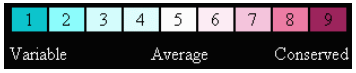
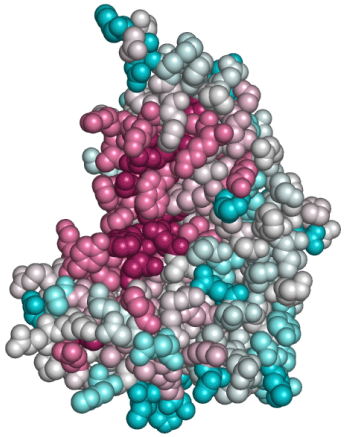
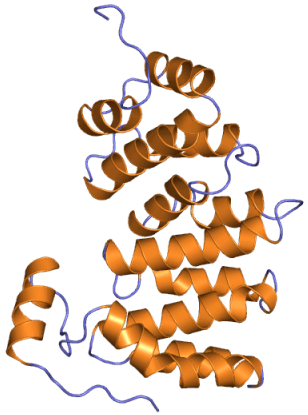
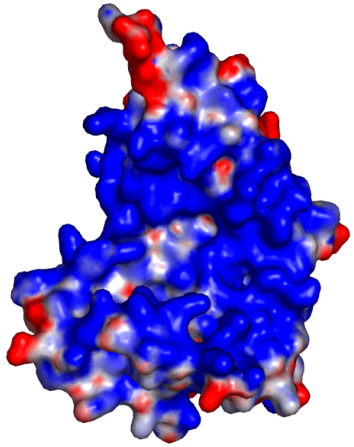
Bioinformatics for molecular biology

Structural bioinformatics tools, predictors, and 3D modeling







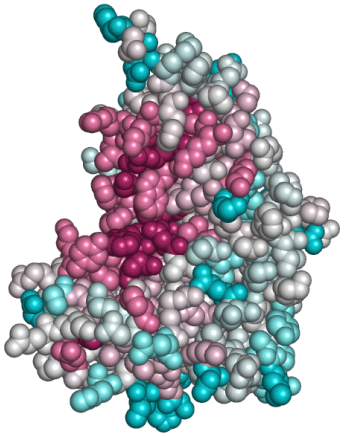


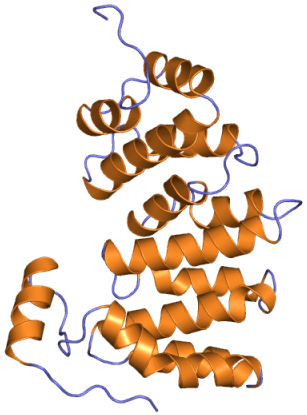
Jon K. Lærdahl, Research Scientist, Centre for Molecular Biology and Neuroscience (CMBN) and Institute of Medical Microbiology, Rikshospitalet
E-mail: jonkl@medisin.uio.no
Phone: 22844784
Main research area: Structural and Applied Bioinformatics

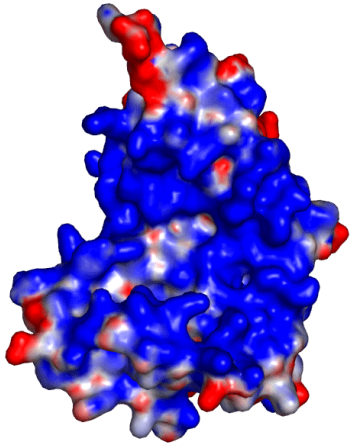
RIKSHOSPITALET

CENTRE FOR MOLECULAR BIOLOGY AND NEUROSCIENCE

Structural bioinformatics







In order to really understand what is going on in a biological system we need the 3D structures of the macromolecules

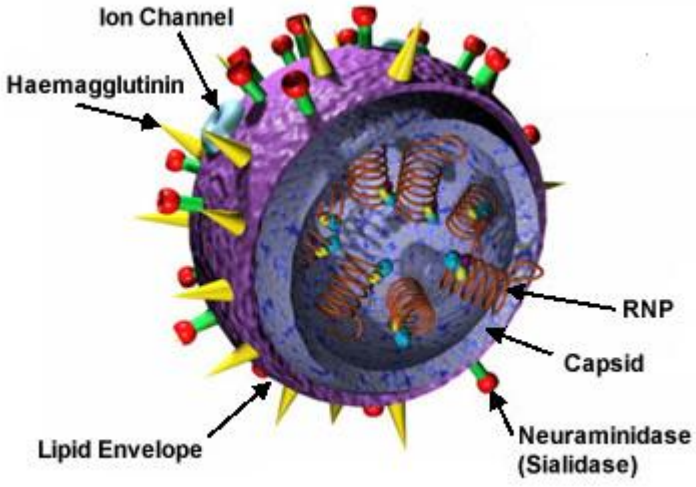
Very expensive (time and money) and difficult to determine 3D structure by experimental methods

Approx.
3 days

Structural bioinformatics:
- Determine 3D structure with computers
- Understand structure through computations
- Work with 3D structures, compare, classify, etc.

