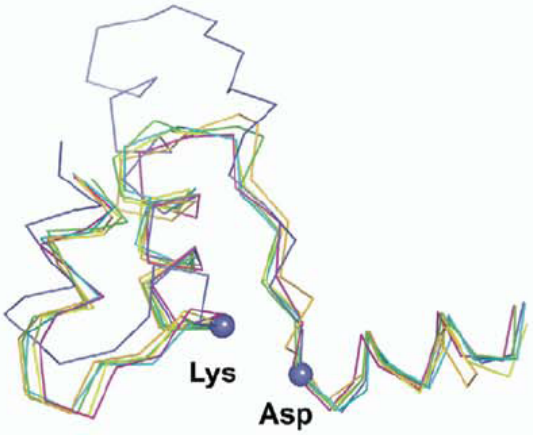
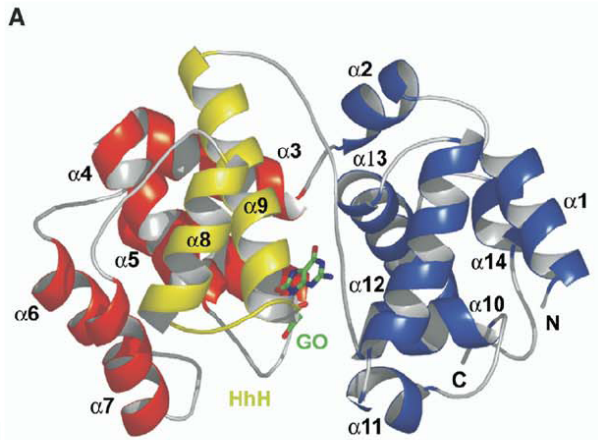
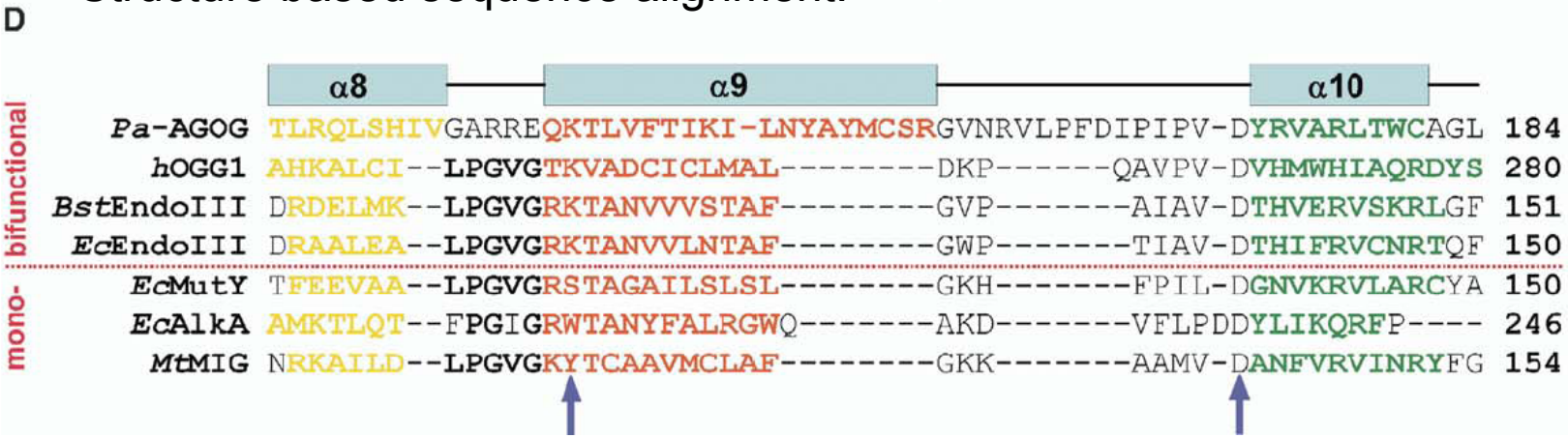
Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics





Structure based sequence alignment:



G.M. Lingaraju *et al. Structure* **13**, 87 (2005)

Hardly any detectable sequence similarity to human OGG1, *E. coli* EndoIII and MutY, and other homologs

Evolution has “eroded away” sequence similarity but left the structure intact

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure alignments

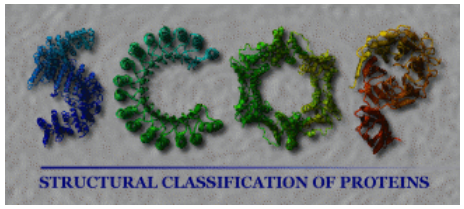
Proteins that fold in the same way, i.e. ”have the same fold” are often homologs.
Structure evolves slower than sequence
Sequence is less conserved than structure

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

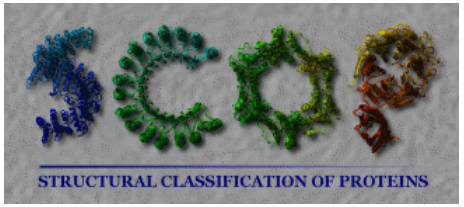
Use structure comparisons to classify, group and cluster proteins. Build protein structure families and hierarchies



Jon K. Lærdahl,
Structural Bioinformatics

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

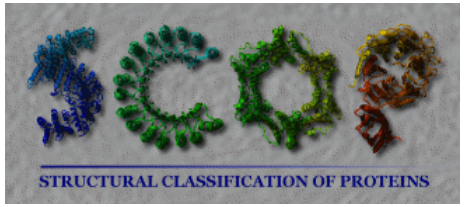
- Almost 100% manually generated
- Proteins grouped into hierarchy of classes, folds, superfamilies and families

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

Jon K. Lærdahl,
Structural Bioinformatics

- 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

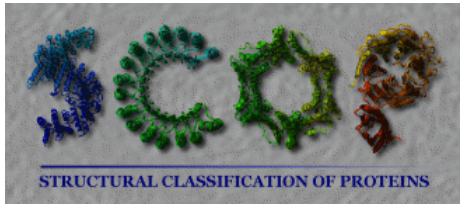
- Classes
  - Upper level of classification (4 major, 3 minor)
  - Contains folds
  - Based on secondary structure composition and “general features”
  - *e.g.* all-α, all-β, ”membrane and cell surface” and ”small proteins”
  - α/β: One β-sheet with strands connected by single α-helices
  - α+β: α-helical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

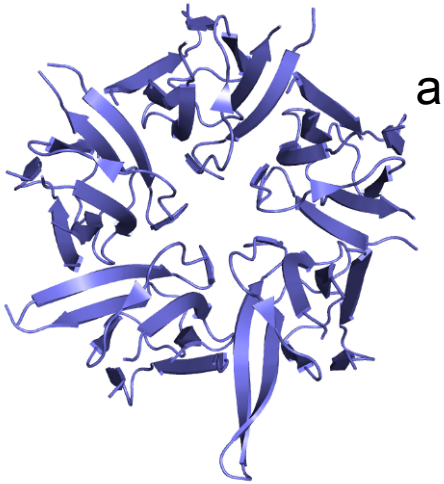
Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins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

