Protein structure alignments

Proteins that fold in the same way, i.e. ”have the same fold” are often homologs.
Structure evolves slower than sequence
Sequence is less conserved than structure

If BLAST gives no homologs (*i.e.* sequence based)

Instead: Search with protein *structure* (pdb-file) in *structure database* (e.g. PDB) to find more remote homologs

- For example using DALI
- Much more sensitive than sequence search
- Problems
  - Much smaller database (PDB vs. Genbank)
  - Need 3D structure of protein

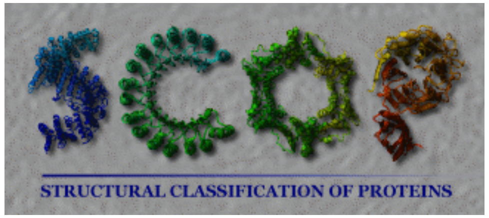
Use structure comparisons to classify, group and cluster proteins. Build protein structure families and hierarchies

CATH
PROTEIN STRUCTURE CLASSIFICATION

STRUCTURAL CLASSIFICATION OF PROTEINS

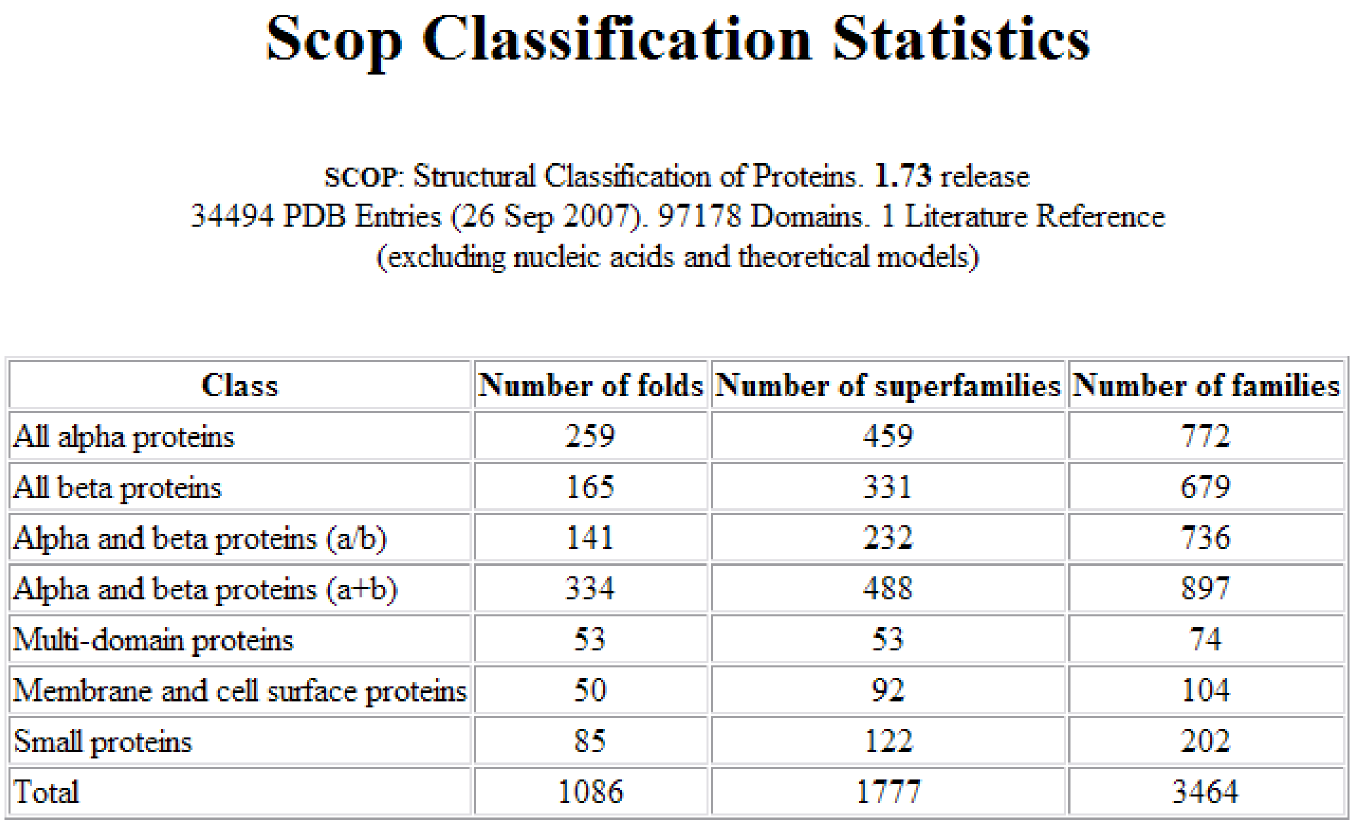
Protein structure classification

- Based on taking all structures of PDB
- Remove redundancy (*i.e.* keep only one copy of “identical” structures)
- Split structures into domains
- Group domains/proteins based on similarity
- Two main classification schemes: SCOP & CATH



Structural Classification of Proteins

- Almost 100% manually generated
- Proteins grouped into hierarchy of classes, folds, superfamilies and families

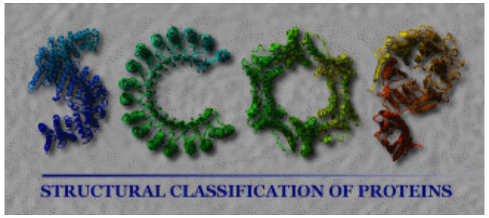


Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.73** release
34494 PDB Entries (26 Sep 2007). 97178 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

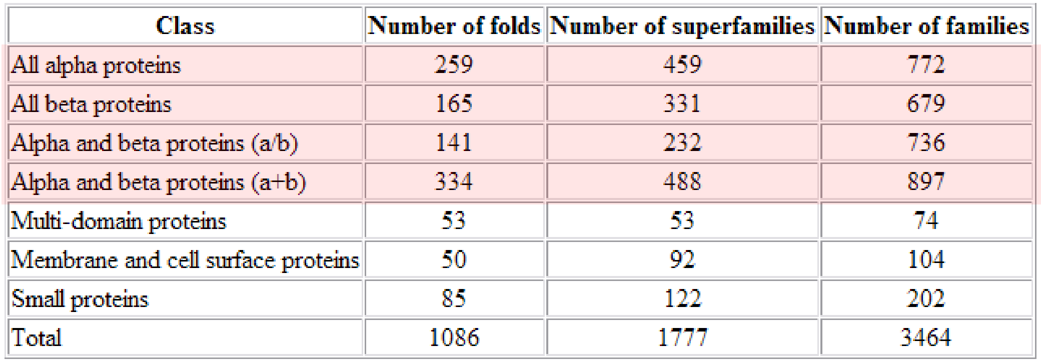
http://scop.mrc-lmb.cam.ac.uk/scop



SCOP

- Families
  - Sequence identity ~30% or higher
  - Very similar structures
  - Clearly homologous proteins
- Superfamilies
  - Contains families
  - May have no or little sequence similarity
  - Common fold
  - Are probably evolutionary related
- Folds
  - Contains superfamilies
  - Difficult level of classification
  - Same major secondary structure elements (α-helices and β-sheets) with same connections
  - Not always homologs

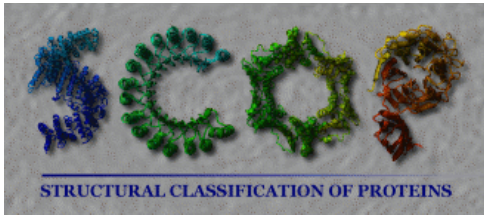
- Classes
  - Upper level of classification (4 major, 3 minor)
  - Contains folds
  - Based on secondary structure composition and “general features”
  - *e.g.* all-α, all-β, ”membrane and cell surface” and ”small proteins”
  - α/β: One β-sheet with strands connected by single α-helices
  - α+β: α-helical and β-sheet part separated in sequence

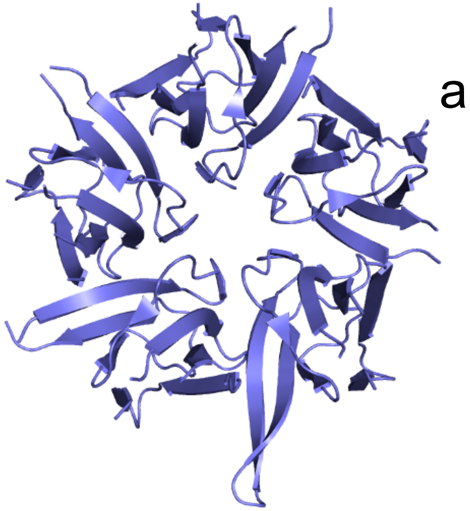


Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

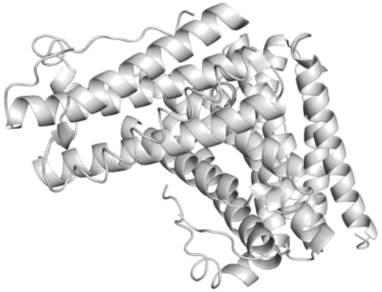
http://scop.mrc-lmb.cam.ac.uk/scop

SCOP

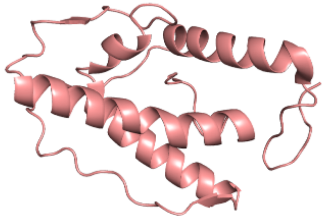




all-β class

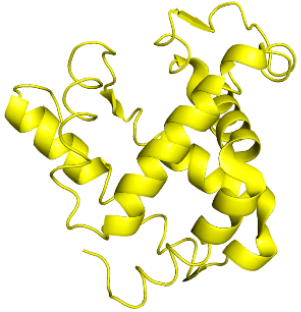


4-helical cytokines

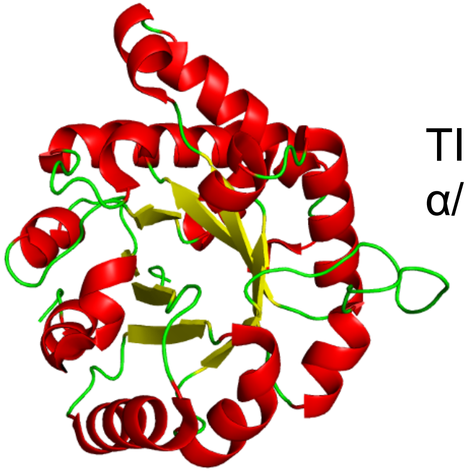


T4 endonuclease V

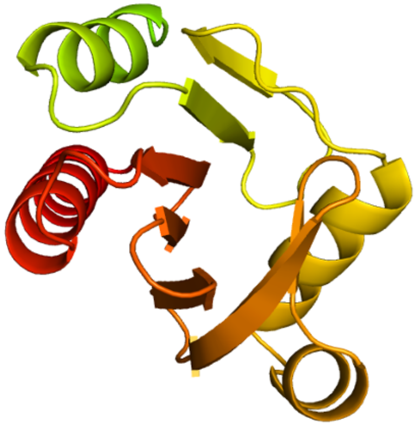
all-α class,
3 different folds



Globin-like



TIM-barrel fold
α/β class



Profilin-like fold
α+β class

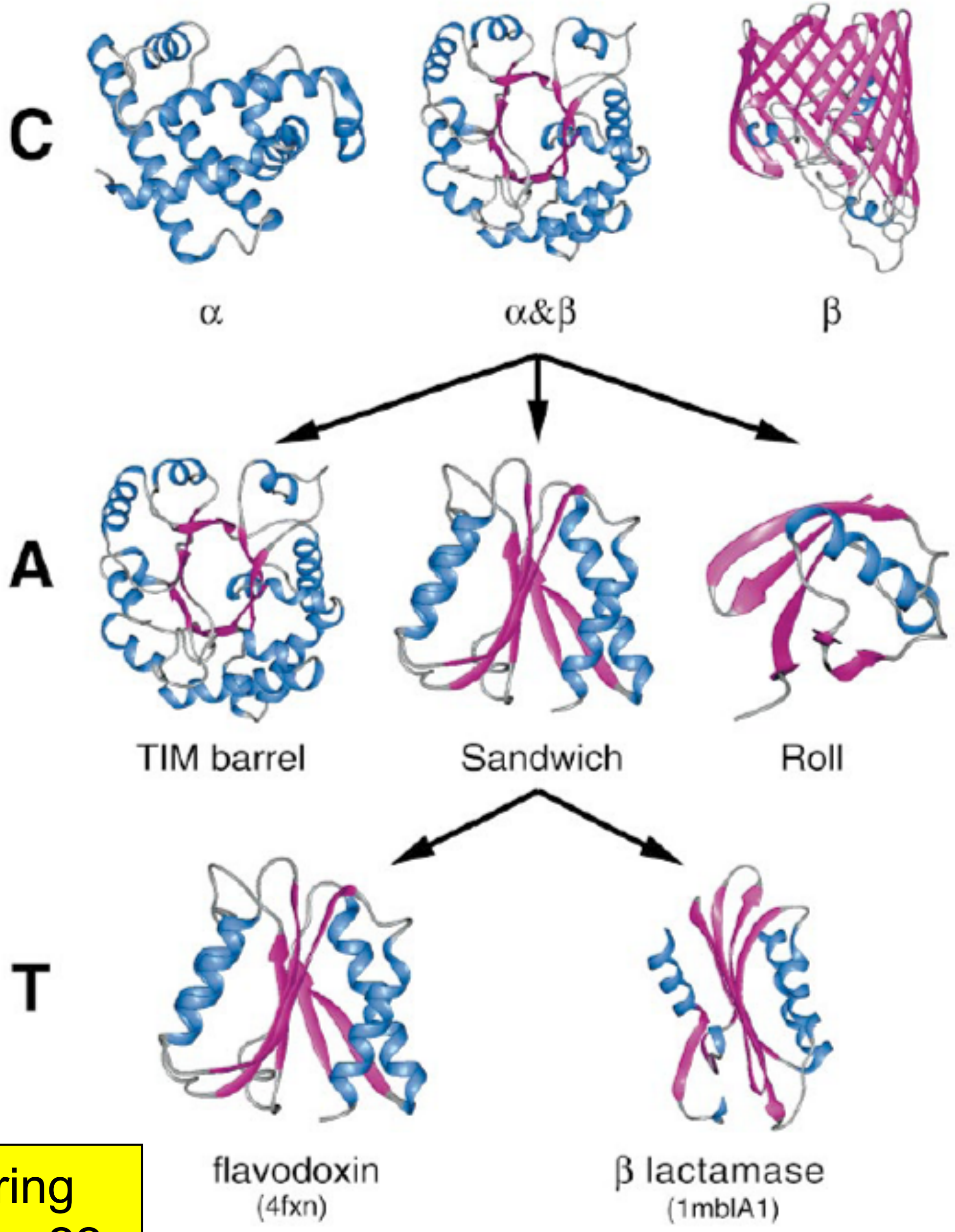
http://scop.mrc-lmb.cam.ac.uk/scop

CATH

Class, Architecture, Topology and Homologous

Both manual structural alignment and automatic alignment with SSAP

5 levels in hierarchy
- Class (as in SCOP)
- Architecture (unique to CATH)
- Fold/Topology (as in SCOP fold)
- Homologous Superfamily (as in SCOP)
- Homologous family (as in SCOP)



