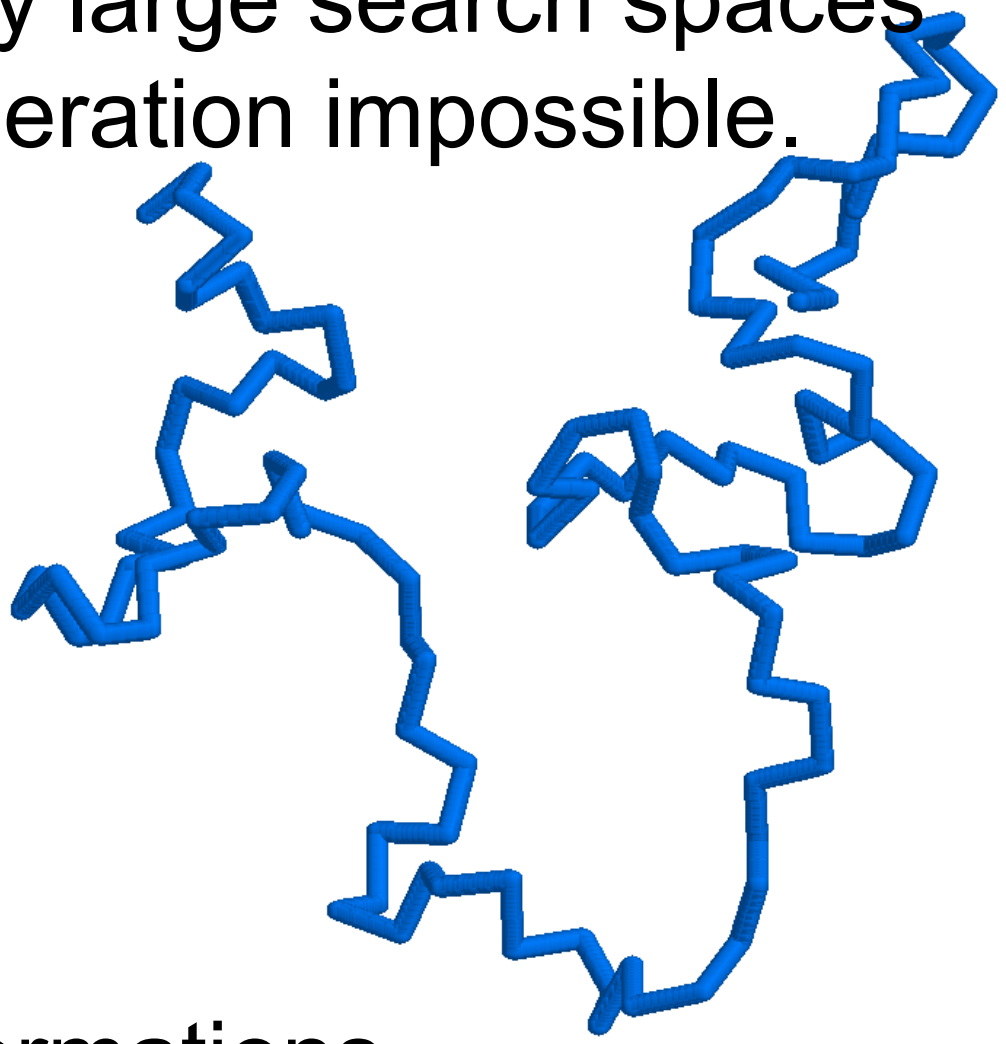
The “Holy Grail” of bioinformatics!

The assumption:
Native structure is a global energy minimum

The algorithm:

1. **Reasonably generate all conformations**
2. **Score with an appropriate scoring function**
3. **Choose the one with best score**

Combinatorially large search spaces make enumeration impossible.