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

Jon K. Lærdahl,
Structural Bioinformatics

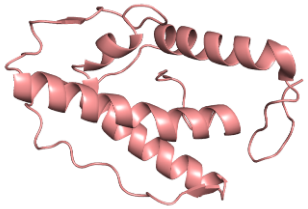
SCOP





all-β class

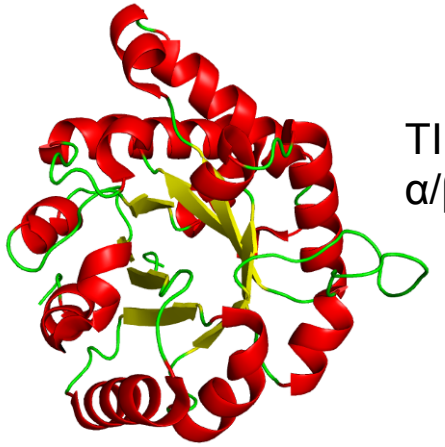
4-helical cytokines



T4 endonuclease V

Globin-like

all-α class,
3 different folds



TIM-barrel fold
α/β class

Profilin-like fold
α+β class

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

Jon K. Lærdahl,
Structural Bioinformatics

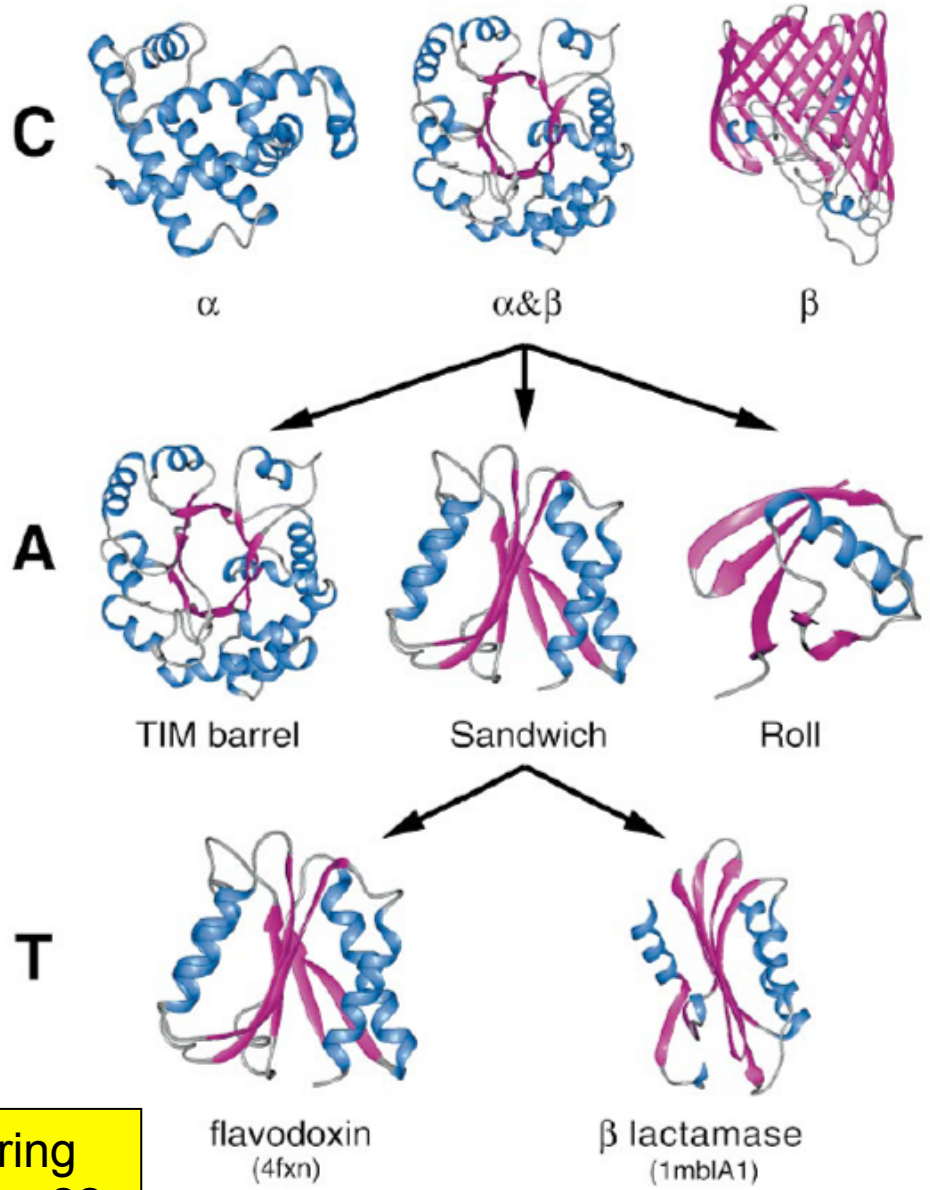
CATH

CATH
PROTEIN STRUCTURE CLASSIFICATION

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)



Explore during the exercises??

http://www.cathdb.info

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

CATH vs. SCOP

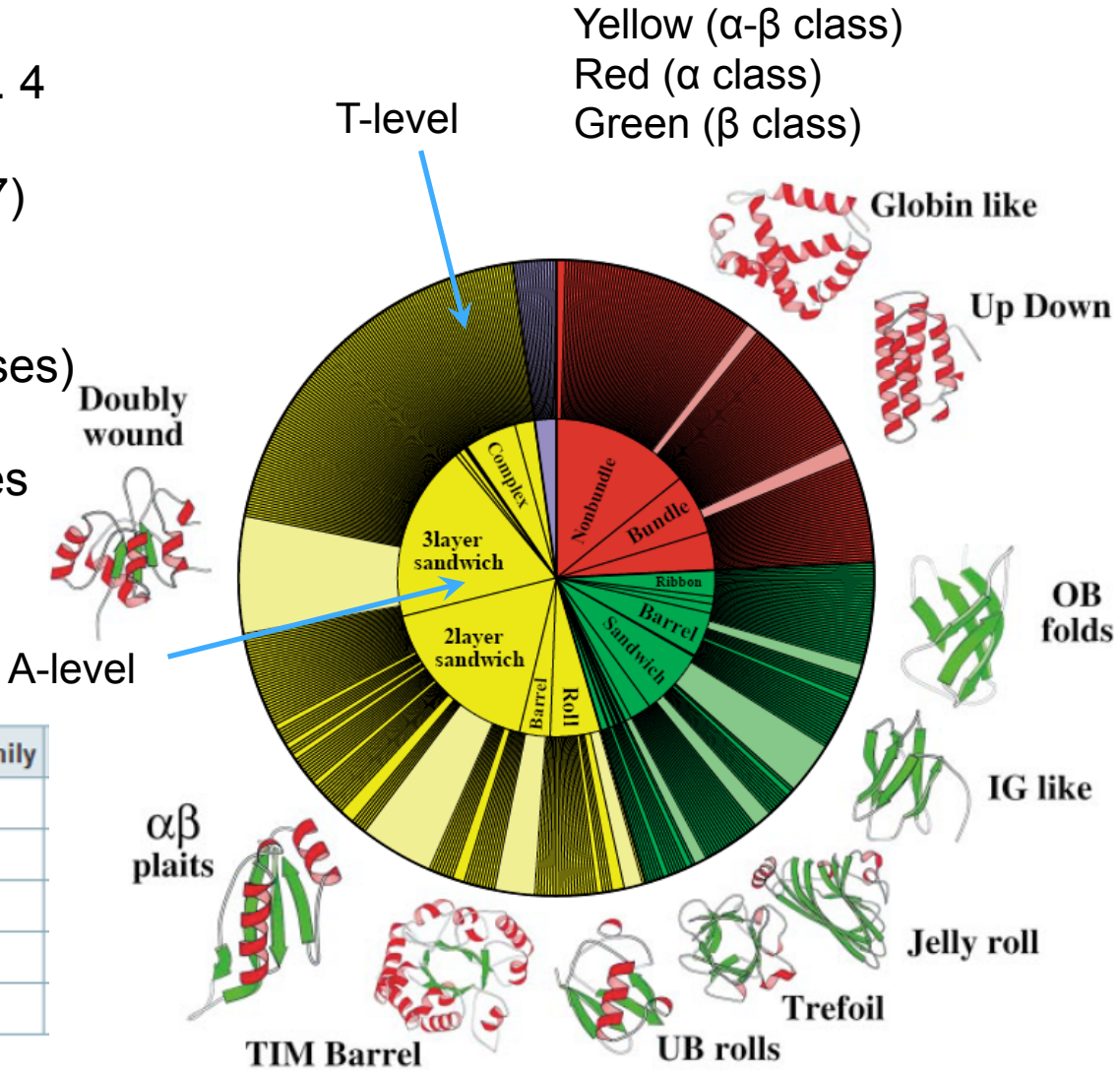
Jon K. Lærdahl,
Structural Bioinformatics

- 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!