C.A. Orengo *et al.* *Structure* **5**, 1093 (1997)

Explore during the exercises??

http://www.cathdb.info

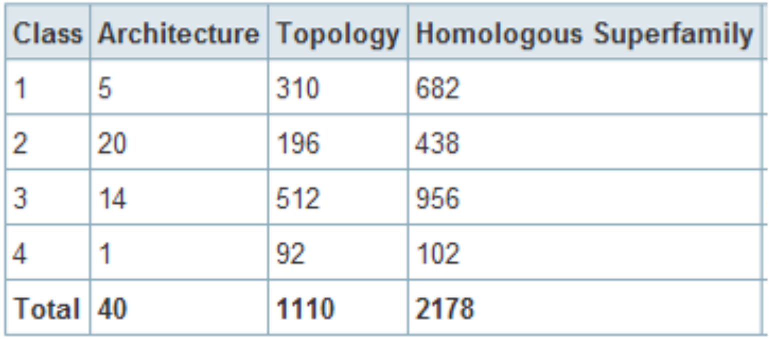
CATH
PROTEIN STRUCTURE CLASSIFICATION

CATH vs. SCOP

STRUCTURAL CLASSIFICATION OF PROTEINS

- Not always same domains
- Differences in hierarchy (5 vs. 4 levels)
- Differences in classes (4 vs. 7)
- Fully manual (SCOP) vs. manual/automatic (CATH)
- Most of the time (~80% of cases) classification is similar
- Both systems has weaknesses and strengths
- Use both!

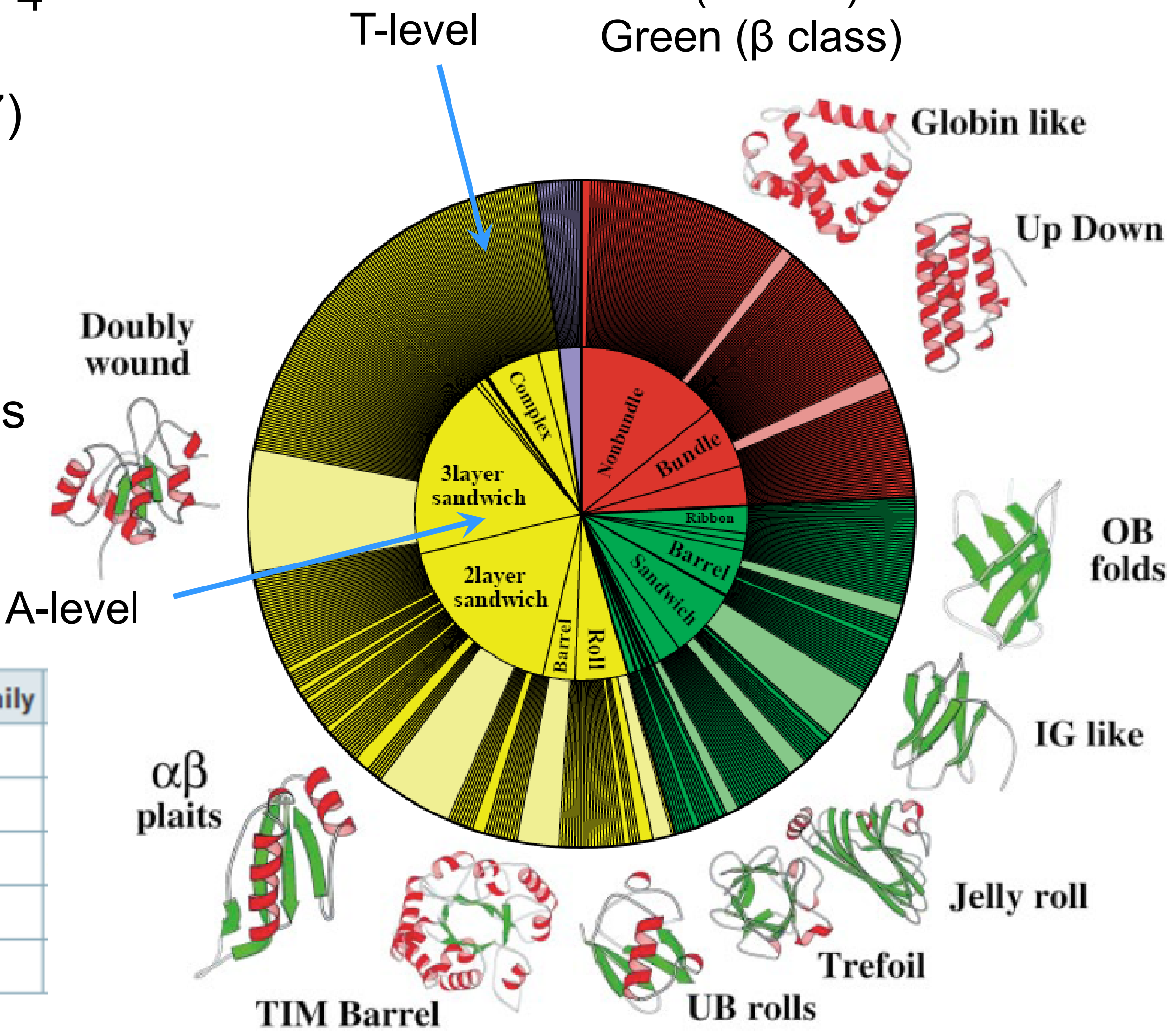
CATH Version 3.2



Class	Architecture	Topology	Homologous Superfamily
1	5	310	682
2	20	196	438
3	14	512	956
4	1	92	102
Total	40	1110	2178

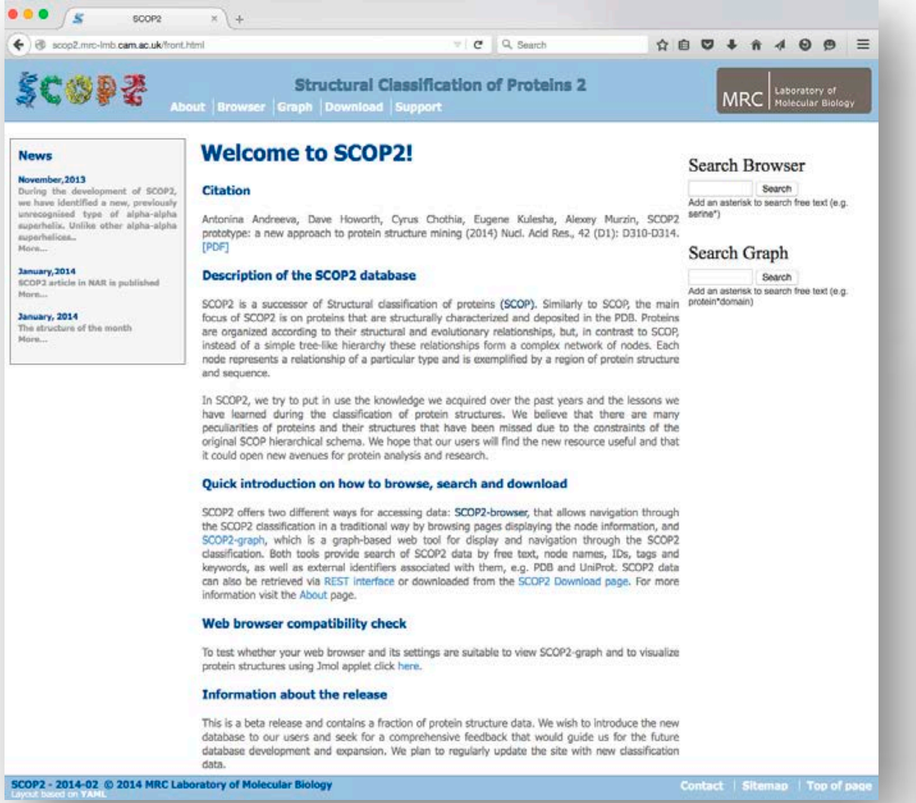
New topologies/folds are not found often!

Yellow (α-β class)
Red (α class)
Green (β class)



C.A. Orengo *et al. Structure* **5**, 1093 (1997)

SCOP2 & SCOPe


SCOP2
scop2.mrc-lmb.cam.ac.uk/front.html
Search
MRC Laboratory of Molecular Biology

Structural Classification of Proteins 2

About | Browser | Graph | Download | Support

News

November, 2013
During the development of SCOP2, we have identified a new, previously unrecognised type of alpha-alpha superhelix. Unlike other alpha-alpha superhelices..
More...

January, 2014
SCOP2 article in NAR is published
More...

January, 2014
The structure of the month
More...

Welcome to SCOP2!

Citation

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey Murzin, SCOP2 prototype: a new approach to protein structure mining (2014) Nucl. Acid Res., 42 (D1): D310-D314. [PDF]

Description of the SCOP2 database

SCOP2 is a successor of Structural classification of proteins (SCOP). Similarly to SCOP, the main focus of SCOP2 is on proteins that are structurally characterized and deposited in the PDB. Proteins are organized according to their structural and evolutionary relationships, but, in contrast to SCOP, instead of a simple tree-like hierarchy these relationships form a complex network of nodes. Each node represents a relationship of a particular type and is exemplified by a region of protein structure and sequence.

In SCOP2, we try to put in use the knowledge we acquired over the past years and the lessons we have learned during the classification of protein structures. We believe that there are many peculiarities of proteins and their structures that have been missed due to the constraints of the original SCOP hierarchical schema. We hope that our users will find the new resource useful and that it could open new avenues for protein analysis and research.

Quick introduction on how to browse, search and download

SCOP2 offers two different ways for accessing data: SCOP2-browser, that allows navigation through the SCOP2 classification in a traditional way by browsing pages displaying the node information, and SCOP2-graph, which is a graph-based web tool for display and navigation through the SCOP2 classification. Both tools provide search of SCOP2 data by free text, node names, IDs, tags and keywords, as well as external identifiers associated with them, e.g. PDB and UniProt. SCOP2 data can also be retrieved via REST interface or downloaded from the SCOP2 Download page. For more information visit the About page.

Web browser compatibility check

To test whether your web browser and its settings are suitable to view SCOP2-graph and to visualize protein structures using Jmol applet click here.

Information about the release

This is a beta release and contains a fraction of protein structure data. We wish to introduce the new database to our users and seek for a comprehensive feedback that would guide us for the future database development and expansion. We plan to regularly update the site with new classification data.