Plan

- Structure comparison and classification
- Predictors
- 3D structure modeling
  - Ab initio
  - Threading/fold recognition
  - Homology modeling
- Lunch
- Exercise
  - PyMOL
  - Homology modeling of influenza neuramidase (Tamiflu resistance?)



Protein domains

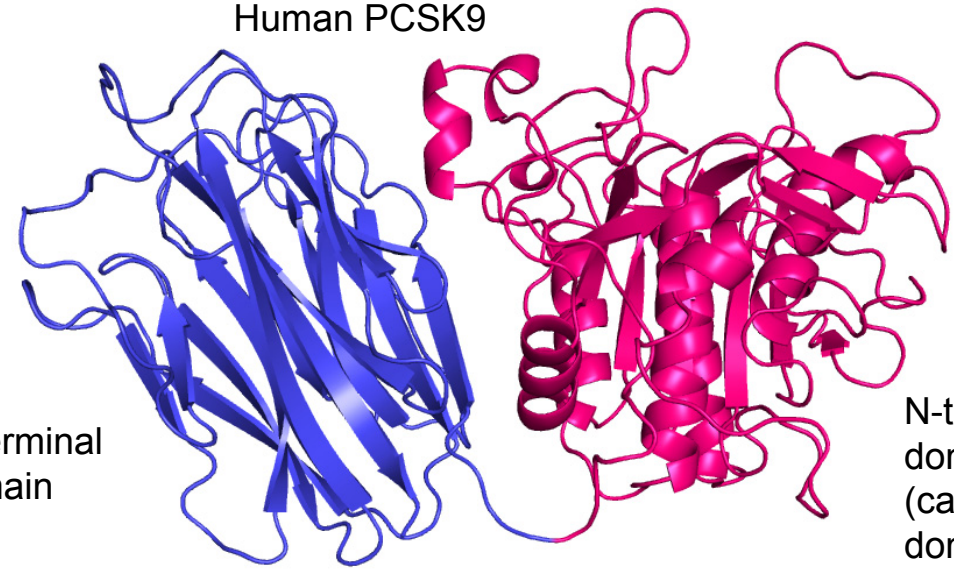
Jon K. Lærdahl,
Structural Bioinformatics

Domain: Compact part of a protein that represents a structurally independent region

Domains are often separate functional units that may be studied separately
- Structure
- Function

Domains can be "switched". They can be viewed as building blocks for building new proteins through evolution.

Human PCSK9



Difficult to detect boundaries between domains from sequence only

Easier to detect domain boundaries when you know the structure, but still no "right way to do it", *i.e.* a lot of subjectivity involved

Very often we model, compare, classify *domains* – not full-length proteins

Comparing structures

Jon K. Lærdahl,
Structural Bioinformatics

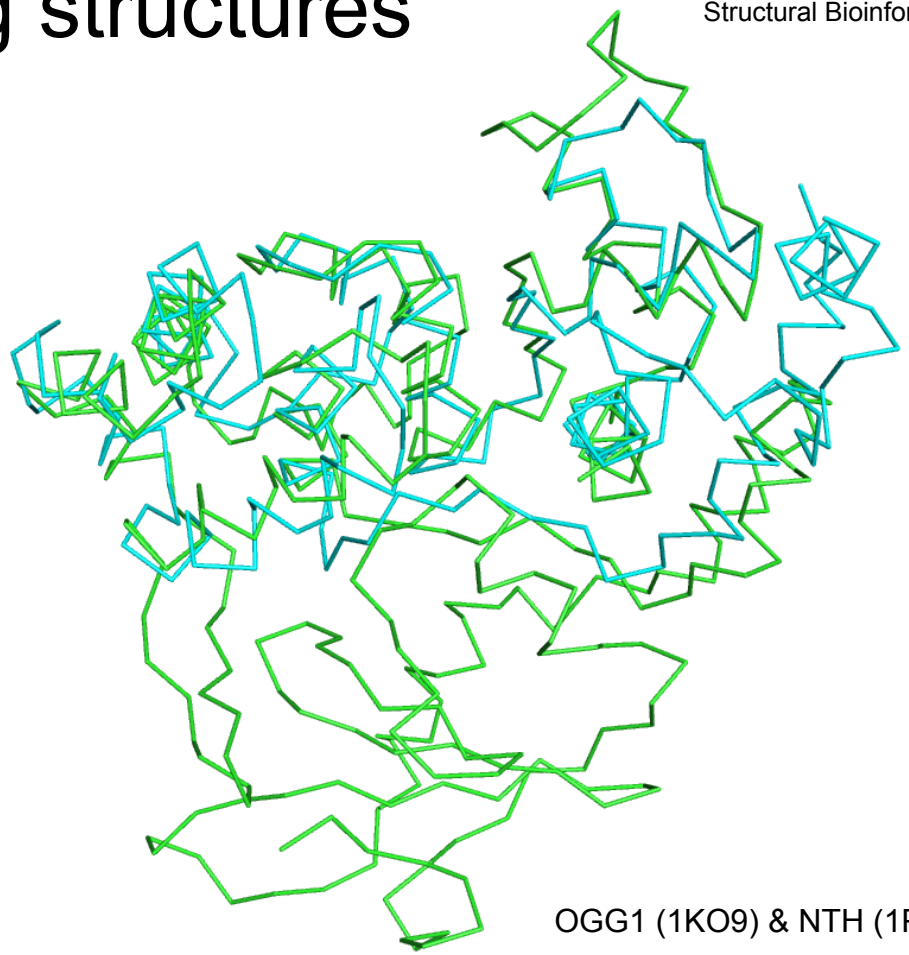
Superimposition/Alignment of 3D structures in space