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

Jon K. Lærdahl,
Structural Bioinformatics

Predictors

Jon K. Lærdahl,
Structural Bioinformatics

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

Jon K. Lærdahl,
Structural Bioinformatics

Secondary structure prediction

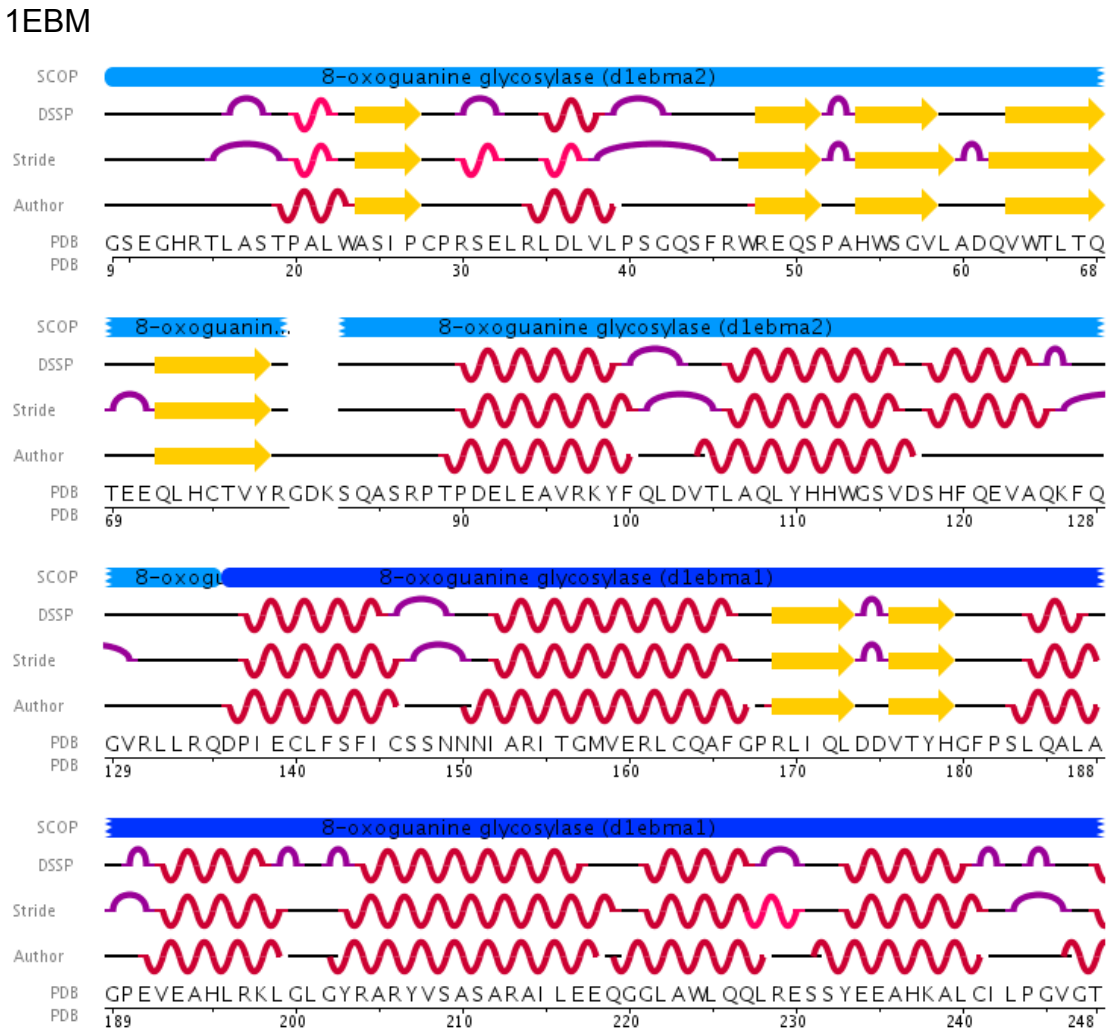
Assigning secondary structure is *not trivial* and there is *no single consensus method* even when 3D structure is known