Consider:

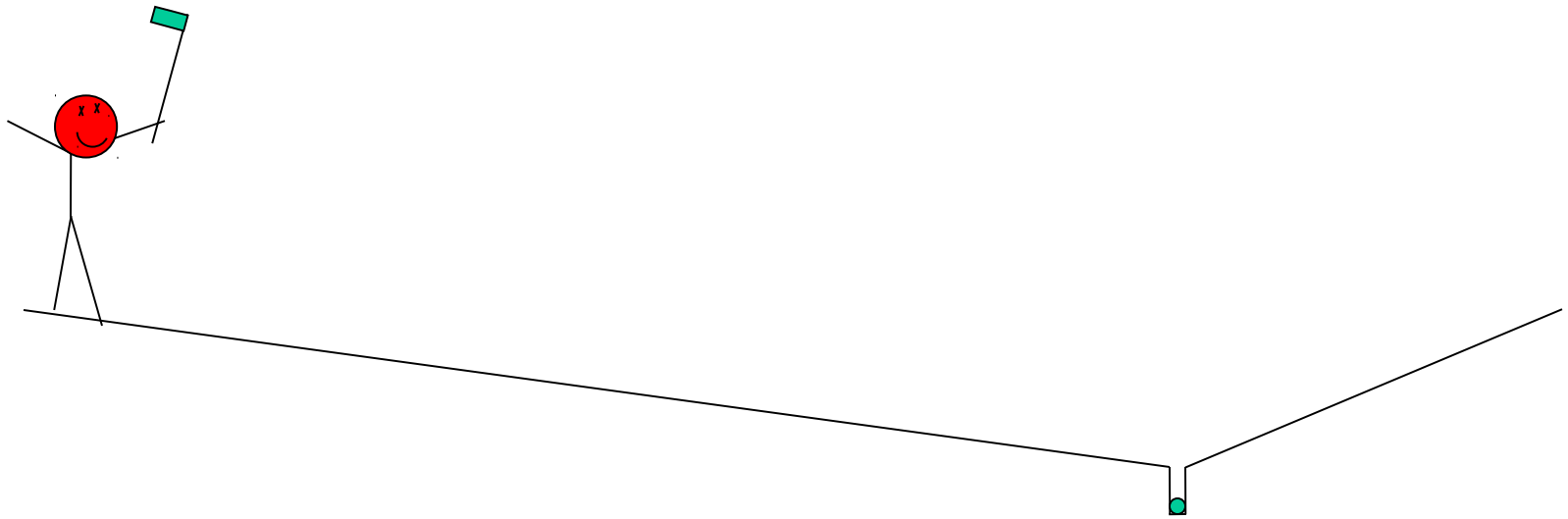
100 residues
3 states:

$3^{100} \approx 10^{47}$ conformations

A Blind Golfer's view of global optimization: I

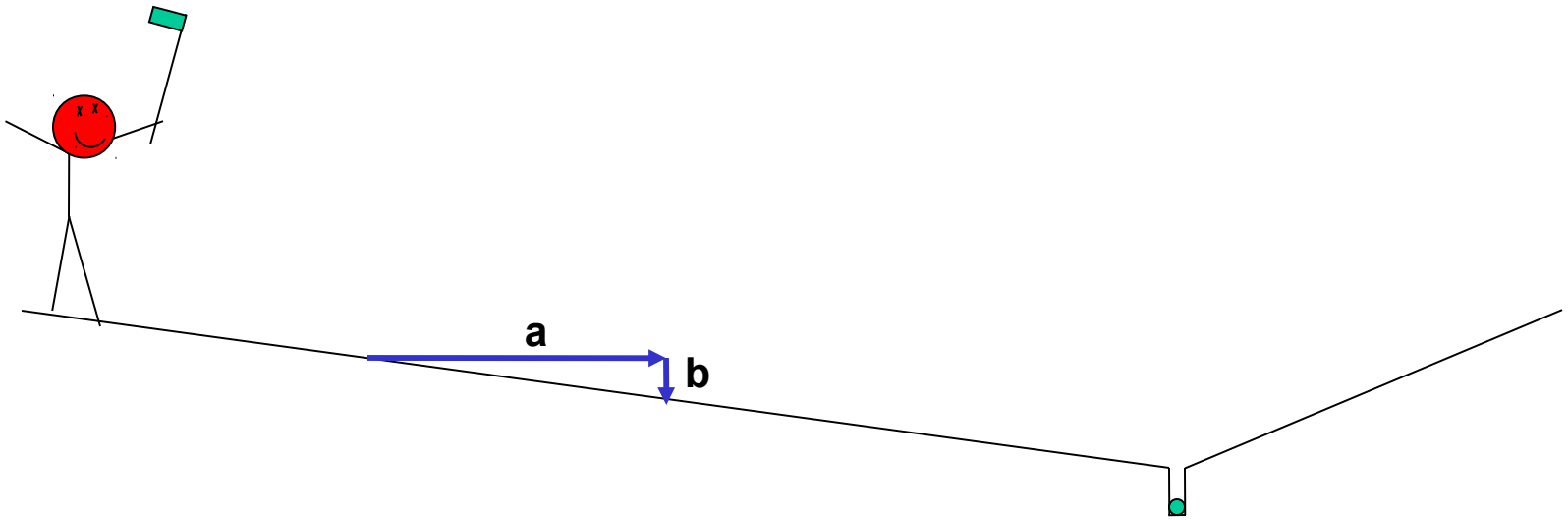
How do you hit a hole-in-one,
when you can't even see the hole ?

A Blind Golfer's view of global optimization: II



Change the shape of the golf course !

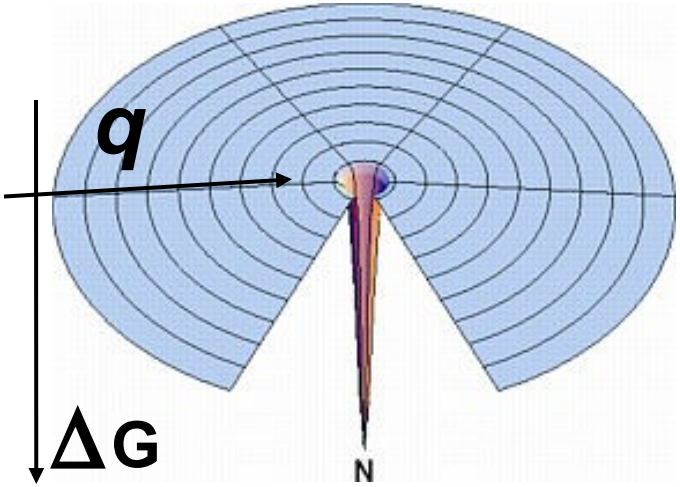
An analysis of why the Blind Golfer's strategy works



<u>Local</u> improvements in position (a) lead to incremental improvements in energy (b) !!!

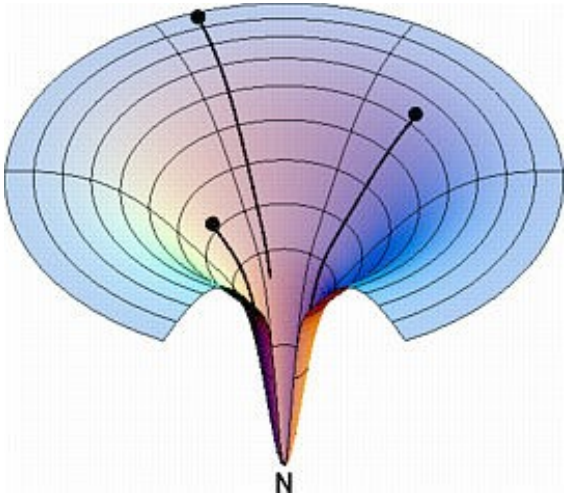
How does nature fold proteins ?

The funnel model reconciles the thermodynamic and the kinetic view !



In a flat folding landscape, a thermodynamic minimum is kinetically **inaccessible**.

But ...