The structures are superimposed in order to get corresponding main chain atoms as closely together as possible

If identical sequences – align all atoms

Non-identical sequences – align back-bone atoms only (usually *only* aligning Cα atoms!)

Structure is more conserved than sequence. **A structural alignment can therefore be used to define the ”correct” sequence alignment**



OGG1 (1KO9) & NTH (1P59)

Above, the two domains of NTH (blue) aligns nicely with the two C-terminal domains of OGG1 (green). The remaining domain of OGG1 is missing in NTH

Comparing structures

Jon K. Lærdahl,
Structural Bioinformatics

Root mean square deviation (RMSD) = square root of averaged sum of the squared differences of atomic (usually Cα) distances

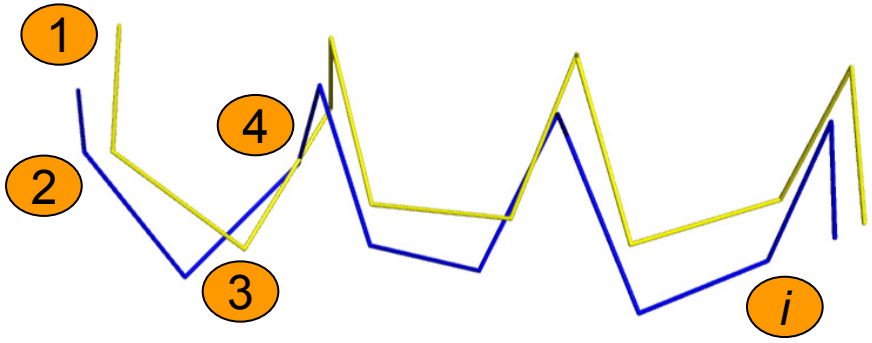
Calculate RMSD by:
Loop over equivalent positions *i*
- Get coordinates for both Cαs
- Calculate distance between Cαs, δ_i
- Square δ_i and add to sum

End loop
Divide sum by number of pairs, *N*, and take square root

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

RMSD tells you how similar two structures are

RMSD of ~0.5 Å or less for ”identical” structures



Comparing structures

Jon K. Lærdahl,
Structural Bioinformatics

Comparison of protein structures

Human NEIL1

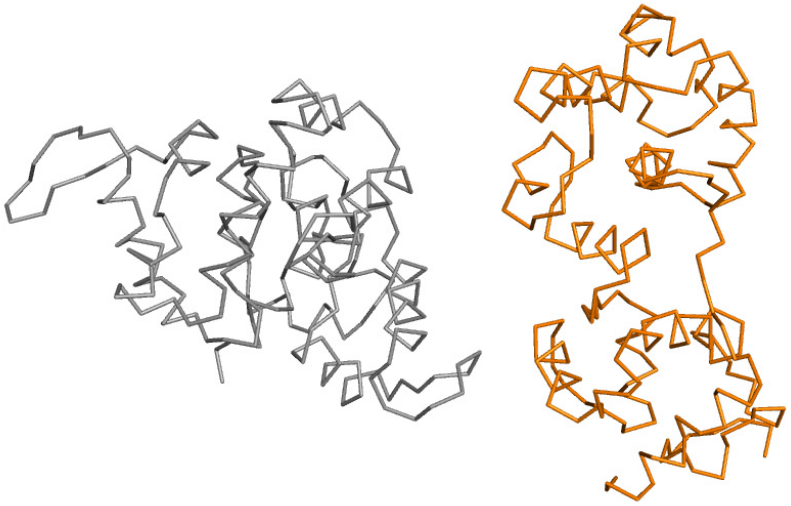
E. coli endonuclease VIII

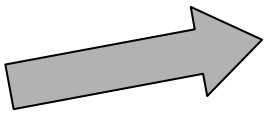
Aligned with RMSD = 1.41 Å

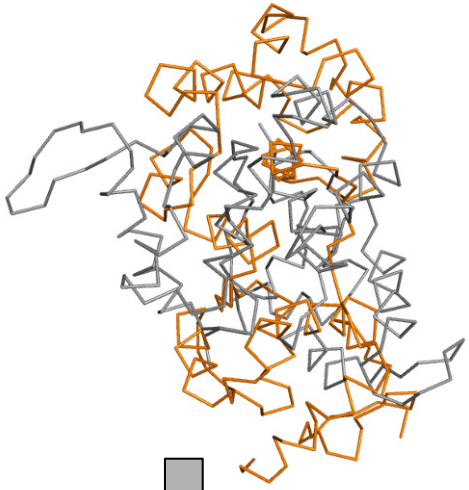
$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