- 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

β-strand

α-helix

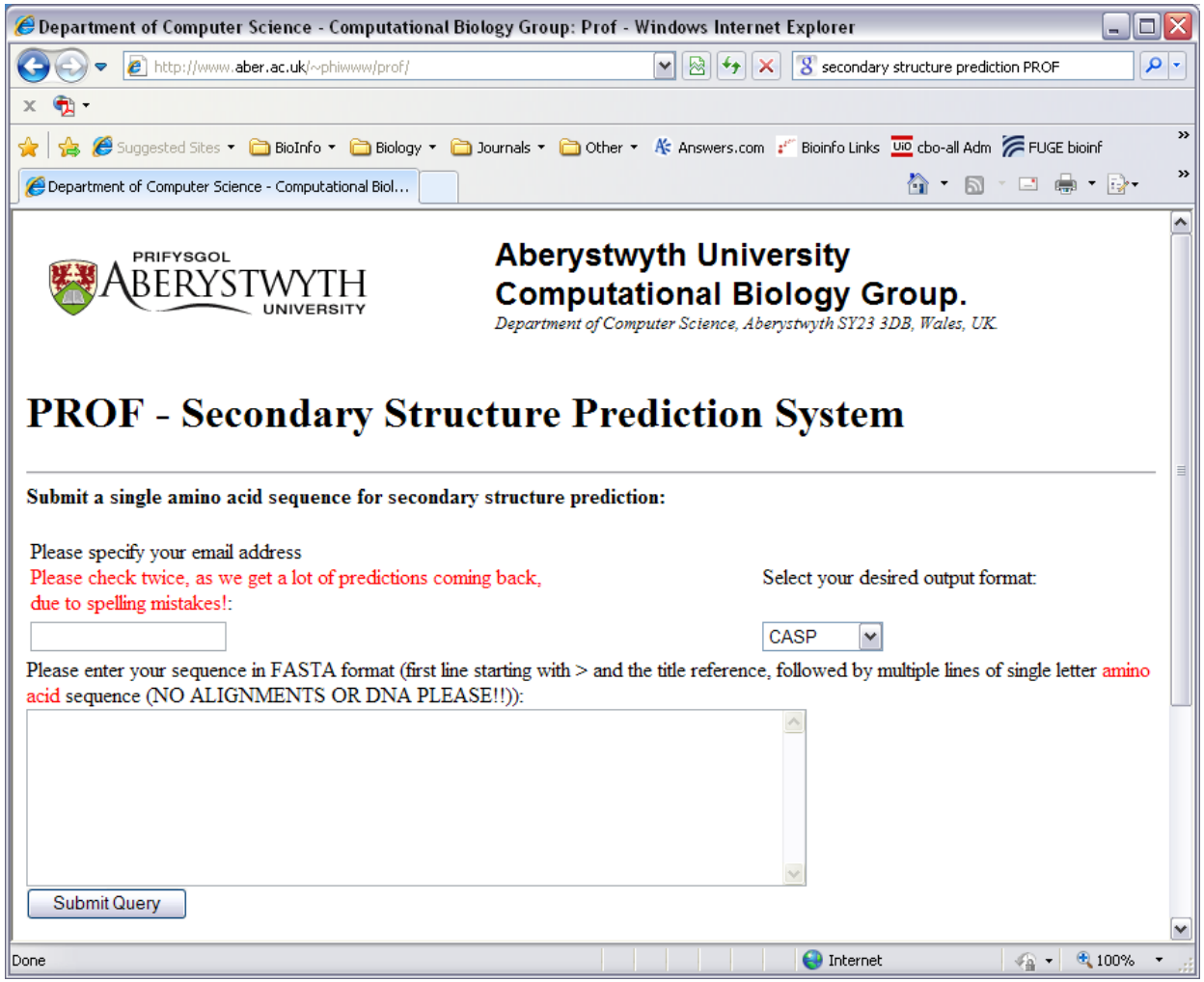
Everything else loop/coil



Jon K. Lærdahl,
Structural Bioinformatics

Secondary structure prediction

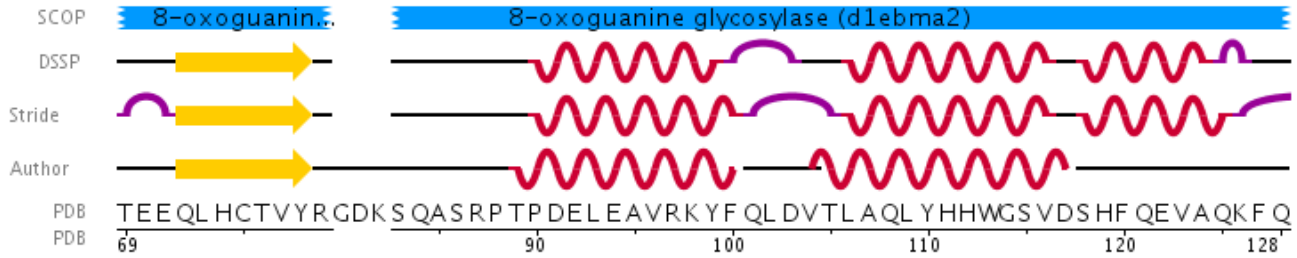
Tools/programs that accept a primary sequence and predicts the secondary structure state (H/helix, E/sheet, or C/Loop&Coil) for each residue



Jon K. Lærdahl,
Structural Bioinformatics

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




```
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

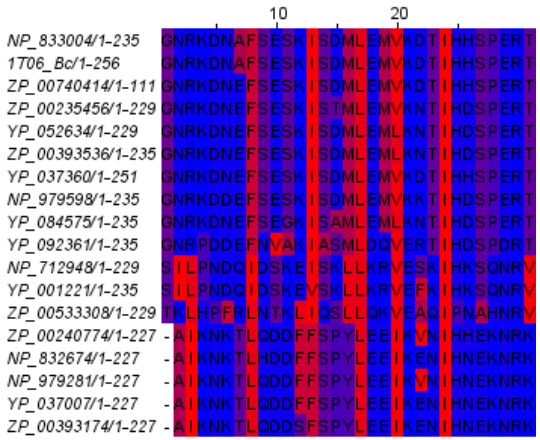
- 3rd generation prediction (until present time) approaching ~80%
- Two main improvements:
    - Machine learning/neural networks
    - Combines information from predictions for single sequence with information from homologous sequences (multiple sequence alignment)

    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)



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

Jon K. Lærdahl,
Structural Bioinformatics

- 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%

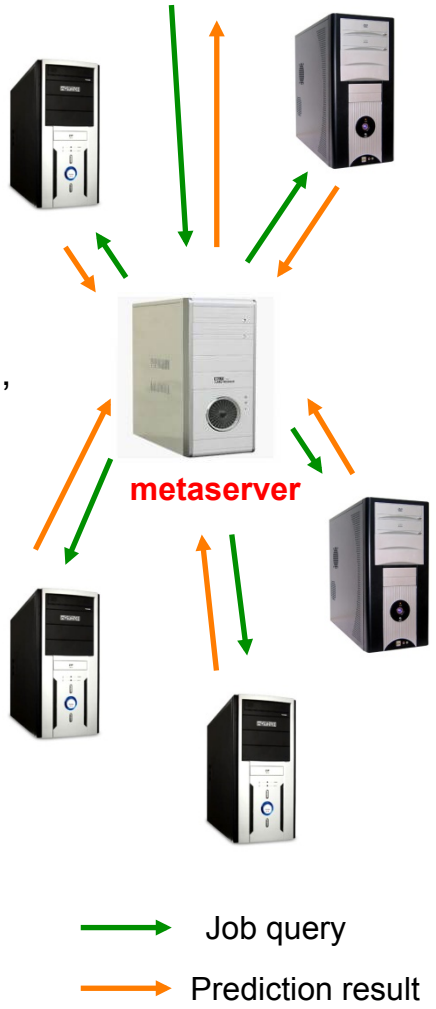
Many (more than 70 different published algorithms!) programs for secondary structure prediction:

- **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)



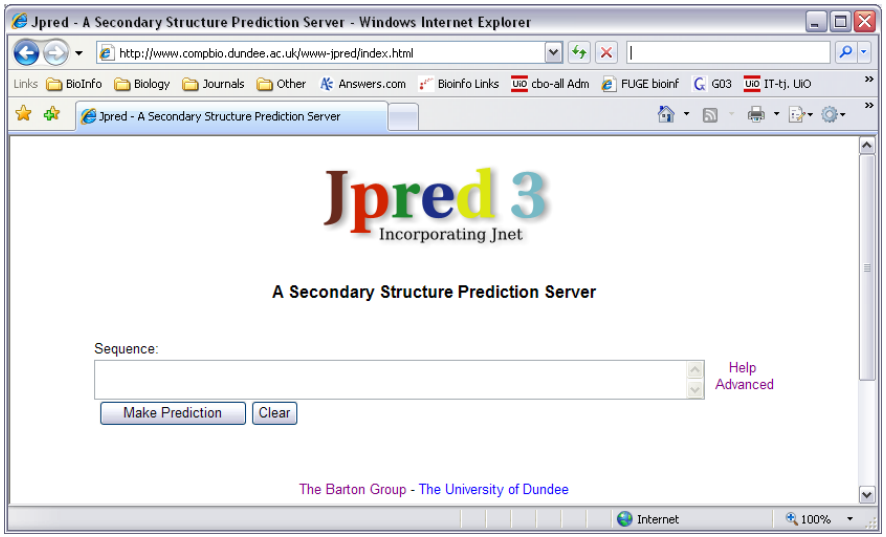
Secondary structure prediction - consensus-based

Jon K. Lærdahl,
Structural Bioinformatics

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

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

jpred            : ------------------HHHHH---EEEE-----EE--HHHHHHHHHH----HHHHHHHHHHH-------HHHHHHHHHHHH---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         : 6888877411038983120225457079955884164432599999882684148999999999758799818789999999886052587 4 : Jnet Rel
```