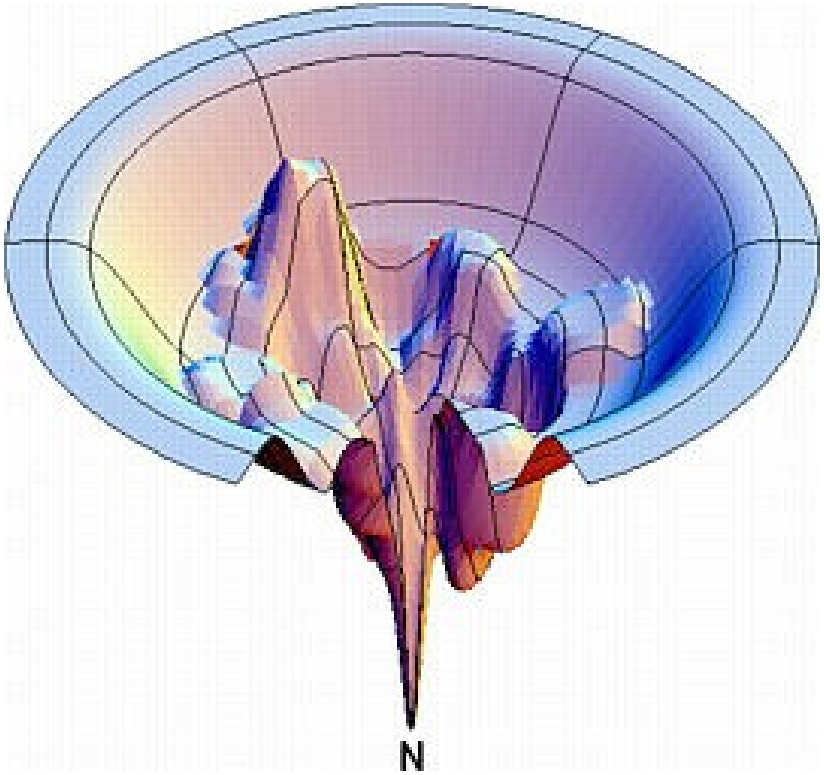


An ideal funnel results in fast, two-state folding through **many possible pathways**.

Dill KA & Chan HS (1997) From Levinthal to pathways to funnels. *Nature Struct Biol* **4**:10-19

How does nature fold proteins ?

Real folding landscapes appear to be more complex - robust folding is possible, but so are populated intermediate states and kinetic traps.



Dill KA & Chan HS (1997) From Levinthal to pathways to funnels. *Nature Struct Biol* **4**:10-19

Ongoing research

CASP: Critical Assessment of protein Structure Prediction

Competition held every 2 years since 1994.