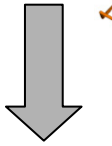
Comparing structures - Intermolecular method

Jon K. Lærdahl,
Structural Bioinformatics









- Two fairly similar structures
- Align *sequences* to find equivalent positions
- Do translation of one structure onto other structure
- Rotate one structure in space in order to minimize RMSD for aligned residues (Usually Cα atoms only)

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

Comparing structures - Intermolecular method

Jon K. Lærdahl,
Structural Bioinformatics

Problems with Intermolecular method:

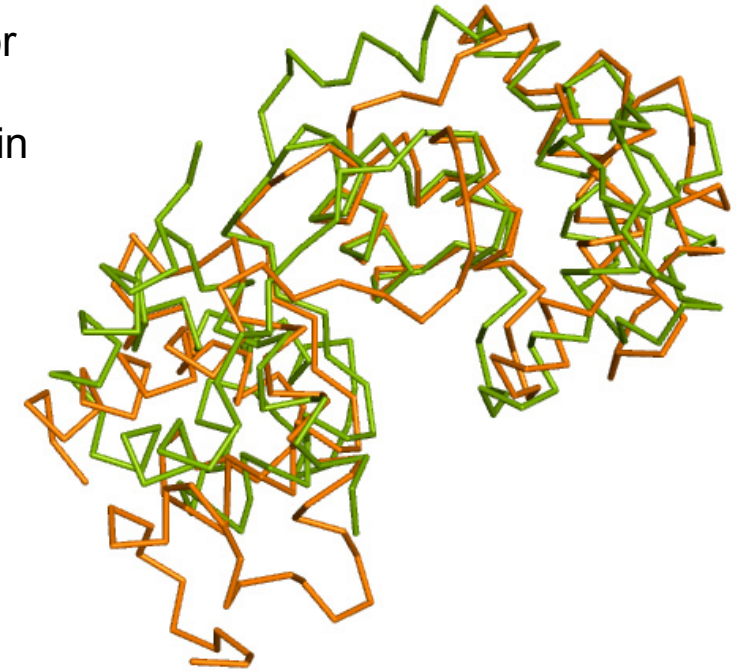
- RMSD depends on protein size
- Tricky to identify “equivalent residues” in the beginning
- Usually means that a sequence alignment is done first
    - Aligned residues are considered “equivalent”
    - Means the method is only useful for sequences that can be aligned by sequence comparison
- Several solutions suggested, but may give strange and non-optimal solutions
    - **Important to check alignments visually!**
- Iterative optimization:
    - First detect (often small) segments that can be aligned based on sequence
    - Do 3D superimposition based on residues in these segments
    - Based on 3D alignment, identify more residues that are close together and that are at “equivalent positions”. Use this larger set of pairs to do a new 3D superimposition.
    - Repeat until RMSD is converged

Important!

Comparing structures - Intramolecular method

Jon K. Lærdahl,
Structural Bioinformatics

- May be used for any two or more structures
- Does not depend on sequence similarity
- Does not necessarily generate physical superimposition
- Instead structural similarity measure based on internal structural statistic for each protein chain
- Based on building and comparing distance matrices for the structures
    - For example matrix A of all Cα distances in protein A
    - matrix B for protein B
    - ”Align” matrices to get best overlap
- Used in the most popular structure comparison tools, for example DALI
- Used for example to find which protein in the PDB is most similar to a new structure



- Intermolecular method:
    - Similar structures
    - Gives physical superimposition
- Intramolecular method:
    - Can be used for any two or more structures