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

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

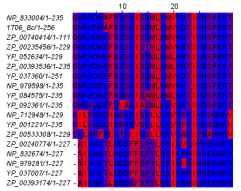


Predictors - *common features*

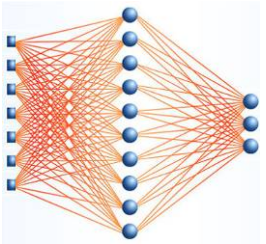
Jon K. Lærdahl,
Structural Bioinformatics

• 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



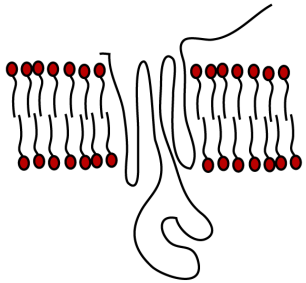
• Use neural networks or similar methods in predictions

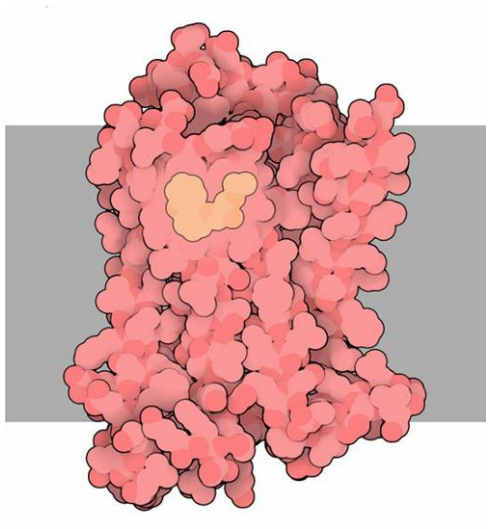
• Consensus from many tools is better than just a single result (e.g. metaservers)

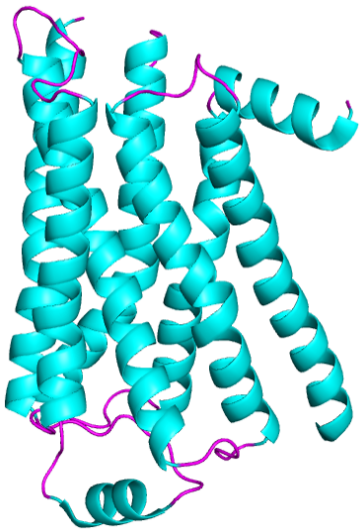
Jon K. Lærdahl,
Structural Bioinformatics

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



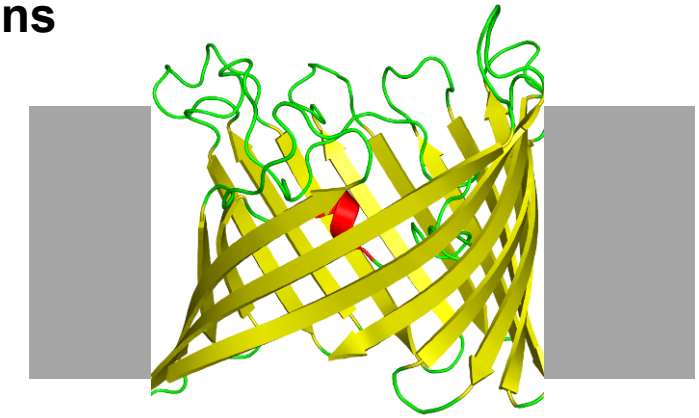




PDB Apr. 08 “Molecule of the Month”

2RH1, Human adrenergic receptor

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

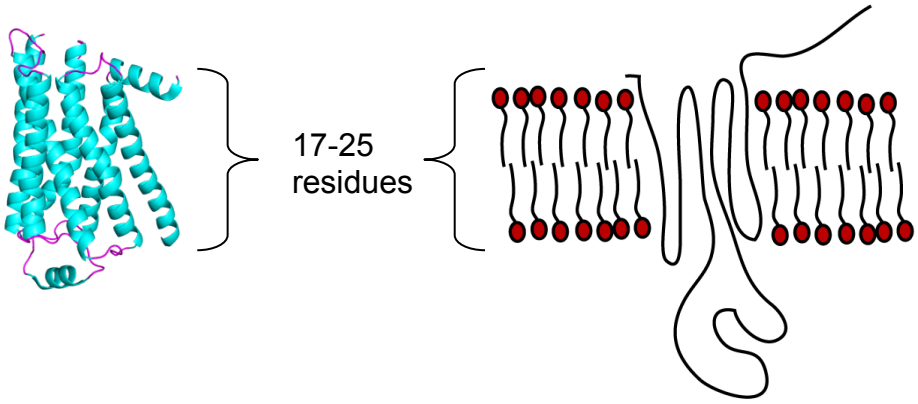


Porin

Transmembrane (TM) proteins

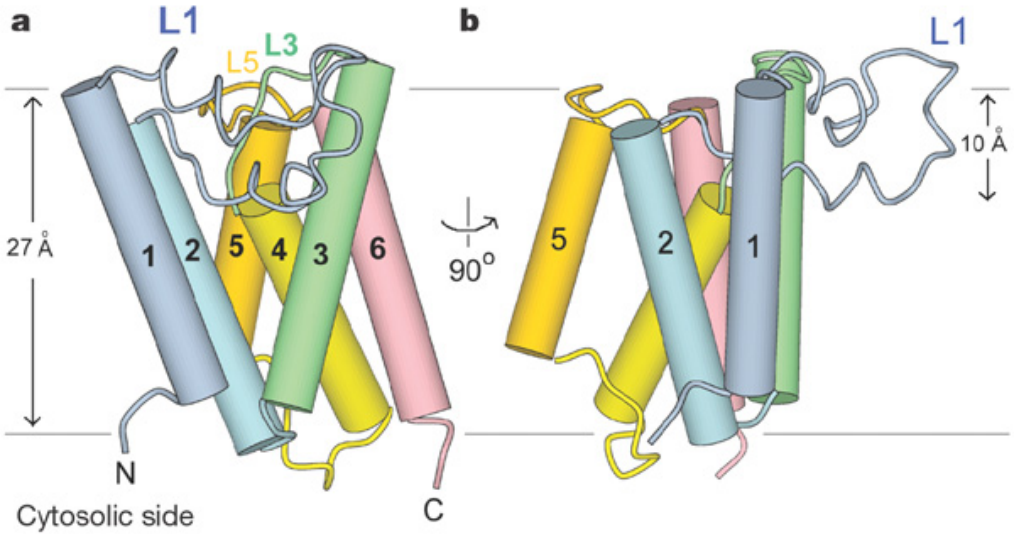
- Extremely difficult to solve membrane structures experimentally!
    - Only a few hundred structures in the PDB
- Can not use the same predictors for secondary structure as for globular proteins
- Special predictors for
    - helical membrane proteins
    - β-barrel proteins
- Pattern in TM α-helical 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

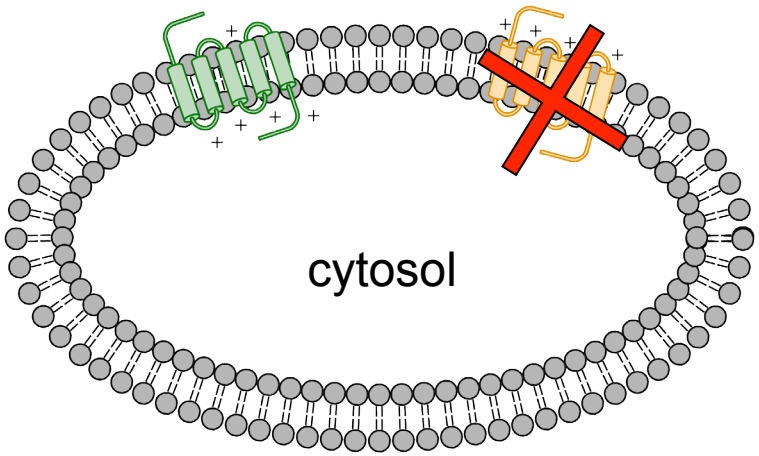


Transmembrane (TM) proteins – Secondary structure prediction

- 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)