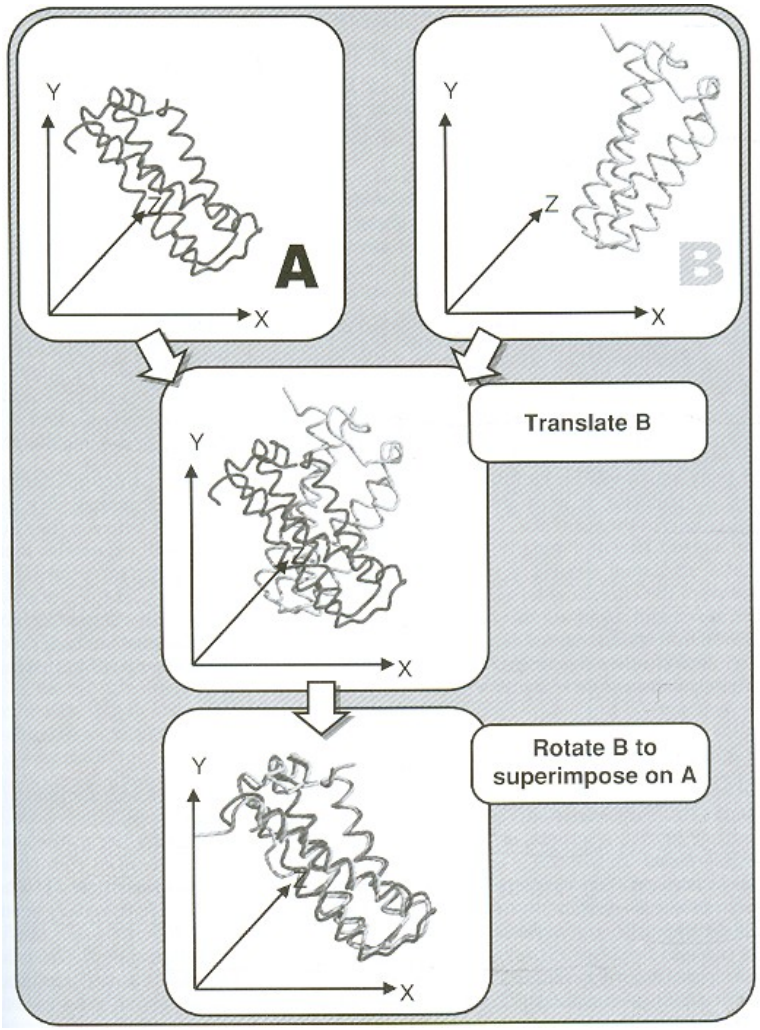
Protein sequence is given and closest to real structure wins.

Structure Similarity

Rigid body Superimposition

Principle

For two protein structures, there is a geometric transformation that, once applied to one structure, minimizes the overall distance between pairs of atoms.

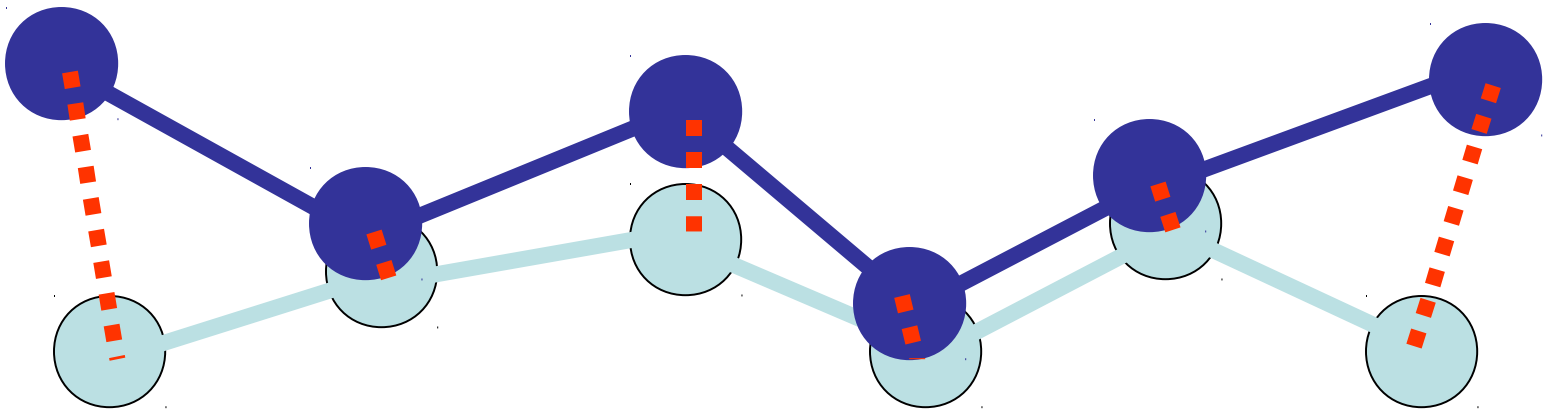


RMSD

Equation

An distance that is somehow normalized to the number of pairs of coordinates.