Search Browser

[ ] Search

Add an asterisk to search free text (e.g. serine*)

Search Graph

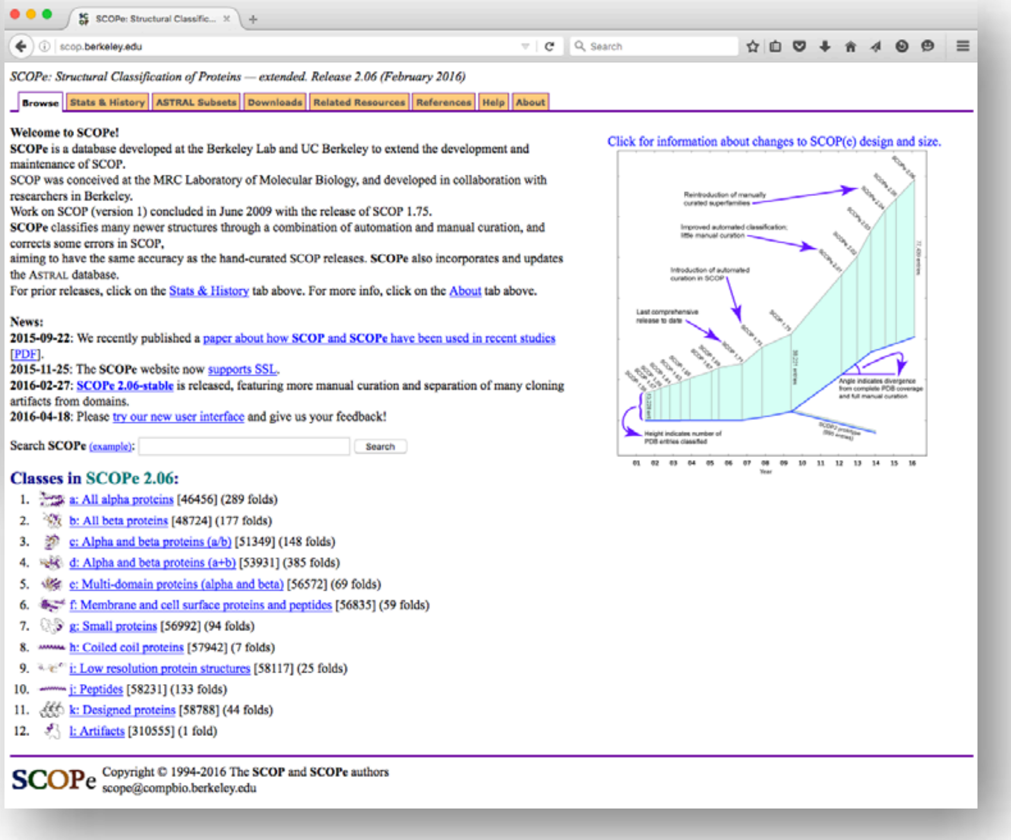
[ ] Search

Add an asterisk to search free text (e.g. protein*domain)

SCOP2 - 2014-02 © 2014 MRC Laboratory of Molecular Biology
Layout based on YAML

Contact | Sitemap | Top of page

Use this, most likely


SCOPe: Structural Classific...
scop.berkeley.edu
Search
Browse
Stats & History
ASTRAL Subsets
Downloads
Related Resources
References
Help
About
Reintroduction of manually curated superfamilies
Improved automated classification; little manual curation
Introduction of automated curation in SCOP
Last comprehensive release to date
Height indicates number of PDB entries classified
Angle indicates divergence from complete PDB coverage and full manual curation
SCOP2 prototype (995 entries)
SCOP 1.55
SCOP 1.57
SCOP 1.59
SCOP 1.61
SCOP 1.63
SCOP 1.65
SCOP 1.67
SCOP 1.69
SCOP 1.71
SCOP 1.73
SCOP 1.75
SCOPe 2.01
SCOPe 2.02
SCOPe 2.03
SCOPe 2.04
SCOPe 2.05
SCOPe 2.06
38,221 entries
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
Year

SCOPe: Structural Classification of Proteins — extended. Release 2.06 (February 2016)

Welcome to SCOPe!
SCOPe is a database developed at the Berkeley Lab and UC Berkeley to extend the development and maintenance of SCOP.
SCOP was conceived at the MRC Laboratory of Molecular Biology, and developed in collaboration with researchers in Berkeley.
Work on SCOP (version 1) concluded in June 2009 with the release of SCOP 1.75.
SCOPe classifies many newer structures through a combination of automation and manual curation, and corrects some errors in SCOP,
aiming to have the same accuracy as the hand-curated SCOP releases. **SCOPe** also incorporates and updates the ASTRAL database.
For prior releases, click on the Stats & History tab above. For more info, click on the About tab above.

Click for information about changes to SCOP(e) design and size.

News:
2015-09-22: We recently published a paper about how **SCOP** and **SCOPe** have been used in recent studies [PDF].
2015-11-25: The **SCOPe** website now supports SSL.
2016-02-27: **SCOPe 2.06-stable** is released, featuring more manual curation and separation of many cloning artifacts from domains.
2016-04-18: Please try our new user interface and give us your feedback!

Search **SCOPe** (example): [ ] Search

Classes in SCOPe 2.06:

1. a: All alpha proteins [46456] (289 folds)
2. b: All beta proteins [48724] (177 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (385 folds)
5. e: Multi-domain proteins (alpha and beta) [56572] (69 folds)
6. f: Membrane and cell surface proteins and peptides [56835] (59 folds)
7. g: Small proteins [56992] (94 folds)
8. h: Coiled coil proteins [57942] (7 folds)
9. i: Low resolution protein structures [58117] (25 folds)
10. j: Peptides [58231] (133 folds)
11. k: Designed proteins [58788] (44 folds)
12. l: Artifacts [310555] (1 fold)

SCOPe Copyright © 1994-2016 The **SCOP** and **SCOPe** authors
scope@compbio.berkeley.edu

http://scop.berkeley.edu

Predictors

Prediction tools

- Predictors are available
  - on the web (in public web servers)
  - as (usually) free or commercial software
  - packaged in large (often commercial) software suites
- Predictors have been made for determining all kinds of features from sequence
  - Secondary structure
  - Structural disorder
  - Domain boundaries
  - Membrane protein or not
  - Number of transmembrane α-helices
  - Metal ion binding sites
  - Post-translational modifications
    - Phosphorylation sites
    - Cleavage sites
  - And many more
- Subcellular localization
  - Nuclear protein?
  - Secreted protein?
- Interaction with other proteins, DNA etc. (usually with some knowledge of 3D structure)

These tools are often extremely useful to biologists!

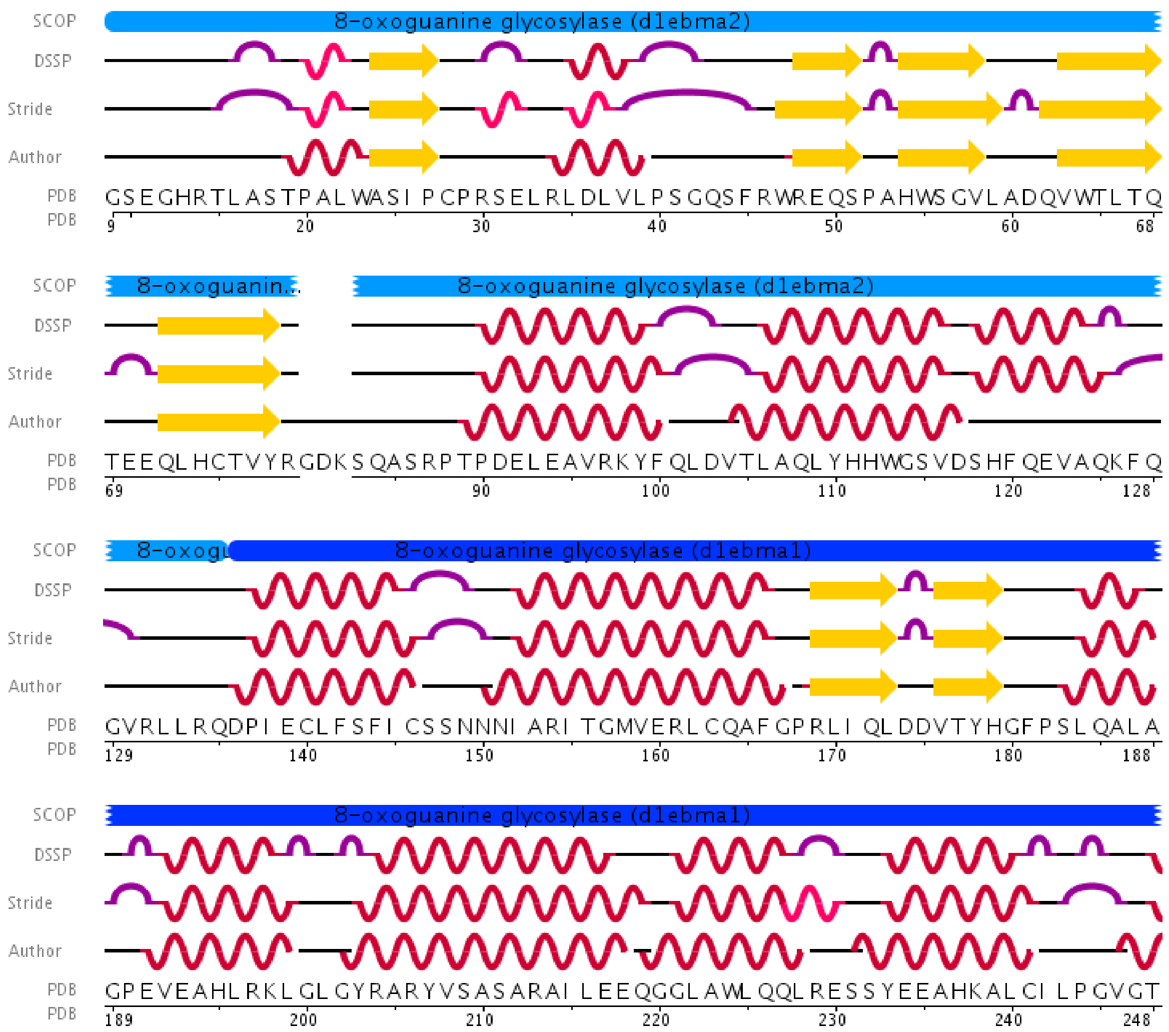
Example here is *secondary structure prediction* but similar or related methods/algorithms are used in most predictors

Secondary structure prediction

Assigning secondary structure is *not trivial* and there is *no single consensus method* even when 3D structure is known

- Secondary structure may be put in manually by the authors behind a PDB-file
- Algorithms based on calculated H-bonds, Ramachandran plot, etc.
    - DSSP
    - STRIDE
    - DEFINE

1EBM


SCOP
8-oxoguanine glycosylase (d1ebma2)
DSSP
Stride
Author
PDB
GSEGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQ
9 20 30 40 50 60 68
8-oxoguanin...
8-oxoguanine glycosylase (d1ebma2)
TEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQ
69 90 100 110 120 128
8-oxogu
8-oxoguanine glycosylase (d1ebma1)
GVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALA
129 140 150 160 170 180 188
8-oxoguanine glycosylase (d1ebma1)
GPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGT
189 200 210 220 230 240 248

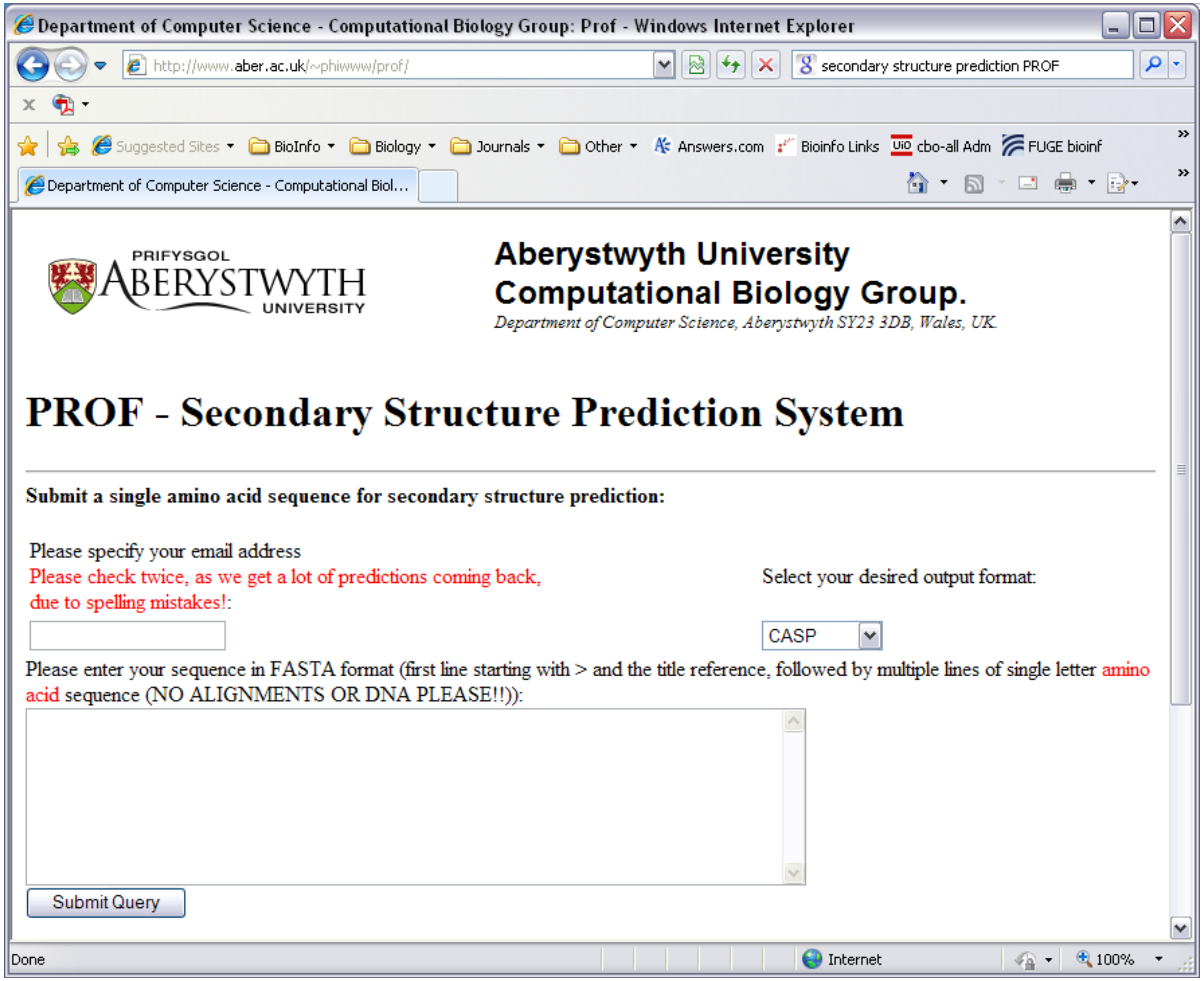
➡ β-strand

〰 α-helix

Everything else loop/coil

Secondary structure prediction

Tools/programs that accept a primary sequence and predicts the secondary structure state (H/helix, E/sheet, or C/Loop&Coil) for each residue


Department of Computer Science - Computational Biology Group: Prof - Windows Internet Explorer
http://www.aber.ac.uk/~phiwww/prof/
secondary structure prediction PROF
Suggested Sites
BioInfo
Biology
Journals
Other
Answers.com
Bioinfo Links
cbo-all Adm
FUGE bioinf
Department of Computer Science - Computational Biol...
PRIFYSGOL ABERYSTWYTH UNIVERSITY
Done
Internet
100%