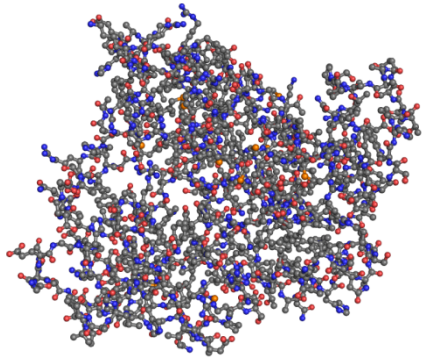
Jon K. Lærdahl,
Structural Bioinformatics

3D structure modeling

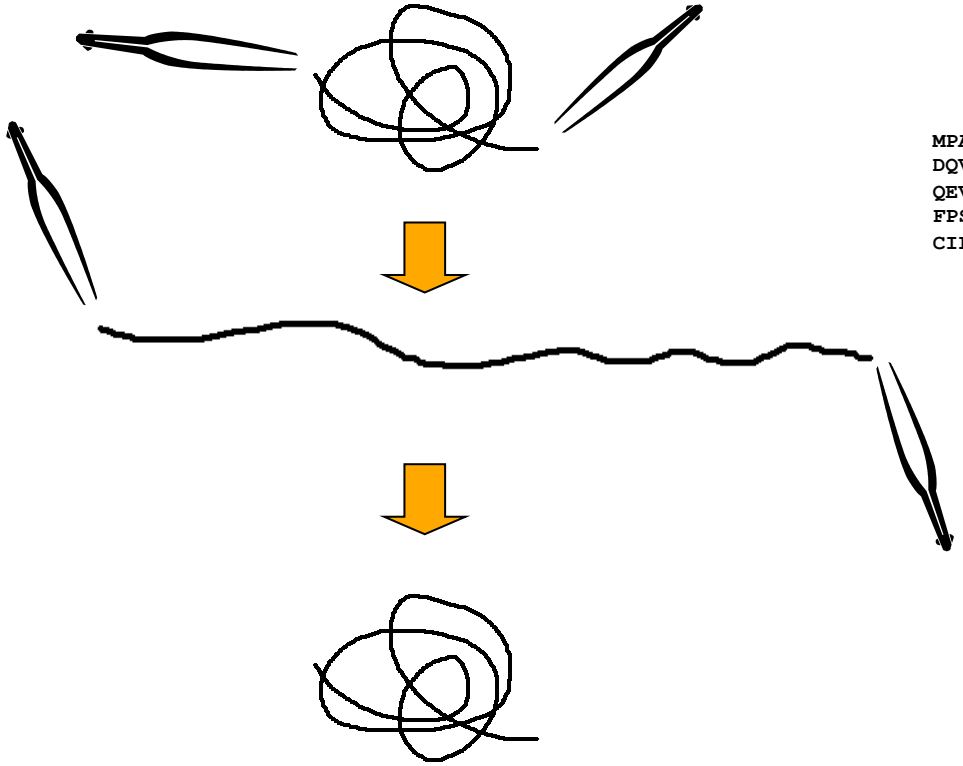
Modeling of 3D structure

Jon K. Lærdahl,
Structural Bioinformatics

- ~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*
- ~104,000 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)



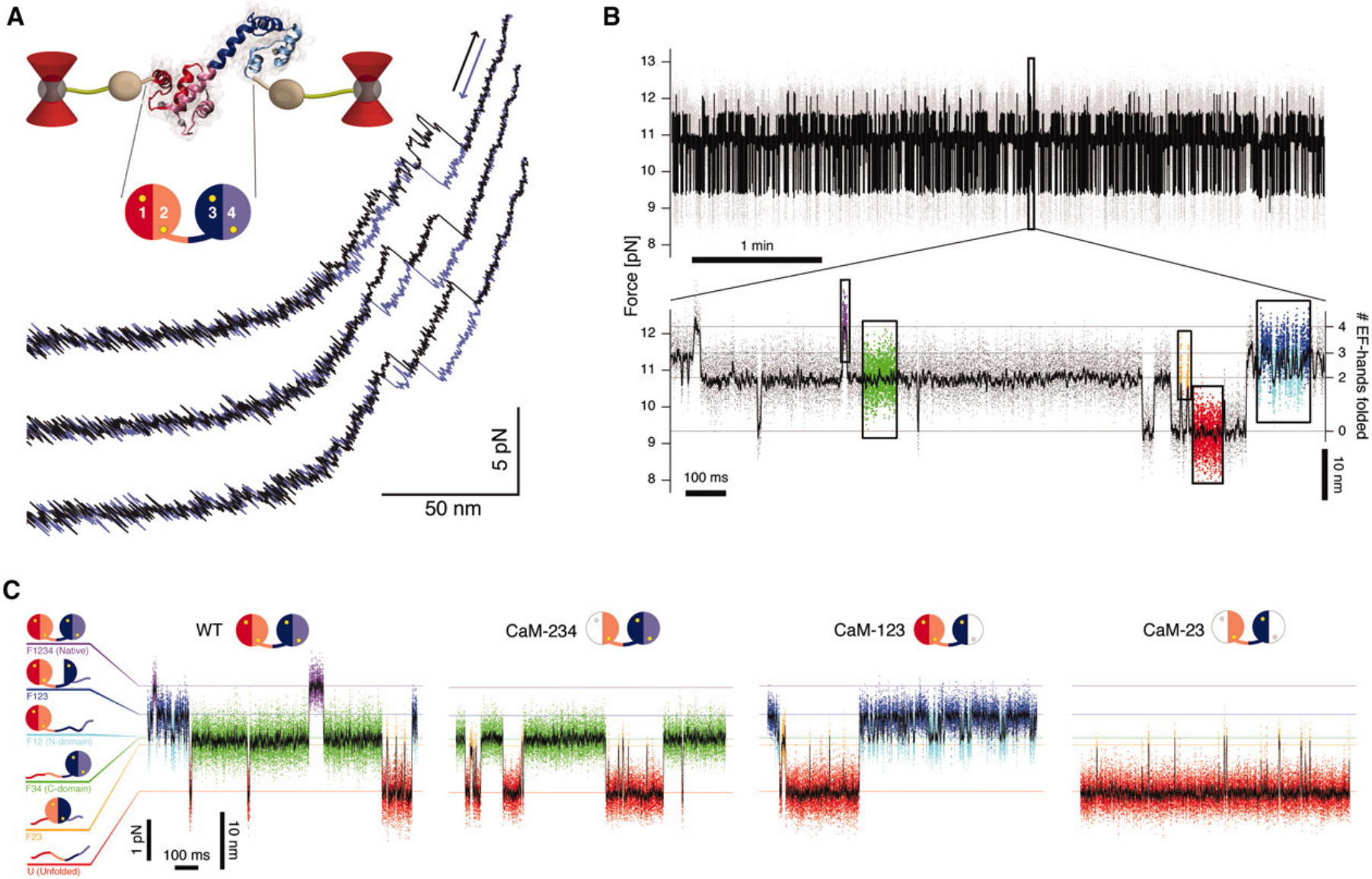
```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
```

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).