Combined methods

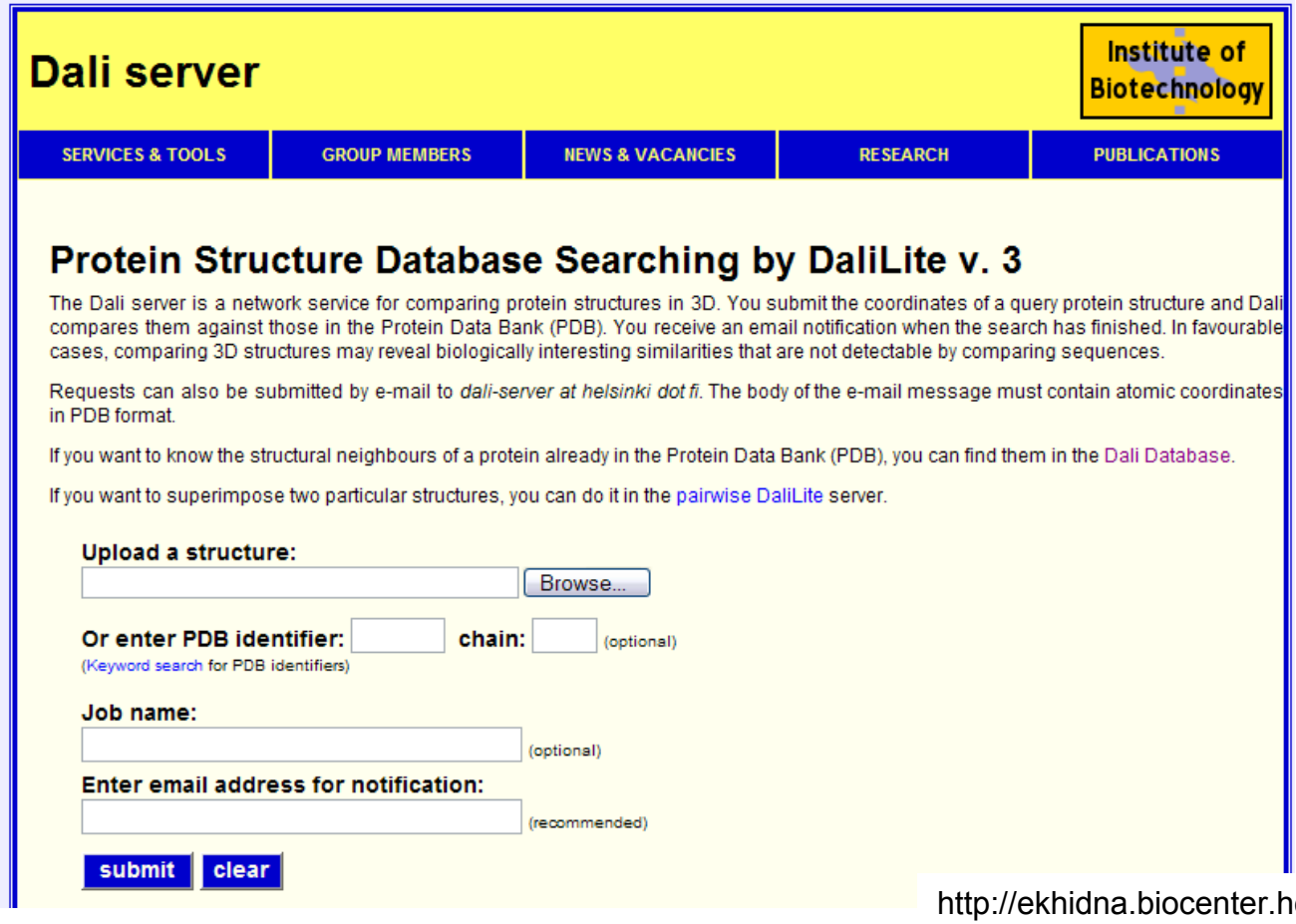
Comparing structures – Some tools:

Jon K. Lærdahl,
Structural Bioinformatics

STAMP (http://www.compbio.dundee.ac.uk/Software/Stamp/stamp.html): Unix program for iterative intermolecular alignment

Similar algorithms are often included in Viewers (*e.g.* DeepView & PyMOL)

DALI



Intramolecular method

Liisa Holm (Finland)

http://ekhidna.biocenter.helsinki.fi/dali_server

Jon K. Lærdahl,
Structural Bioinformatics

Comparing structures – Some tools:

Dali

(http://ekhidna.biocenter.helsinki.fi/dali_server)

Results

- Parseable data
- Matches to PDB90

The match list is truncated at 500 hits.

Query: 1ebmA

MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | Expand gaps | 3D Superimposition (Jmol Applet) | Reset Selection

Summary

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	1ebm-A	99.9	0.0	0	314	0	PDB	MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
2:	1n3a-A	51.9	0.0	0	314	0	PDB	MOLECULE: N-GLYCOSYLASE/DNA LYASE;
3:	1m3q-A	51.8	0.0	0	314	0	PDB	MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
4:	1lww-A	51.8	0.0	0	314	0	PDB	MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
13:	2jhj-B	24.3	0.0	0	291	0	PDB	MOLECULE: 3-METHYLADENINE DNA-GLYCOSYLASE;
14:	2jhn-A	23.9	0.0	0	293	0	PDB	MOLECULE: 3-METHYLADENINE DNA-GLYCOSYLASE;
15:	3d4v-D	22.9	0.0	0	281	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
16:	3d4v-B	22.8	0.0	0	282	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
17:	3cwu-D	22.8	0.0	0	282	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
18:	3cwt-D	22.8	0.0	0	281	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
19:	3cvs-D	22.7	0.0	0	282	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
38:	1pu8-A	14.7	0.0	0	215	0	PDB	MOLECULE: 3-METHYLADENINE DNA GLYCOSYLASE;
39:	2h56-B	14.1	0.0	0	217	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSIDASE;
40:	2abk	14.0	0.0	0	211	0	PDB	MOLECULE: ENDONUCLEASE III;
41:	1rrt-A	14.0	0.0	0	346	0	PDB	MOLECULE: MUTY;
42:	1vrl-A	14.0	0.0	0	346	0	PDB	MOLECULE: 5'-D(*AP*AP*GP*AP*CP*(8OG)P*TP*GP*GP*AP
43:	1kea-A	13.5	0.0	0	217	0	PDB	MOLECULE: POSSIBLE G-T MISMATCHES REPAIR ENZYME;
44:	1kg6-A	13.0	0.0	0	224	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
45:	1kg5-A	12.9	0.0	0	225	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
46:	1kg3-A	12.8	0.0	0	224	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
47:	1kqj-A	12.7	0.0	0	225	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
48:	1muy-A	12.6	0.0	0	225	0	PDB	MOLECULE: ADENINE GLYCOSYLASE;