Aberystwyth University
Computational Biology Group.
Department of Computer Science, Aberystwyth SY23 3DB, Wales, UK

PROF - Secondary Structure Prediction System

Submit a single amino acid sequence for secondary structure prediction:

Please specify your email address
Please check twice, as we get a lot of predictions coming back, due to spelling mistakes!:

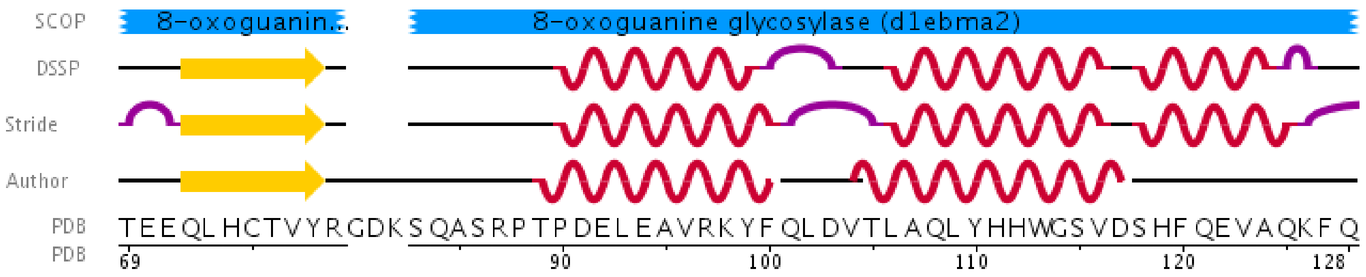
Select your desired output format: CASP

Please enter your sequence in FASTA format (first line starting with > and the title reference, followed by multiple lines of single letter amino acid sequence (NO ALIGNMENTS OR DNA PLEASE!!)):

Submit Query

Secondary structure prediction

Tools/programs that accept a primary sequence and predicts the secondary structure state (H/helix, E/sheet, or C/Loop&Coil) for each residue


SCOP
DSSP
Stride
Author
PDB
PDB
8-oxoguanin...
8-oxoguanine glycosylase (d1ebma2)
TEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQ
69
90
100
110
120
128

```
Human OGG1        TEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQ
PROF Prediction   CEEEEEEEEECCCCCCCCCCHHHHHHHHHHHCCCCHHHHHHCCCCCCHHHHHHHHHHHCC
```

Uses:

- Correct and guide sequence alignments since secondary structure is more conserved than primary sequence
- Classify proteins
  - If you think your protein is a TIM-barrel, but your prediction suggests it has only α-helices, you probably are wrong
- Important step towards predicting 3D structure

Globular and transmembrane proteins have quite different properties and should be tackled with different algorithms

Secondary structure prediction

- Random prediction ~40% accuracy
- 1st generation prediction (1970’s) ~50%
    - Based on relative *propensities*/intrinsic tendencies of each amino acid to be in a state X (= H, E, or C)
    - Ala, Glu & Met often in state H
    - Pro & Gly often in state C

- 2nd generation prediction (until mid 1990’s) ~60%
    - Proper inclusion of propensities for neighboring residues
    - Larger experimental data set

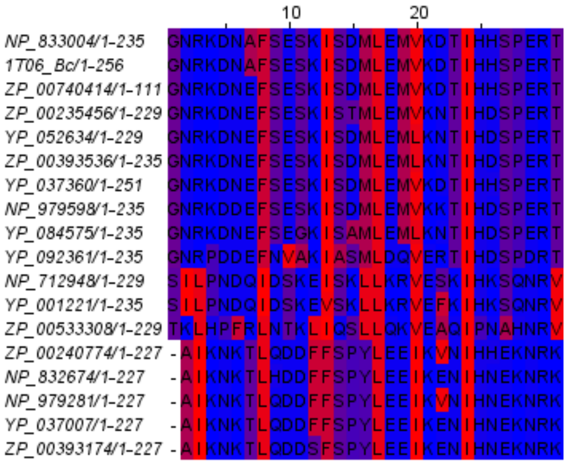
- 3rd generation prediction (until present time) approaching ~80%
- Two main improvements:
    - Machine learning/neural networks
    - Combines information from predictions for single sequence with information from homologous sequences (multiple sequence alignment)

    Since structure is more conserved than sequence homologs (>35% identity) are likely to have same secondary structure

Secondary structure prediction

Jon K. Lærdahl,
Structural Bioinformatics

- 3rd generation prediction (until present time) approaching ~80%
- Two main improvements:
  - Machine learning/neural networks
  - Combines information from predictions for single sequence with information from homologous sequences (For example sequences with >35% identity in multiple sequence alignment)



```
                                 10        20
NP_833004/1-235     GNRKDNAFSESKISDMLEMVKDTIHHSPERT
1T06_Bc/1-256       GNRKDNAFSESKISDMLEMVKDTIHHSPERT
ZP_00740414/1-111   GNRKDNEFSESKISDMLEMVKDTIHHSPERT
ZP_00235456/1-229   GNRKDNEFSESKISTMLEMVKNTIHDSPERT
YP_052634/1-229     GNRKDNEFSESKISDMLEMLKNTIHDSPERT
ZP_00393536/1-235   GNRKDNEFSESKISDMLEMLKNTIHDSPERT
YP_037360/1-251     GNRKDNEFSESKISDMLEMVKDTIHHSPERT
NP_979598/1-235     GNRKDDEFSESKISDMLEMVKKTIHDSPERT
YP_084575/1-235     GNRKDNEFSEGKISAMLEMLKNTIHDSPERT
YP_092361/1-235     GNRPDDEFNVAKIASMLDQVERTIHDSPDRT
NP_712948/1-229     SILPNDQIDSKEISKLLKRVESKIHKSQNRV
YP_001221/1-235     SILPNDQIDSKEVSKLLKRVEFKIHKSQNRV
ZP_00533308/1-229   TKLHPFRLNTKLIQSLLQKVEAQIPNAHNRV
ZP_00240774/1-227   -AIKNKTLQDDFFSPYLEEIKVNIHHEKNRK
NP_832674/1-227     -AIKNKTLHDDFFSPYLEEIKENIHNEKNRK
NP_979281/1-227     -AIKNKTLQDDFFSPYLEEIKVNIHNEKNRK
YP_037007/1-227     -AIKNKTLQDDFFSPYLEEIKENIHNEKNRK
ZP_00393174/1-227   -AIKNKTLQDDSFSPYLEEIKENIHNEKNRK
```

Predict secondary structure for all these and fit onto alignment

Generate prediction based on consensus

Structure is more conserved than sequence!
More sequences available than structures (PDB vs GenBank)!

Sequences
& *known* secondary structures
from PDB

Neural network is trained on these data

Sequences

Trained neural network

Predicted secondary structures

Secondary structure prediction – consensus-based

- Random prediction ~40% accuracy
- 1st generation prediction (1970's) ~50%
- 2nd generation prediction (until mid 1990's) ~60%
- 3rd generation prediction (until present time) approaching ~80%

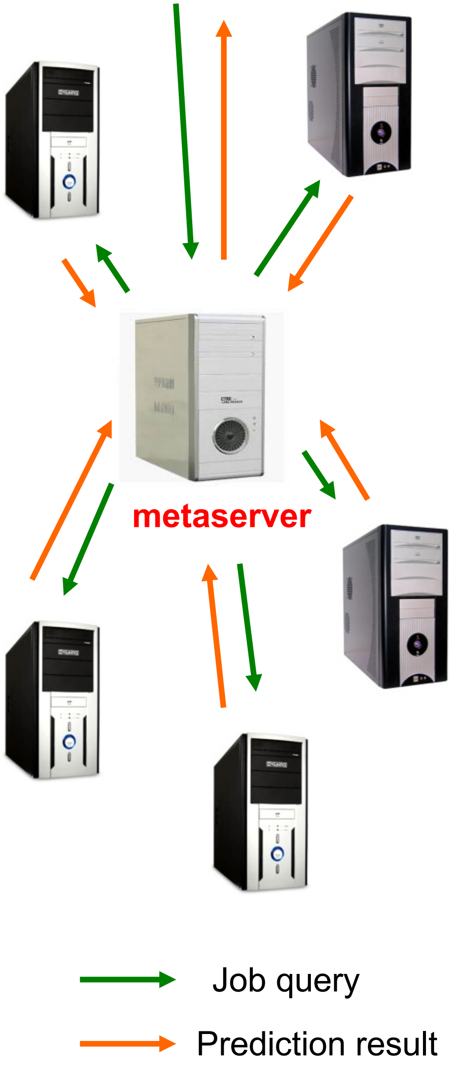
Many (more than 70 different published algorithms!) programs for secondary structure prediction:

- **PHD** – BLASTP to find homologs, MSA of homologs, neural networks used for prediction, web server
- **PSIPRED** – PSI-BLAST for homologs, MSA generated, neural network prediction, filtering, web server
- **PROF** – PSI-BLAST, MSA, neural network

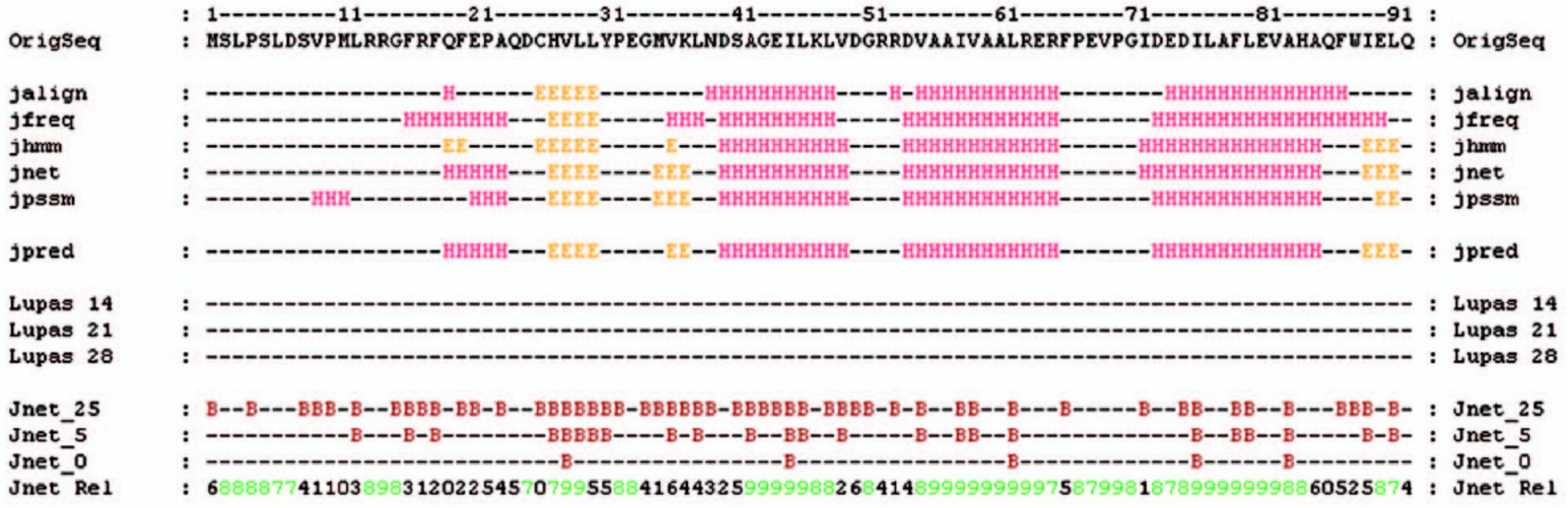
Very good idea to use *not one tool* and trust the results, but instead *use several unrelated tools* and compare/use the consensus

Some web servers do this automatically and generates a consensus based on several algorithms (*e.g.* Jpred & PredictProtein)

- Several programs run and the results are presented to the user as
  - one consensus result
  - all results and the interpretation is left to the user
- The individual programs may be
  - run locally
  - on web servers other places on the internet with the results collected and combined on the consensus-server (metaserver)