Jon K. Lærdahl,
Structural Bioinformatics

Protein folding

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

Folding is spontaneous in the cell

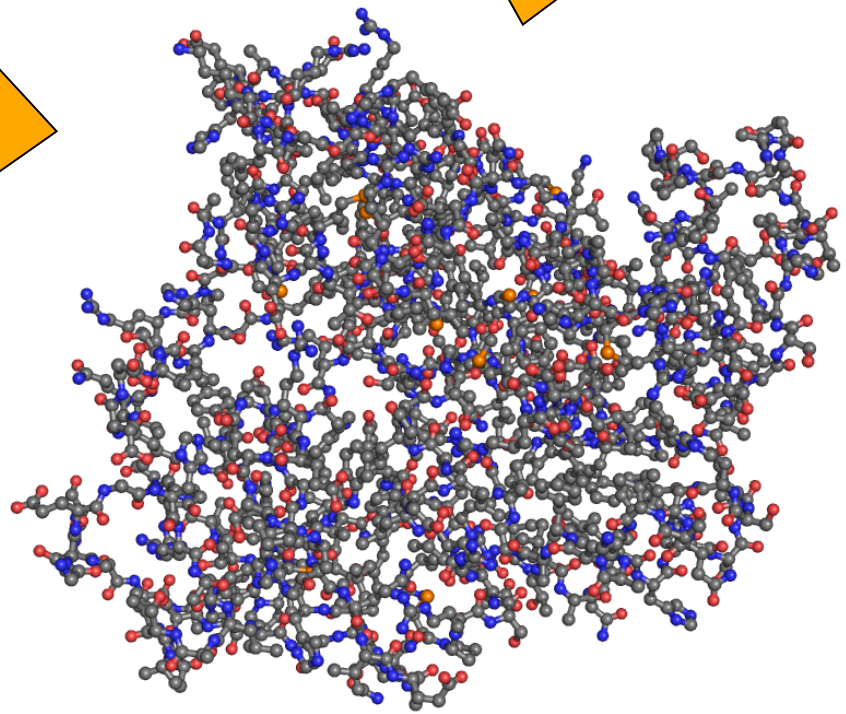
In the cell

The sequence determines the 3D structure!

In the computer

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

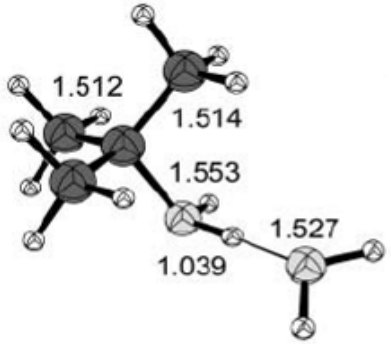
- Based on physical/chemical laws and not already published experimental structures

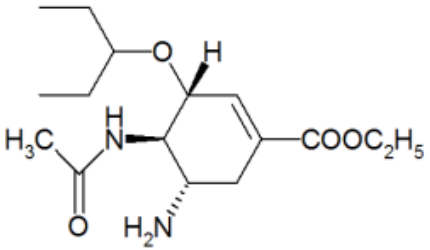


Jon K. Lærdahl,
Structural Bioinformatics

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 for small (poly)peptides (less than 10-100 residues?)

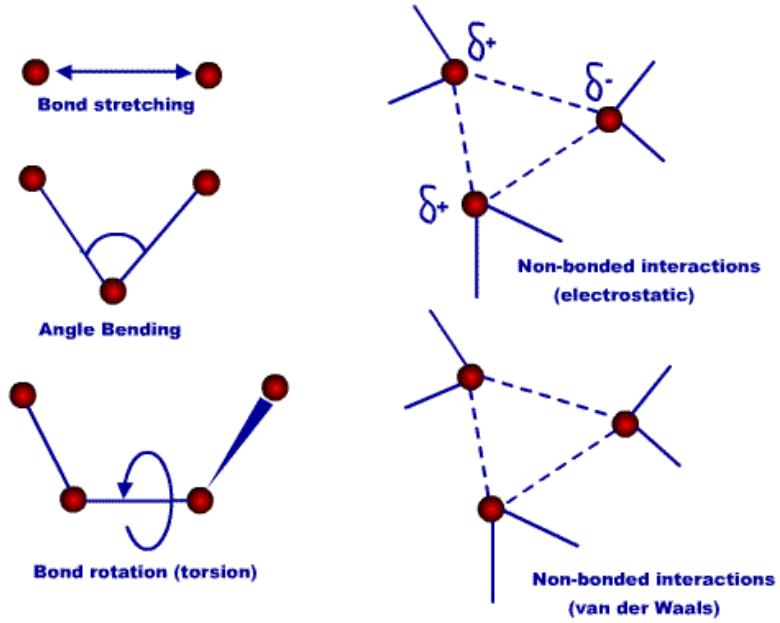




Jon K. Lærdahl,
Structural Bioinformatics

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.



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

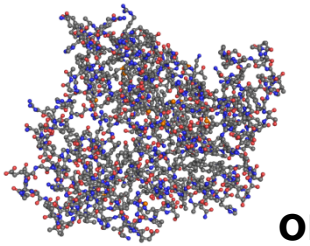
Brooks *et al.*, J. Comput. Chem. **30**, 1545 (2009).

Ab initio structural prediction

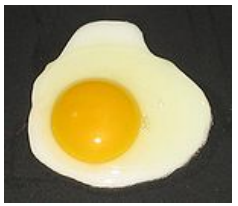
Jon K. Lærdahl,
Structural Bioinformatics

- 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

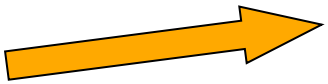
- All problems with *ab initio* modeling will never be completely solved?
- They have certainly not been solved yet!



or



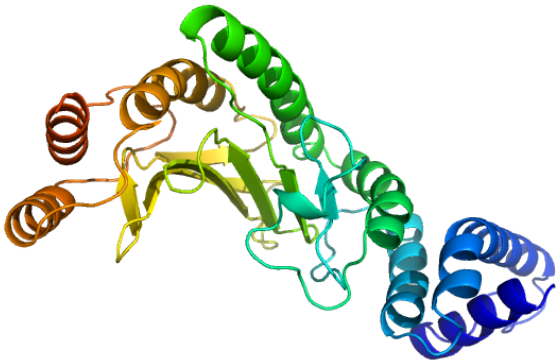
```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
```



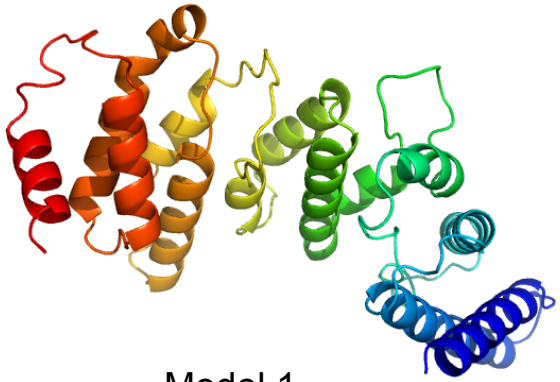
?

Jon K. Lærdahl,
Structural Bioinformatics

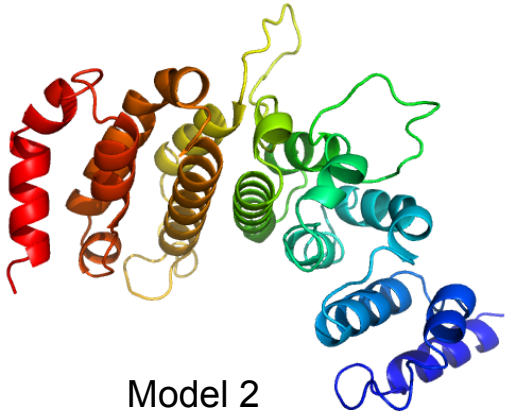
I-TASSER from Yang Zhang-lab is another possibility. Ranked as no. 1 in ”structure prediction competition” in 2006, 2008, and 2010.