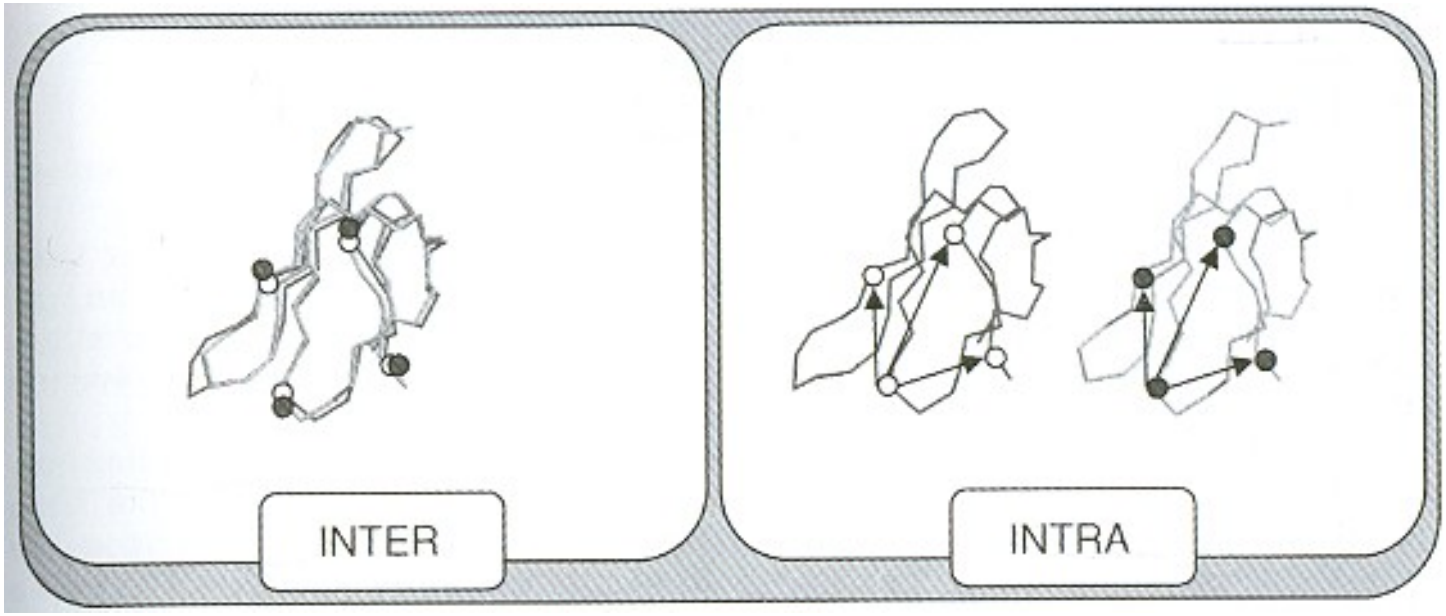
$$RMSD = \sqrt{\frac{\sum_{i}^{N} d_{i,i'}^{2}}{N}}$$



Problem with Rigid Body Superimposition

There is no way to know which sites to pair together in the first place, or even whether a given site should be paired.

Inter and intramolecular distances



Intermolecular comparisons : an absolute distance between two paired sites.

Intramolecular comparisons : a difference between two sets of relative descriptions of a site's context.

Structure Prediction Tools

CE & VAST – identify large groups of pairs of atoms.

DALI – identifies rigid similar segments

FATCAT – allows some regions to be flexible

Combinatorial Extension (CE)

Shindyalov and Bourne, 1998

Principle

Breaks the protein in 8-mers.

Find all good 8-mer pairs as potential starting points.

From each starting point, extend the alignment with the 8-mer which is the most consistent with the growing path.

Finally, find the overall path through the structure

Applications

Compare the structure of the protein in solution versus the same protein in the bound conformation to a ligand.

- Identify what changes and what stays the same (active sites).

Detection of distant evolutionary relationships.

Many protein families (related) have sequence similarity that is too low to reliably use sequence similarity measures.

Structural variability within a group of related structures.

This can tell what is important and what is unique to a group with an exclusive function.

Common structural motifs.