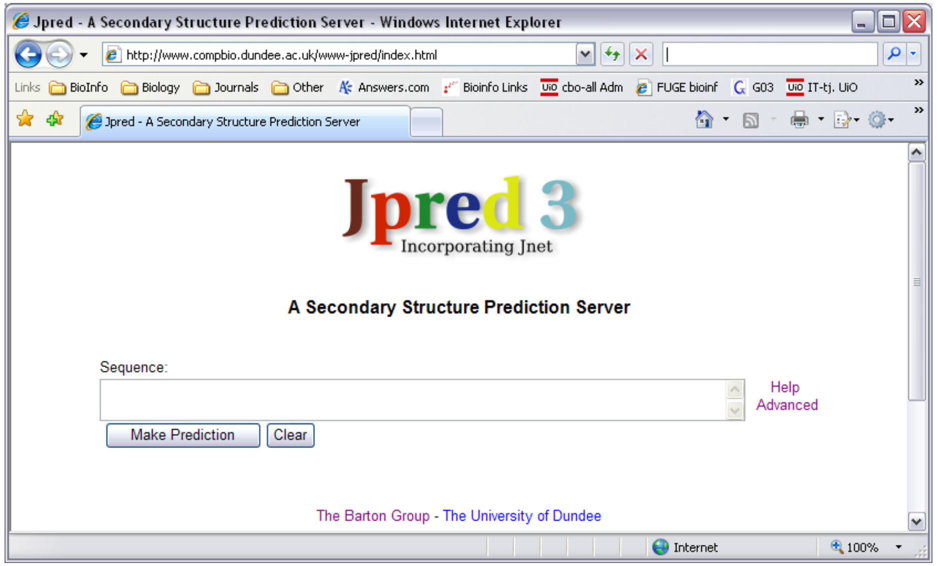


Secondary structure prediction – consensus-based



```
             : 1---------11--------21--------31--------41--------51--------61--------71--------81--------91 :
OrigSeq      : MSLPSLDSVPMLRRGFRFQFEPAQDCHVLLYPEGMVKLNDSAGEILKLVDGRRDVAAIVAALRERFPEVPGIDEDILAFLEVAHAQFWIELQ : OrigSeq

jalign       : ------------------H------EEEEE--------HHHHHHHHHH----H-HHHHHHHHHHH--------HHHHHHHHHHHHHH----- : jalign
jfreq        : ---------------HHHHHHHH---EEEE-----HHH-HHHHHHHHH-----HHHHHHHHHHHH-------HHHHHHHHHHHHHHHHHH-- : jfreq
jhmm         : ------------------EE-----EEEEE-----E---HHHHHHHHHH----HHHHHHHHHHHH------HHHHHHHHHHHHHH---EEE- : jhmm
jnet         : ------------------HHHHH---EEEE----EEE--HHHHHHHHHH----HHHHHHHHHHHH------HHHHHHHHHHHHHH---EEE- : jnet
jpssm        : --------HHH---------HHH---EEEE----EEE--HHHHHHHHHH----HHHHHHHHHHHH-------HHHHHHHHHHHHH----EE- : jpssm

jpred        : ------------------HHHHH---EEEE-----EE--HHHHHHHHHH----HHHHHHHHHHHH-------HHHHHHHHHHHHH---EEE- : jpred

Lupas 14     : -------------------------------------------------------------------------------------------- : Lupas 14
Lupas 21     : -------------------------------------------------------------------------------------------- : Lupas 21
Lupas 28     : -------------------------------------------------------------------------------------------- : Lupas 28

Jnet_25      : B--B---BBB-B--BBBB-BB-B--BBBBBBB-BBBBBB-BBBBBB-BBBB-B-B--BB--B---B-----B--BB--BB--B---BBB-B- : Jnet_25
Jnet_5       : -----------B---B-B--------BBBBB----B-B---B--BB--B-----B--BB--B-------------B--BB--B-----B-B- : Jnet_5
Jnet_0       : ---------------------------B----------------B----------------B-------------B------B--------- : Jnet_0
Jnet Rel     : 68888774110389831202254570799558841644325999998826841489999999997587998187899999998860525874 : Jnet Rel
```

Puehringer *et al. BMC Biochemistry* **9**:8 (2008)

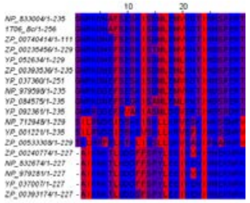

Jpred - A Secondary Structure Prediction Server - Windows Internet Explorer
http://www.compbio.dundee.ac.uk/www-jpred/index.html
Links
BioInfo
Biology
Journals
Other
Answers.com
Bioinfo Links
cbo-all Adm
FUGE bioinf
G03
IT-tj. UiO
Jpred - A Secondary Structure Prediction Server
Jpred 3
Incorporating Jnet
A Secondary Structure Prediction Server
Sequence:
Help
Advanced
Make Prediction
Clear
The Barton Group - The University of Dundee
Internet
100%

C. Cole *et al. Nucleic Acids Res.* ***36***, W197 (2008)

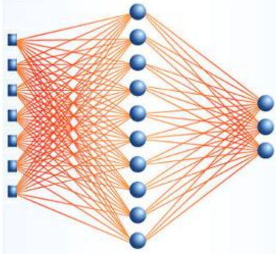
Predictors – *common features*

- Use propensities/intrinsic tendencies of single residues or short sequence segments to be in a certain state (*e.g.* secondary structure state, order/disorder state, signal sequence)

- Include local interactions, *i.e.* take into account states in up- and downstream sequence

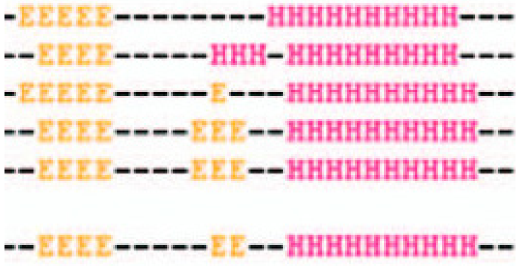


- Use homologous sequences to get predictions from many sequences with same structure/function



- Use neural networks or similar methods in predictions

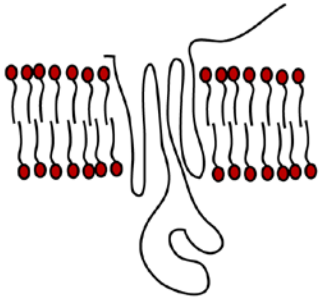
- Consensus from many tools is better than just a single result (e.g. metaservers)



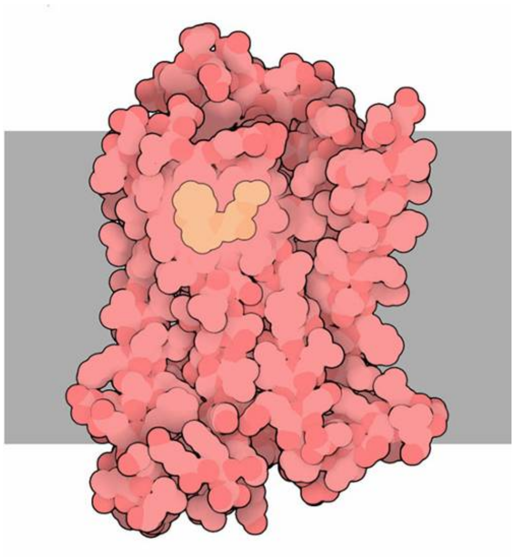
```
-EEEEE--------HHHHHHHHHH---
--EEEE-----HHH-HHHHHHHHH---
-EEEEE-----E---HHHHHHHHHH--
--EEEE----EEE--HHHHHHHHHH--
--EEEE----EEE--HHHHHHHHHH--

--EEEE-----EE--HHHHHHHHHH--
```

Transmembrane (TM) proteins

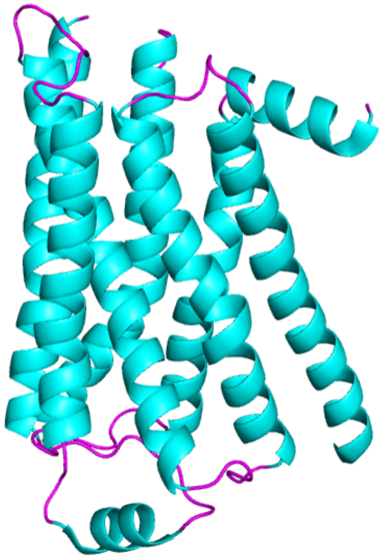
- ~30% of proteins in cells (but more than 50% of proteins interacts with membranes)
    - α-helical type: all membranes and organisms
    - β-barrel type: only outer membranes of Gram-negative bacteria, lipid-rich cell walls of a few Gram-positive bacteria, and outer membranes of mitochondria and chloroplasts



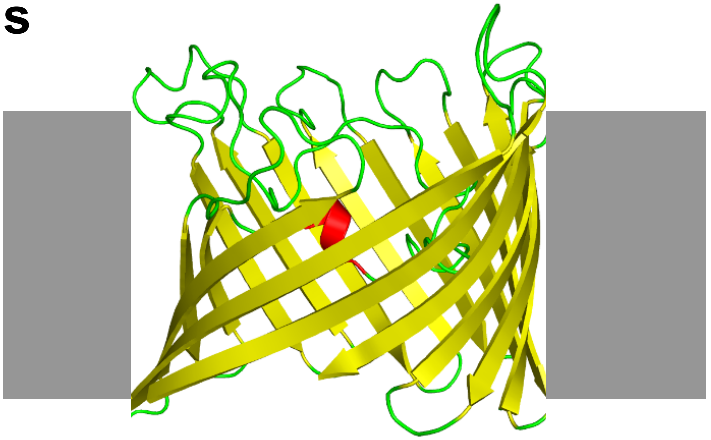
Can usually NOT use the same predictors for secondary structure and other properties as for globular proteins



PDB Apr. 08 “Molecule of the Month”



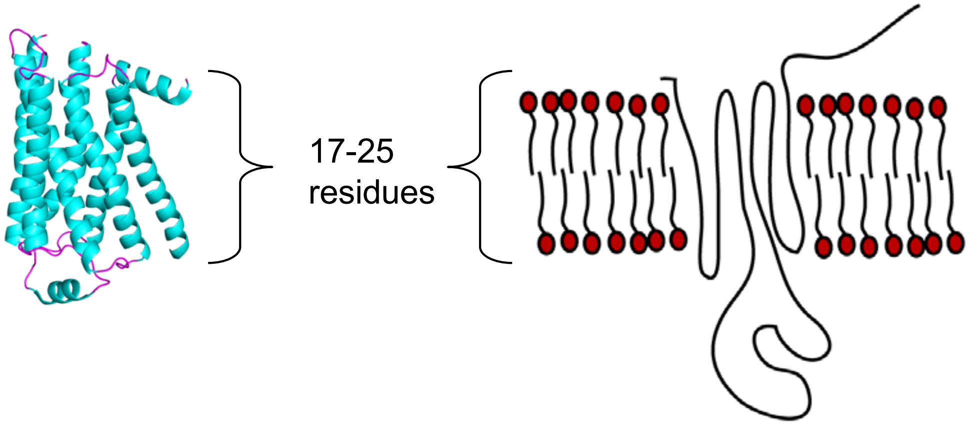
2RH1, Human adrenergic receptor



Porin

Transmembrane (TM) proteins

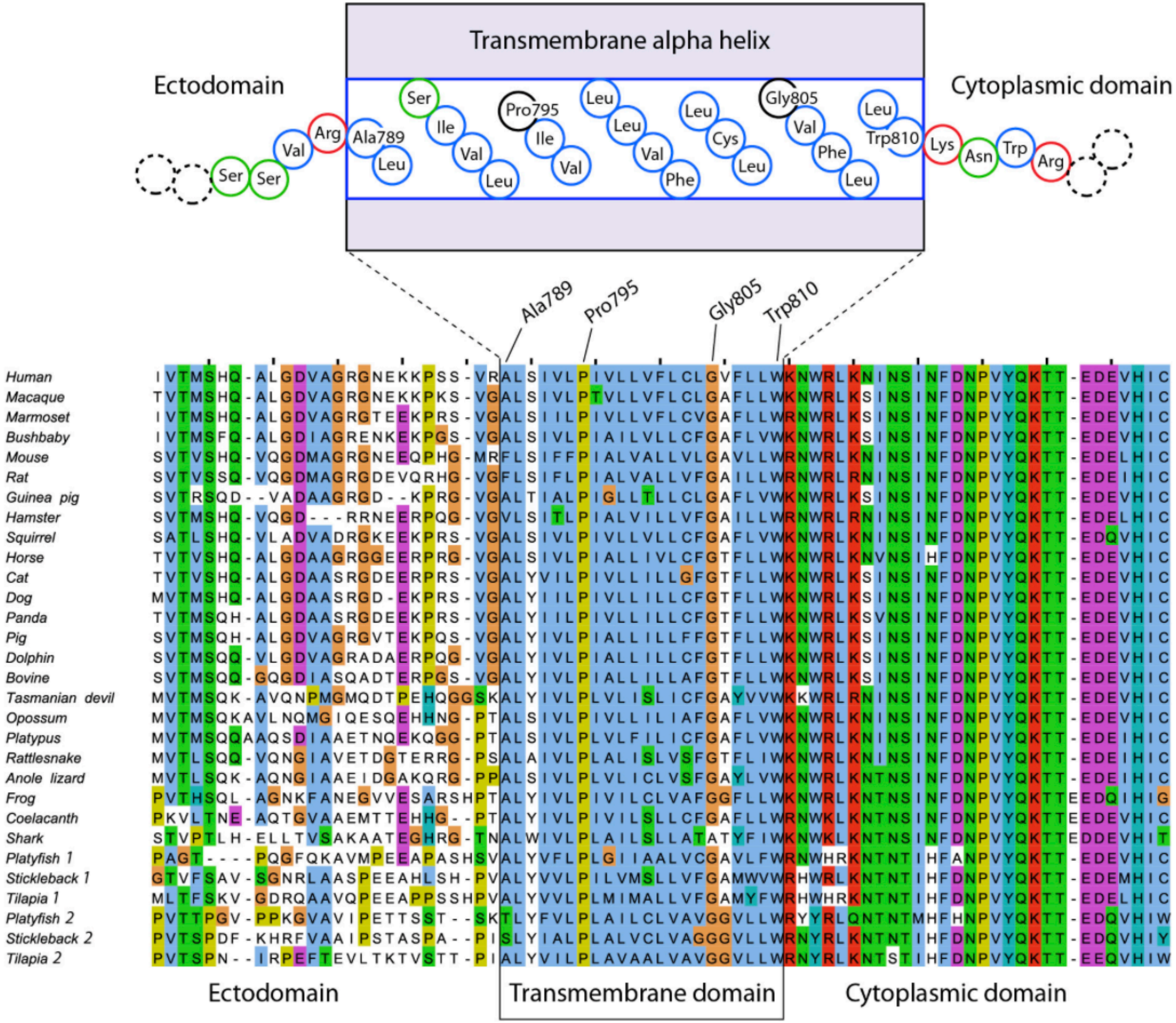
- Extremely difficult to solve membrane structures experimentally!
  - Only a few hundred structures in the PDB
- Can not use the same predictors for secondary structure as for globular proteins
- Special predictors for
  - helical membrane proteins
  - β-barrel proteins
- Pattern in TM α-helical proteins is:
  - 17-25 mainly hydrophobic TM helices
  - <60 residues polar connectors
- Predictions based on scanning for segments with high score for hydrophobicity
- Improved with neural networks