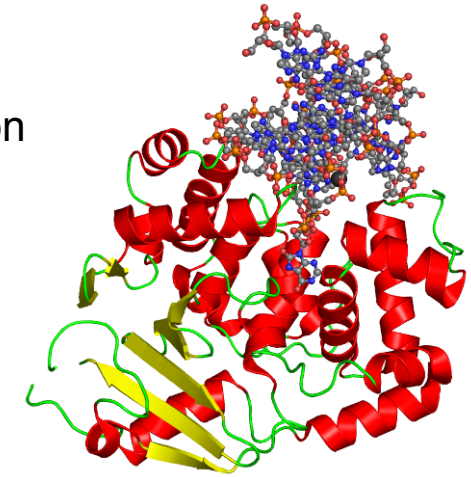
- Compare 2 structures
- Compare multiple structures
- Search a database of structures for the most similar structures with a pdb-file query
- Search database with PDB id query
- Z-score > 4 usually indicates significant level of similarity

Alternatives:
VAST (at NCBI)
CE
SSAP
More…

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure evolution

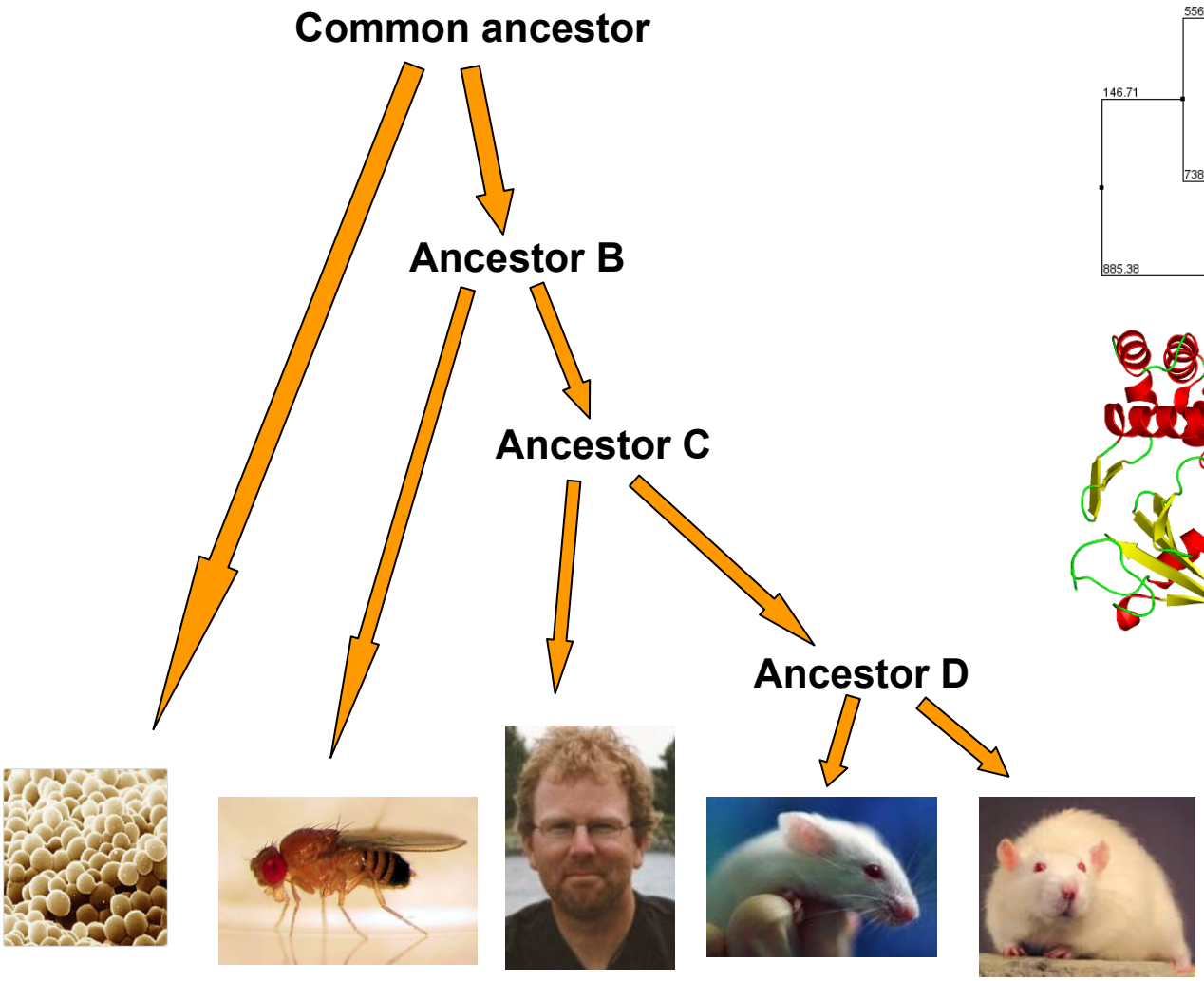
- The origin of this gene/protein is (very likely) before the last common ancestor of *S. cerevisiae* (yeast), human, mouse, rat, and fruit fly
- Some of the amino acids have not mutated in ~1 billion years
- Neutral mutation rate in mammals is ~0.01 base pair/5 million yr