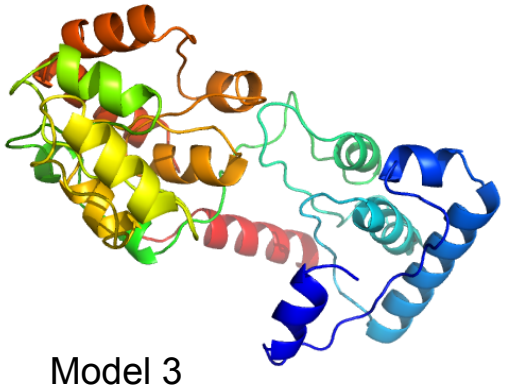
Experimental 3D structure of my colleague



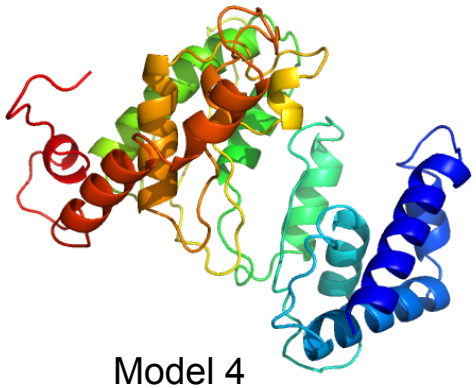
Model 1



Model 2

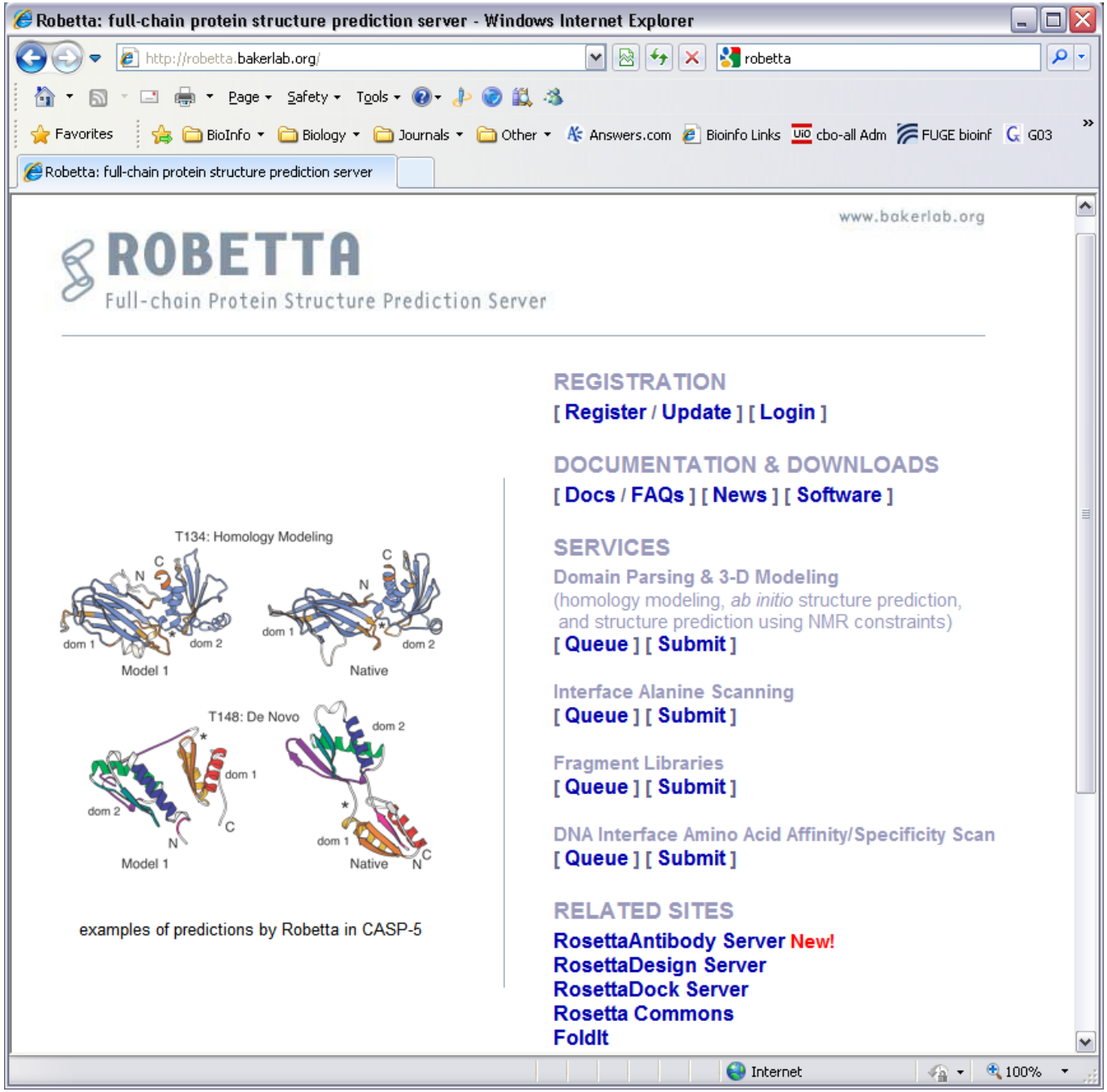


Model 3



Model 4

Model 5


Robetta: full-chain protein structure prediction server - Windows Internet Explorer
http://robetta.bakerlab.org/
robetta
Page
Safety
Tools
Favorites
BioInfo
Biology
Journals
Other
Answers.com
Bioinfo Links
cbo-all Adm
FUGE bioinf
G03
Robetta: full-chain protein structure prediction server
www.bakerlab.org
ROBETTA
Full-chain Protein Structure Prediction Server
T134: Homology Modeling
dom 1
dom 2
Model 1
Native
T148: De Novo
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
Internet
100%

David Baker

3D structure modeling

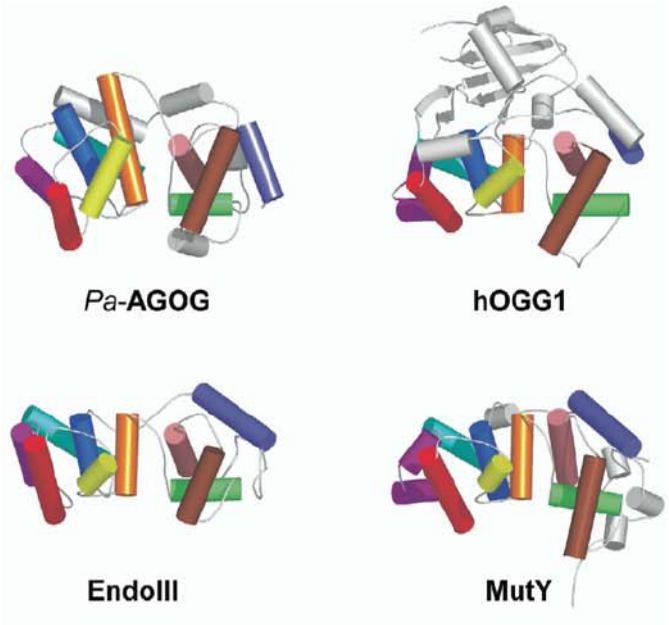
- *Ab initio/de novo* – very hard…

- 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**