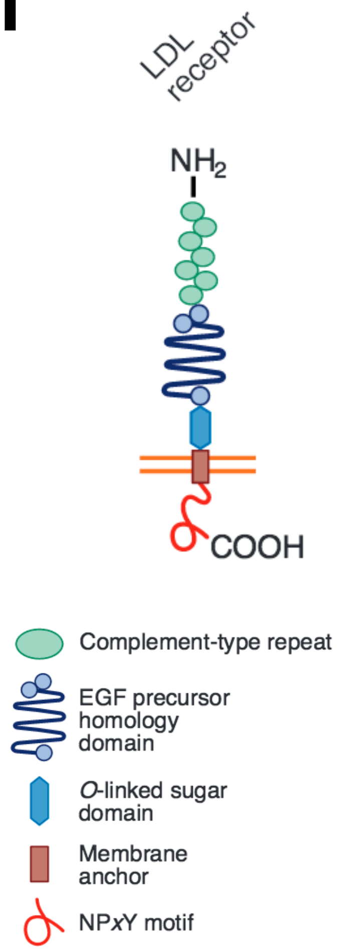
Tools:

- TMHMM
- Phobius

TM single-pass protein



```
Human            IVTMSHQ-ALGDVAGRGNEKKPSS-VRALSIVLPIVLLVFLCLGVFLLWKNWRLKNINSINFDNPVYQKTT-EDEVHIC
Macaque          TVTMSHQ-ALGDVAGRGNEKKPKS-VGALSIVLPTVLLVFLCLGAFLLWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Marmoset         IVTMSHQ-ALGDVAGRGTEEKPRS-VGALSIILPIVLLVFLCVGAFLLWRNWRLKSINSINFDNPVYQKTT-EDEVHIC
Bushbaby         IVTMSFQ-ALGDIAGRENKEKPGS-VGALSIVLPIAILVLLCFGAFLVWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Mouse            SVTVSHQ-VQGDMAGRGNEEQPHG-MRFLSIFFPIALVALLVLGAVLLWRNWRLKNINSINFDNPVYQKTT-EDELHIC
Rat              SVTVSSQ-VQGDMAGRGDEVQRHG-VGFLSIFLPIALVALLVFGAILLWRNWRLRNINSINFDNPVYQKTT-EDEIHIC
Guinea pig       SVTRSQD--VADAAGRGD--KPRG-VGALTIALPIGLLTLLCLGAFLVWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Hamster          SVTMSHQ-VQGD---RRNEERPQG-VGVLSITLPIALVILLVFGAILLWRNWRLRNINSINFDNPVYQKTT-EDELHIC
Squirrel         SATLSHQ-VLADVADRGKEEKPRS-VGALSIVLPIVLLVLLCFGAFLVWKNWRLKNINSINFDNPVYQKTT-EDQVHIC
Horse            TVTVSHQ-ALGDAAGRGGEERPRG-VGALSIVLPIALLIVLCFGTFLLWKNWRLKNVNSIHFDNPVYQKTT-EDEVHIC
Cat              TVTVSHQ-ALGDAASRGDEERPRS-VGALYVILPIVLLILLGFGTFLLWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Dog              MVTMSHQ-ALGDAASRGD-EKPRS-VGALYIILPIVLLILLCFGTFLLWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Panda            TVTMSQH-ALGDAASRGDEEKPRS-VGALYIILPIVLLILLCFGTFLLWKNWRLKSVNSINFDNPVYQKTT-EDEVHIC
Pig              SVTMSQH-ALGDVAGRGVTEKPQS-VGALYIVLPIALLILLFFGTFLLWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Dolphin          SVTMSQQ-VLGDVAGRADAERPQG-VGALYIVLPIALLILLCFGTFLLWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Bovine           SVTMSQQ-GQGDIASQADTERPGS-VGALYIVLPIALLILLAFGTFLLWKNWRLKSINSINFDNPVYQKTT-EDEVHIC
Tasmanian devil  MVTMSQK-AVQNPMGMQDTPEHQGGSKALYIVLPLVLISLICFGAYVVWKKWRLRNINSINFDNPVYQKTT-EDEVHIC
Opossum          MVTMSQKAVLNQMGIQESQEHHNG-PTALSIVLPIVLLILIAFGAFLVWKNWRLKNINSINFDNPVYQKTT-EDEVHIC
Platypus         MVTMSQQAAQSDIAAETNQEKQGG-PTALSIVLPLVLFILICFGAFLVWKNWRLKNINSINFDNPVYQKTT-EDEIHIC
Rattlesnake      MVTLSQQ-VQNGIAVETDGTERRG-PSALAIVLPLALISLVSFGTFLIWKNWRLKNINSINFDNPVYQKTT-EDEIHIC
Anole lizard     MVTLSQK-AQNGIAAEIDGAKQRG-PPALSIVLPLVLICLVSFGAYLVWKNWRLKNTNSINFDNPVYQKTT-EDEIHIC
Frog             PVTHSQL-AGNKFANEGVVESARSHPTALYIVLPIVILCLVAFGGFLLWKNWRLKNTNSINFDNPVYQKTTEEDQIHIG
Coelacanth       PKVLTNE-AQTGVAAEMTTEHHG--PTALYIVLPIVILSLLCFGAFLLWRNWKLKNTNSINFDNPVYQKTTEEDEVHIC
Shark            STVPTLH-ELLTVSAKAATEGHRG-TNALWIVLPLAILSLLATATYFIWKNWKLKNTNSINFDNPVYQKTTEDDEVHIT
Platyfish 1      PAGT----PQGFQKAVMPEEAPASHSVALYVFLPLGIIAALVCGAVLFWRNWHRKNTNTIHFANPVYQKTT-EDEVHIC
Stickleback 1    GTVFSAV-SGNRLAASPEEAHLSH-PVALYVVLPILVMSLLVFGAMWVWRHWRLKNTNTIHFDNPVYQKTT-EDEMHIC
Tilapia 1        MLTFSKV-GDRQAAVQPEEAPPSSHPVALYVVLPLMIMALLVFGAMYFWRHWHRKNTNTIHFDNPVYQKTT-EDELHIC
Platyfish 2      PVTTPGV-PPKGVAVIPETTSST--SKTLYFVLPLAILCLVAVGGVLLWRYYRLQNTNTMHFHNPVYQKTT-EDQVHIW
Stickleback 2    PVTSPDF-KHRFVAAIPSTASPA--PISLYIALPLALVCLVAGGGVLLWRNYRLKNTNTIHFDNPVYQKTT-EDQVHIY
Tilapia 2        PVTSPN--IRPEFTEVLTKTVSTT-PIALYVILPLAVAALVAVGGVLLWRNYRLKNTSTIHFDNPVYQKTT-EEQVHIW
```



Complement-type repeat

EGF precursor homology domain

O-linked sugar domain

Membrane anchor

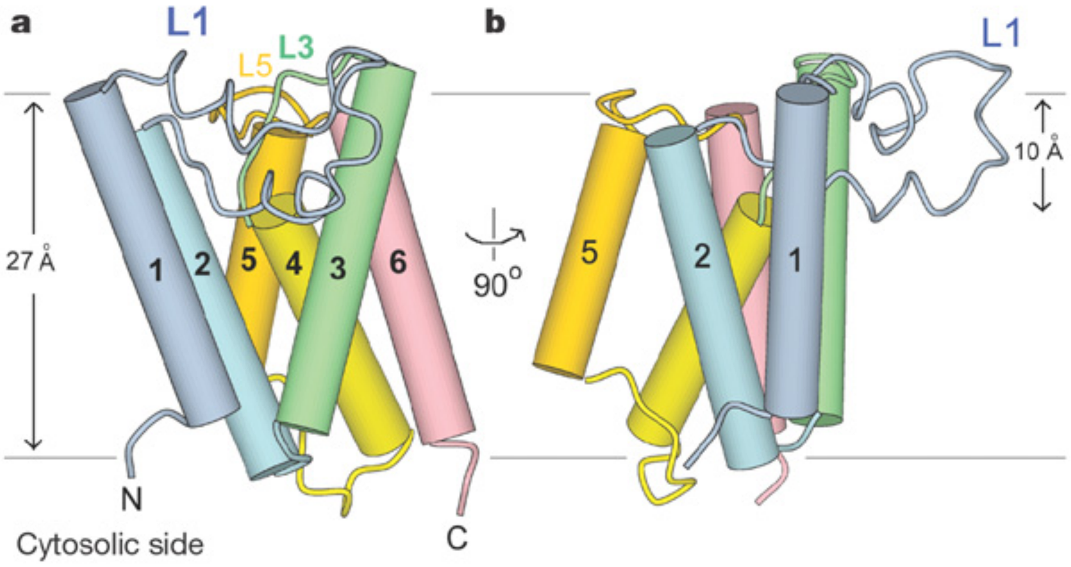
NP*x*Y motif

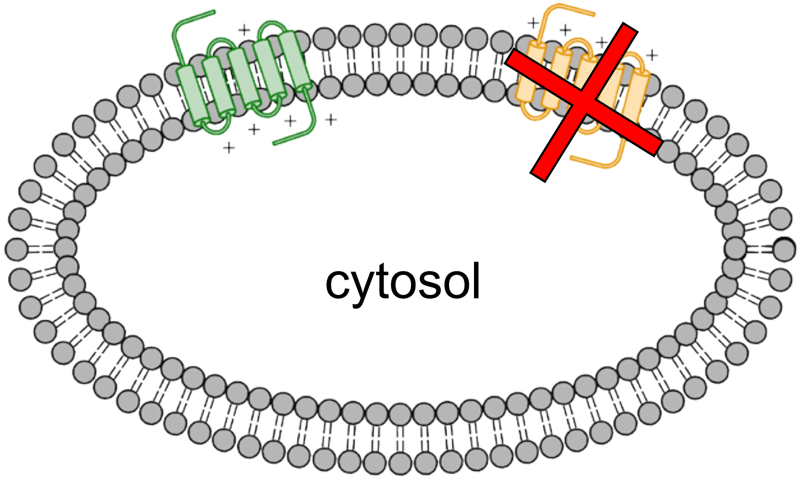
T.E. Willnow *et al. Nat. Cell Biol.* **1**, E157 (1999)

T.B. Strøm *et al. FEBS Open Bio.* **4**, 321 (2014)

Transmembrane (TM) proteins – Secondary structure prediction

- Prediction of membrane orientation (in-out)
- *Positive-inside rule:* Residues at cytosolic side are more positively charged than at the lumenal/periplasmic side





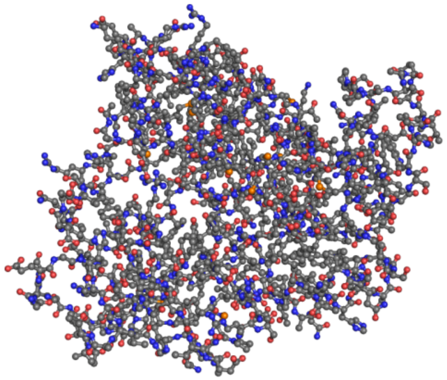
2NRF

Y. Wang, *et al.* Nature **444**, 179 (2006)

3D structure modeling

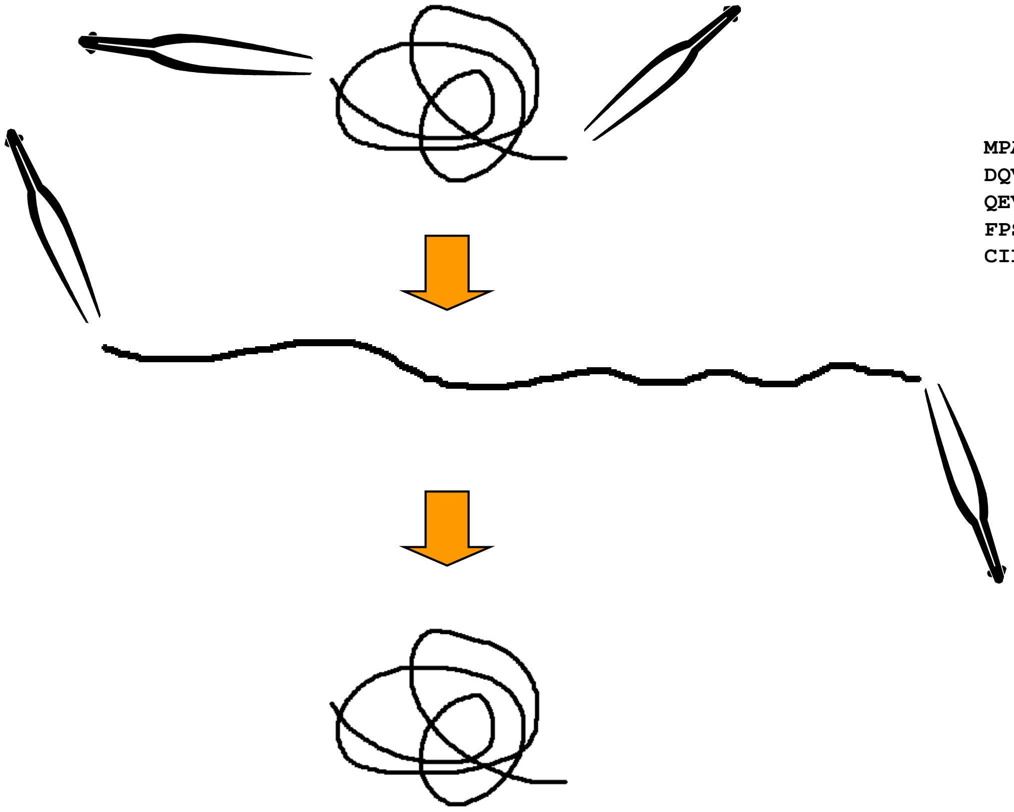
Modeling of 3D structure

- ~135,000,000 sequence records in the traditional GenBank divisions (Apr 2011)
  - Several orders of magnitude more sequences in other public databases
  - Next Generation Sequencing generates ~20 Gb in *a single run*
- ~135,000 3D structures in the PDB (*i.e.* all published structures)
  - Solving a single structure experimentally takes 1-3 yrs
  - Some protein structures are “close to impossible” to solve, *e.g.* many membrane proteins
- In the cell, the sequence determines the 3D structure of the protein