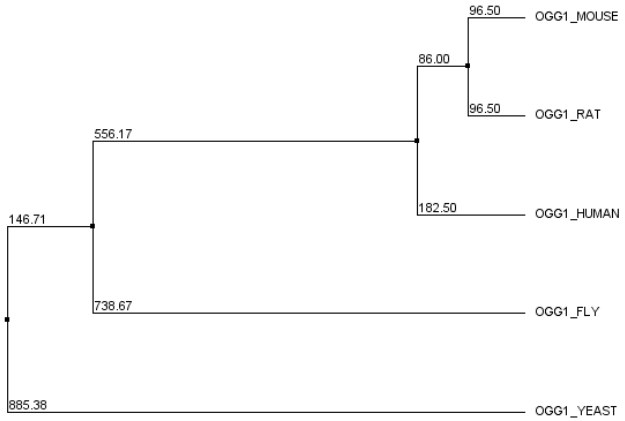


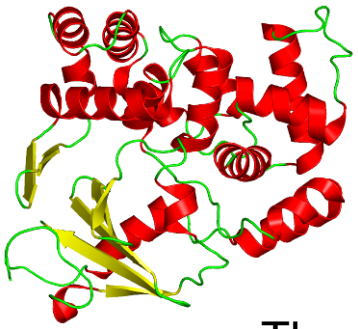
```
OGG1_YEAST/1-376     1 -MSYK------------------FGKLAINKSELCLANVLQAGQSFRW---------IWDEKLNQYSTTMKIGQQEKYSVVILRQDEENEILEFVAVGDCGNQ 75
OGG1_MOUSE/1-345     1 -MLFRSWLPSSMRHRTLSSSPALWASIPCPRSELRLDLVLASGQSFRWK---EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDDSQVSRPTLEEL------ 93
OGG1_RAT/1-345       1 -MLFSSSLSSSMRHRTLTSSPALWASIPCPRSELRLDLVLASGQSFRWR---EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDKGQVGRPTLEEL------ 93
OGG1_HUMAN/1-345     1 -MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWR---EQSPAHWSGVLADQWWTLTQTEEQLHCTVYRGDKSQASRPTPDEL------ 93
OGG1_FLY/1-343       1 MLAHNLGFHKKRLFSNMKAVLQDRGVIGLSLEECDLERTLLGGQSFRWRSICDGNRTKYGGVVFNTYWVLQQEESFITYEAY-GTSSPLATKDYSSL------ 96

OGG1_YEAST/1-376    76 DALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFAKLSP--QGIRILAQEPWETLISFICSSNNNISRITRMCNSLCSNFGNLITTIDGVAYHSFPTS---EELT 173
OGG1_MOUSE/1-345    94 ----ETLHKYFQLDVSLAQLYSH-WASVDSHFQRVAQKFQGVRLLRQDPTECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPNL---HALA 188
OGG1_RAT/1-345      94 ----ETLHKYFQLDVSLTQLYSH-WASVDSHFQSVAQKFQGVRLLRQDPTECLFSFICSSNNNIARITGMVERLCQAFGPRLVQLDDVTYHGFPNL---HALA 188
OGG1_HUMAN/1-345    94 ----EAVRKYFQLDVTLAQLYHH-WGSVDSHFQEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSL---QALA 188
OGG1_FLY/1-343      97 ------ISDYLRVDFDLKVNQKD-WLSKDDNFVKFLS--KPVRLLSQEPFENIFSFLCSQNNNIKRISSMIEWFCATFGTKIGHFNGADAYTFPTINRFHDIP 190

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

OGG1_YEAST/1-376   277 -KNHLKELRTKYNALPISRKKINLELDHIRLMLFKKWGSYAGWAQGVLFSKEIGGTSGSTTTGTIKKRKWDMIKETEAIVTKQMKLKVELSDLHIKEAKID  376
OGG1_MOUSE/1-345   285 -TSQAKGPS------PLANKELG---NFFRNL----WGPYAGWAQAVLFSADLRQPS-LSREPPAKRK----------------------KGSKRPEG---  345
OGG1_RAT/1-345     285 -TSQTKGPS------PLANKELG---NFFRNL----WGPYAGWAQAVLFSADLRQQN-LSREPPAKRK----------------------KGSKKTEG---  345
OGG1_HUMAN/1-345   285 -TSQAKGPS------PQTNKELG---NFFRSL----WGPYAGWAQAVLFSADLRQSR-HAQEPPAKRR----------------------KGSKGPEG---  345
OGG1_FLY/1-343     286 GQKNVT---------KKIYEEVS---KHFQKL----HGKYAGWAQAILFSADLSQFQ-NTSTVACKKK----------------------SNKKPKK---  343
```