Folding is spontaneous in the cell (but often with helper molecules, chaperones)

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
```

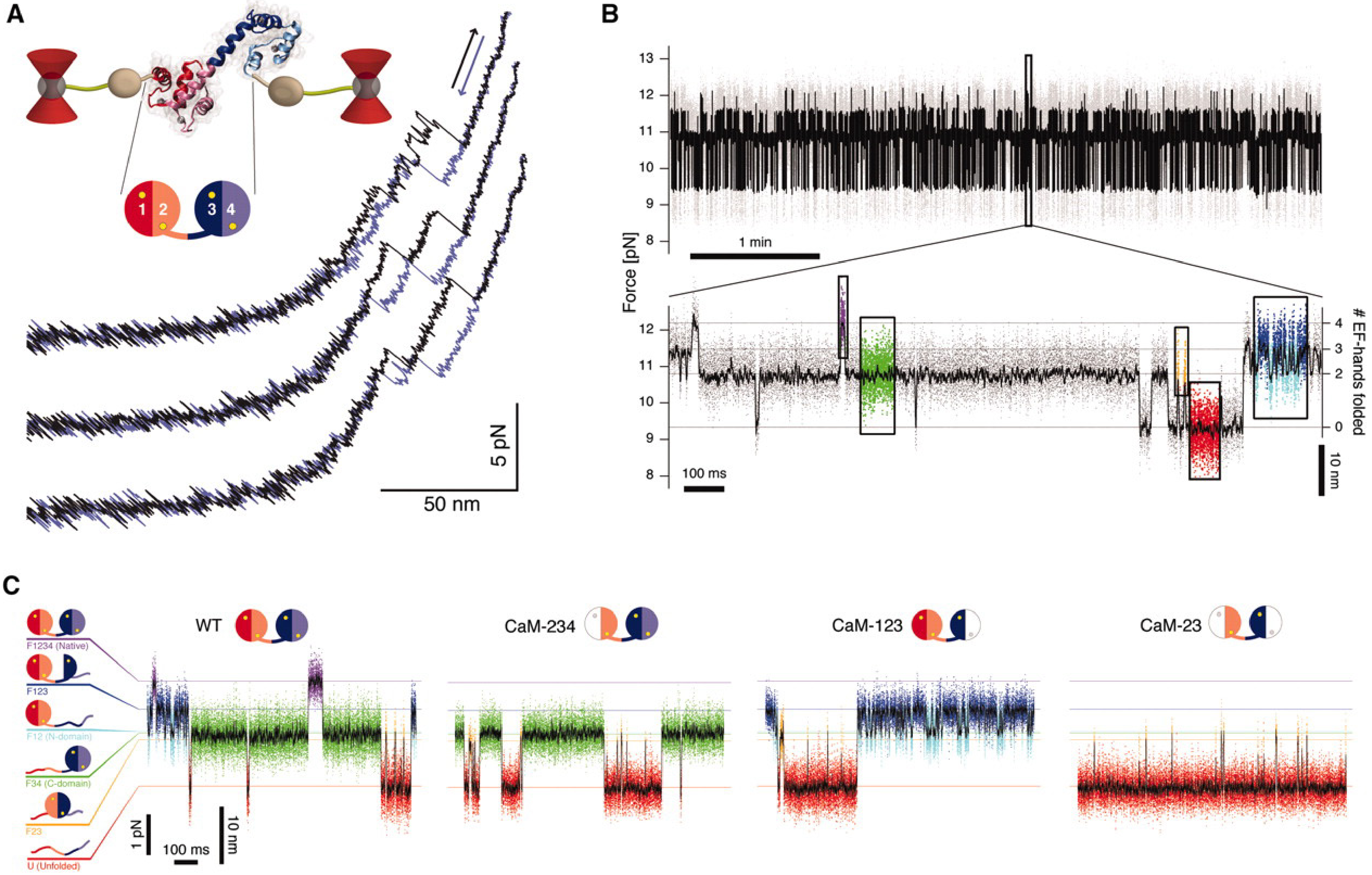


The sequence determines the 3D structure!

Nobel Prize in chemistry 1972 to Christian B. Anfinsen

Optical tweezers

Jon K. Lærdahl,
Structural Bioinformatics



Stigler *et al.*, Science **334**, 512 (2011).

Protein folding

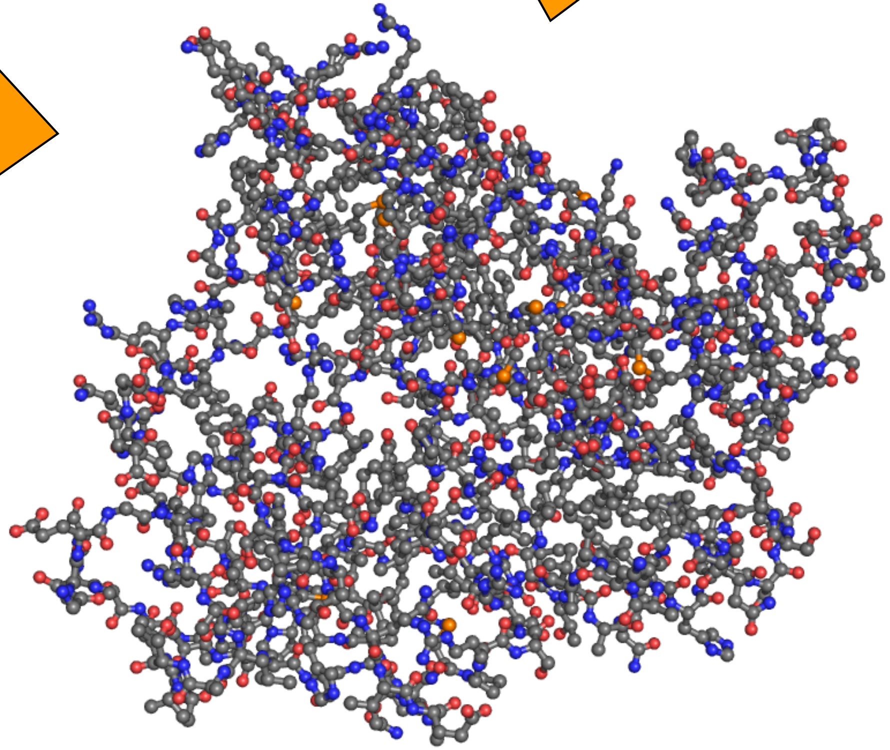
```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
NFFRSLWGPYAGWAQATPPSYRCCSVPTCANPAMLRSHQQSAERVPKGRKARWGTLDKEI
```

Folding is spontaneous in the cell

In the cell

The sequence determines the 3D structure!

In the computer

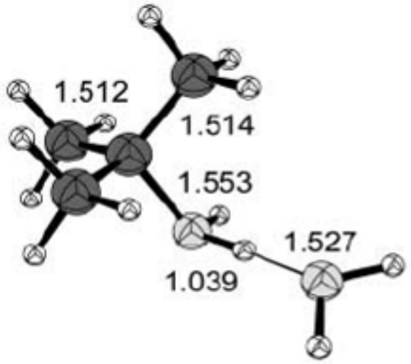


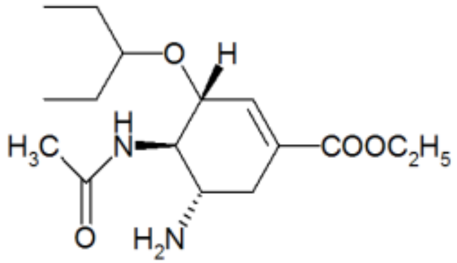
***Ab initio/de novo* structure prediction**

- Based on physical/chemical laws and not already published experimental structures

Ab initio structural prediction

- Determine the tertiary structure for a protein based on amino acid sequence and chemical and physical laws only
- Does *not* use prior knowledge of structure from the PDB
- *Ab initio* quantum chemistry is pure "*ab initio*"
  - Based on solving the Schrödinger equation
  - Is routinely used for chemical systems of up to 20-50 atoms
  - Can be used to compute/model the correct 3D structure for drug candidates, small metabolites or tiny peptides
  - Will *not soon* be applicable for large proteins with 1000s of atoms

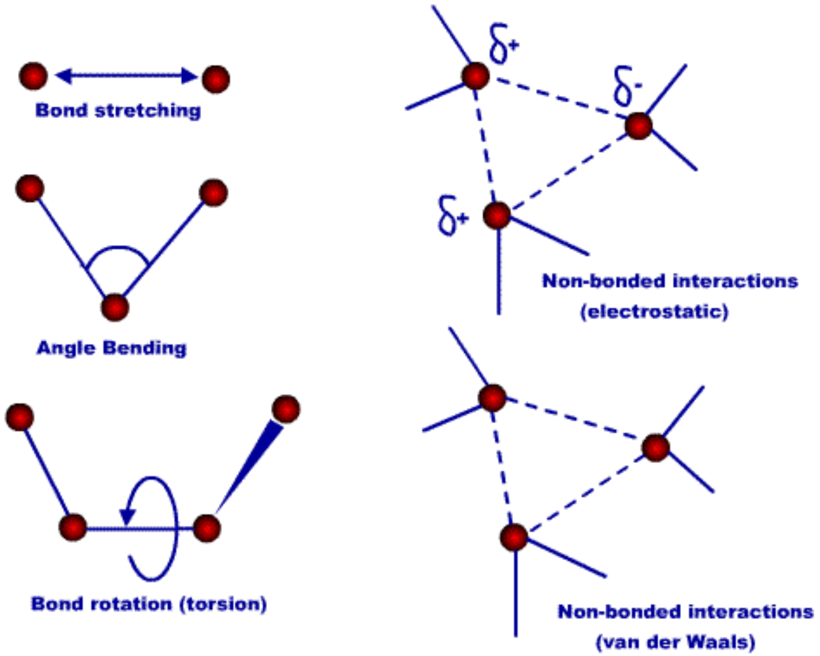




- *Ab initio* protein 3D structure prediction
  - Also called *de novo* structure prediction/protein modeling
  - Is *not* based on solving the Schrödinger equation
  - Instead uses more approximate methods for energy minimization/folding (Confusing: This is exactly what is *not ab initio* quantum chemistry)
  - Extremely computationally intensive
  - Very hard! This field is far from mature…
  - Only possible (useful/reliable) for small (poly)peptides (less than 10-100 residues?)

Ab initio structural prediction

- Molecular mechanics/force field calculations – Newtonian mechanics to model proteins
    - Each atom simulated as a single particle
    - Each particle has a size (van der Waals radius), charge and polarizability
    - Bonded interactions are treated as “springs” with a given equilibrium bond distance – same for bond angles and dihedral angles
    - Additional terms, *e.g.* non-bonded collisions, solvent etc.

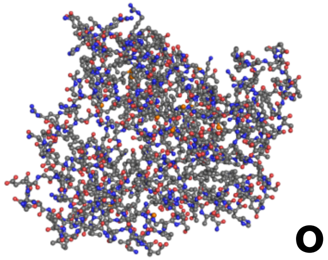


$$
\begin{aligned}
U(\vec{R}) = {} & \sum_{\text{bonds}} K_{\text{b}}(b-b_0)^2 + \sum_{\text{angles}} K_{\theta}(\theta-\theta_0)^2 \\
& + \sum_{\text{Urey-Bradley}} K_{\text{UB}}(S-S_0)^2 \\
& + \sum_{\text{dihedrals}} K_{\varphi}(1+\cos(n\varphi-\delta)) + \sum_{\text{impropers}} K_{\omega}(\omega-\omega_0)^2 \\
& + \sum_{\substack{\text{non-bonded}\\ \text{pairs}}} \left\{ \varepsilon_{ij}^{\text{min}} \left[\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\} \\
& + \sum_{\text{residues}} U_{\text{CMAP}}(\varphi,\psi)
\end{aligned}
$$

Brooks *et al.*, J. Comput. Chem. **30**, 1545 (2009).

Ab initio structural prediction

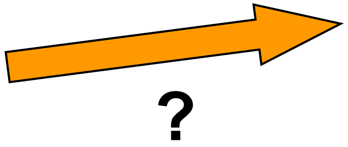
- Does *not* use prior knowledge of structure from the PDB
    - That is why they are known as *ab initio*
- Still, some programs known as *ab initio* protein modeling programs also use *some* information from the PDB, for example structures for small fragments
- At least in some respects based on the "paradigm" of Anfinsen that all information that is needed to determine the tertiary structure is in the primary sequence
    - Is it really correct?
    - Certainly not always!
        - Folding chaperons
        - Ribosomal environment, timing of protein synthesis, solvent, salinity, pH, temperature, metabolites and other macromolecules, etc. may (and do) in many cases contribute to the folding process

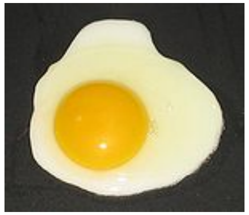


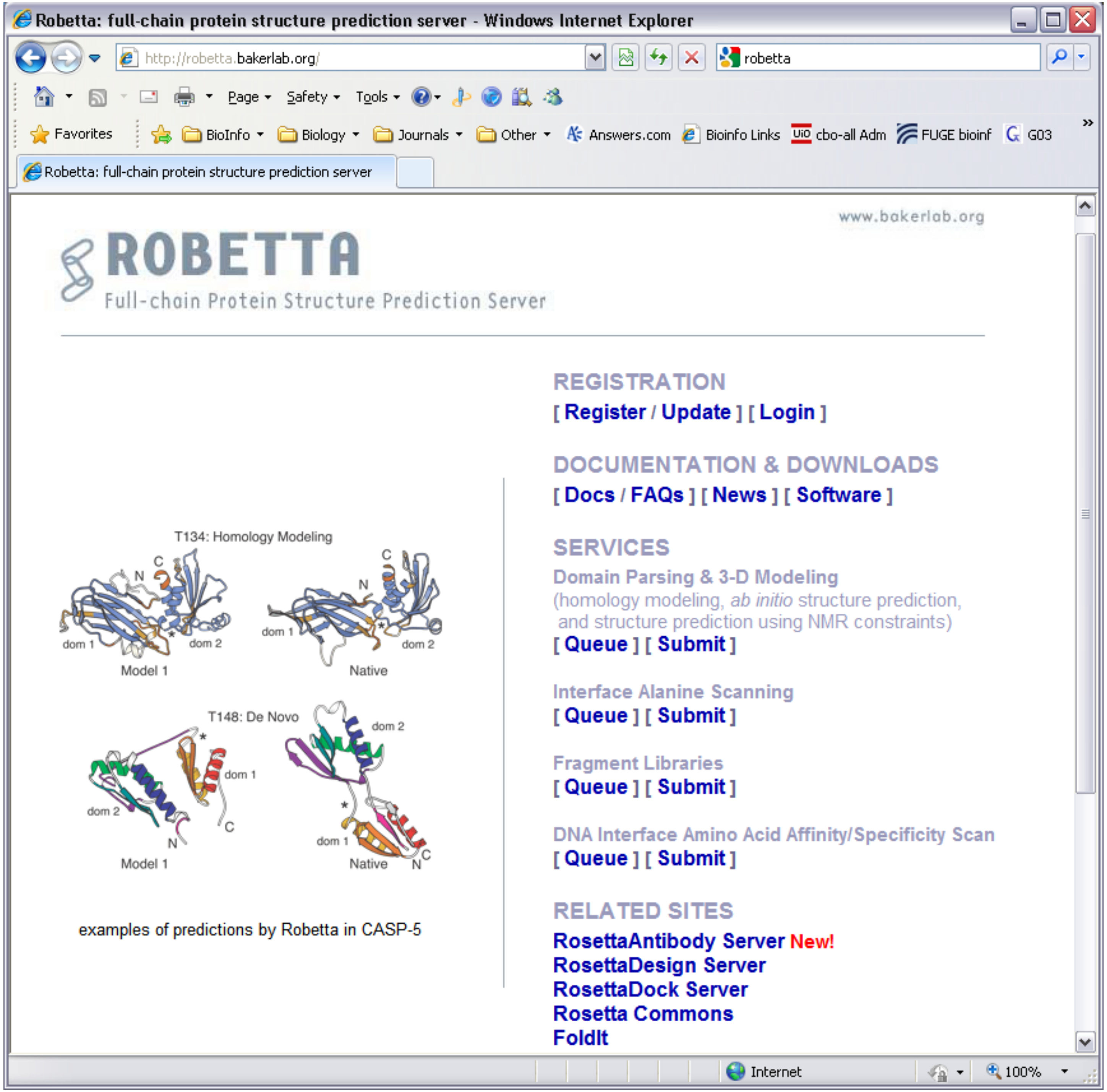
- All problems with *ab initio* modeling will never be completely solved?
- They have certainly not been solved yet!

or

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
```







www.bakerlab.org

ROBETTA

Full-chain Protein Structure Prediction Server

examples of predictions by Robetta in CASP-5

REGISTRATION

[Register / Update] [Login]

DOCUMENTATION & DOWNLOADS

[Docs / FAQs] [News] [Software]

SERVICES

Domain Parsing & 3-D Modeling
(homology modeling, *ab initio* structure prediction, and structure prediction using NMR constraints)
[Queue] [Submit]

Interface Alanine Scanning
[Queue] [Submit]

Fragment Libraries
[Queue] [Submit]

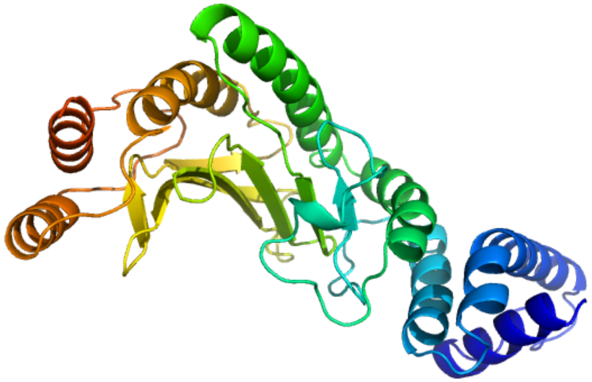
DNA Interface Amino Acid Affinity/Specificity Scan
[Queue] [Submit]

RELATED SITES

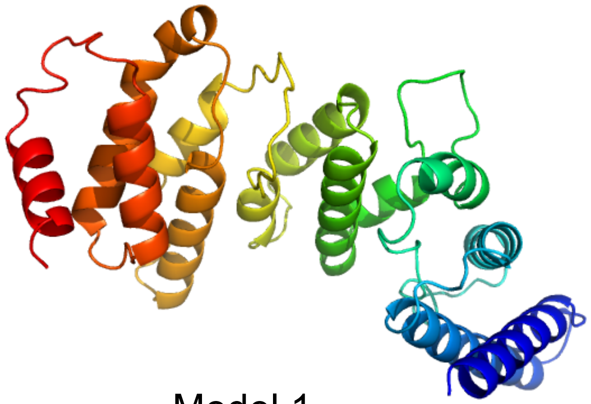
RosettaAntibody Server New!
RosettaDesign Server
RosettaDock Server
Rosetta Commons
FoldIt

David Baker

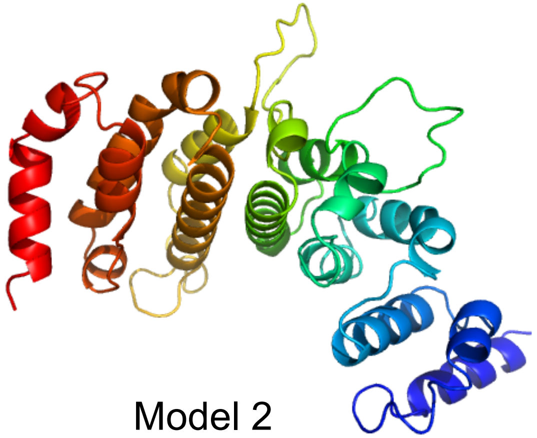
I-TASSER from Yang Zhang-lab is another possibility. Ranked as no. 1 in ”structure prediction competition” in 2006, 2008, 2012, and 2014 (Actually not pure *ab initio*).



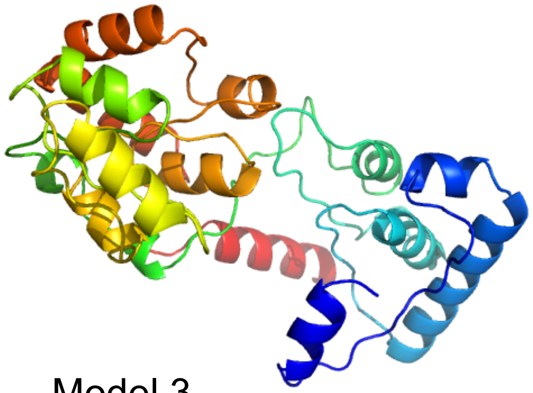
Experimental 3D structure of my colleague



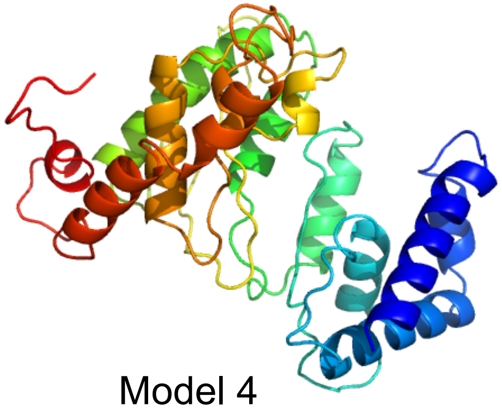
Model 1



Model 2



Model 3



Model 4

Model 5

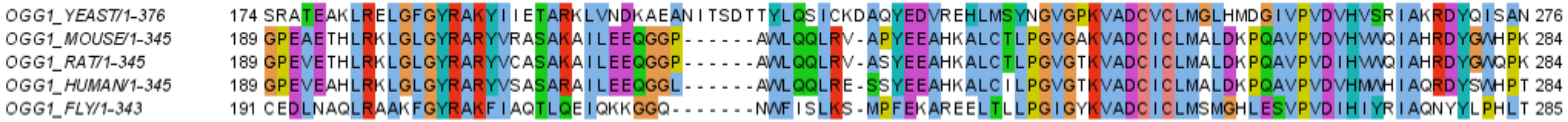
3D structure modeling

- *Ab initio/de novo* – very hard…

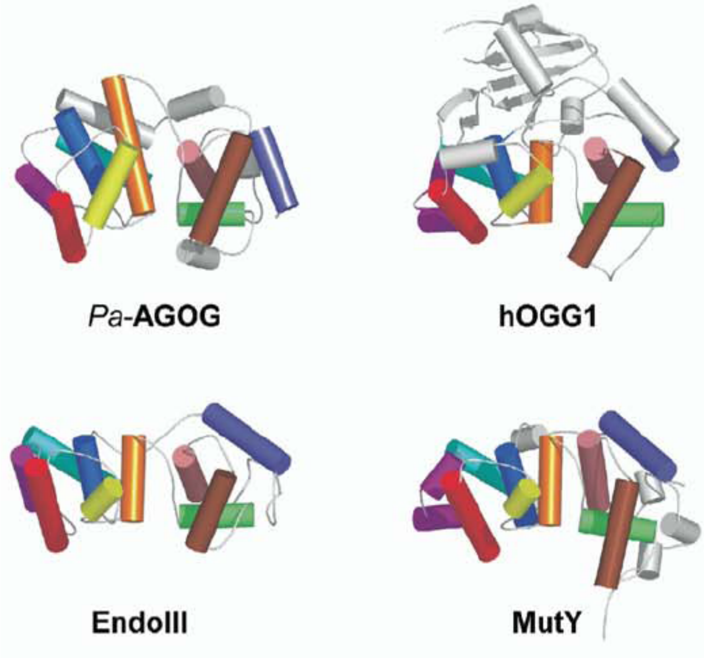
- Threading/fold recognition
- Homology modeling

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics

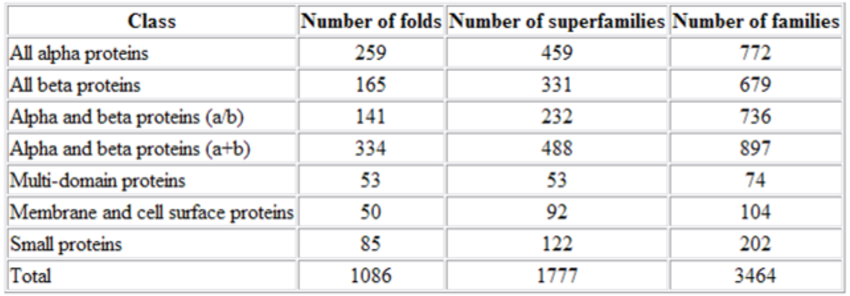


```
OGG1_YEAST/1-376   174 SRATEAKLRELGFGYRAKYIIETARKLVNDKAEANITSDTTYLQSICKDAQYEDVREHLMSYNGVGPKVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISAN 276
OGG1_MOUSE/1-345   189 GPEAETHLRKLGLGYRARYVRASAKAILEEQGGP------AWLQQLRV-APYEEAHKALCTLPGVGAKVADCICLMALDKPQAVPVDVHVWQIAHRDYGWHPK 284
OGG1_RAT/1-345     189 GPEVETHLRKLGLGYRARYVCASAKAILEEQGGP------AWLQQLRV-ASYEEAHKALCTLPGVGTKVADCICLMALDKPQAVPVDIHVWQIAHRDYGWQPK 284
OGG1_HUMAN/1-345   189 GPEVEAHLRKLGLGYRARYVSASARAILEEQGGL------AWLQQLRE-SSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPT 284
OGG1_FLY/1-343     191 CEDLNAQLRAAKFGYRAKFIAQTLQEIQKKGGQ-------NWFISLKS-MPFEKAREELTLLPGIGYKVADCICLMSMGHLESVPVDIHIYRIAQNYYLPHLT 285
```



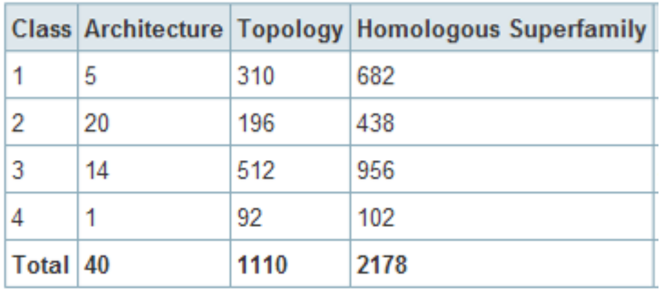
- Reason for similarities in sequence/structure is ***common ancestry***, the sequences/structures are ***homologs***
- Structures evolves slowly
- Sequence evolves faster
  - Many mutations does not change the structure
- Only some few 1000 superfamilies in the PDB
- Only a factor 2-10(???) as many superfamilies in Nature? Some few 1000 folds?

SCOP



Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

CATH



Class	Architecture	Topology	Homologous Superfamily
1	5	310	682
2	20	196	438
3	14	512	956
4	1	92	102
Total	**40**	**1110**	**2178**