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure evolution







The overall structure of this protein is the same in all these organisms – i.e. many mutations does not change the structure and/or function

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure evolution

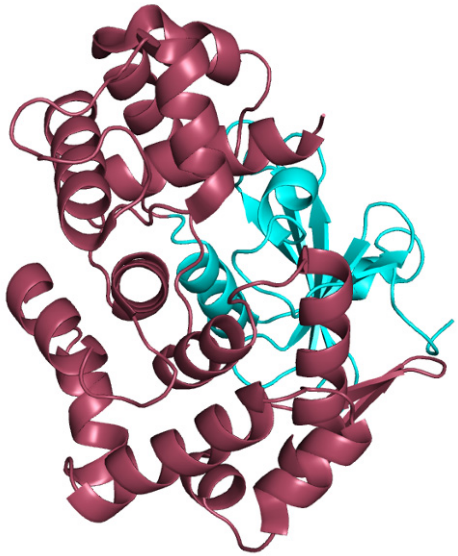
Proteins that fold in the same way, i.e. ”have the same fold” are often homologs.

Structure evolves slower than sequence

Sequence is less conserved than structure

EndoIII (2ABK):

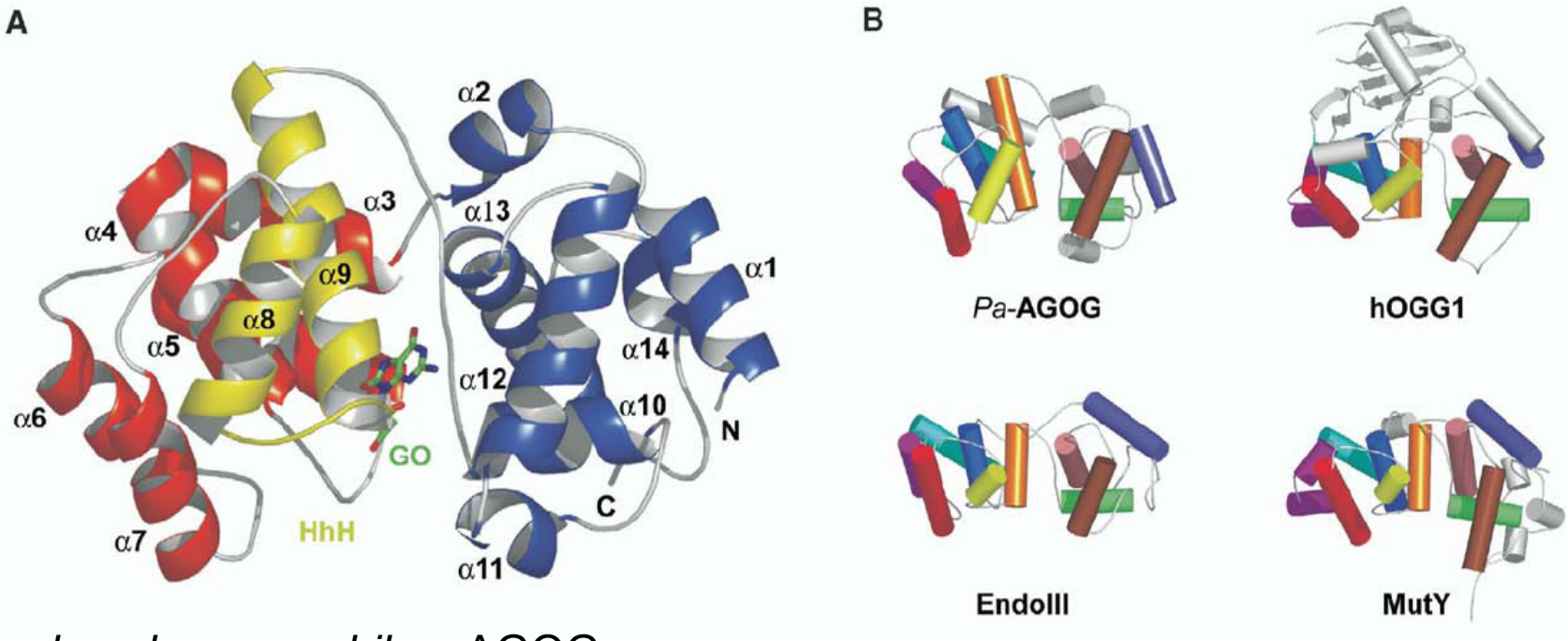
OGG1 (1EBM):



```
OGG1/1-345     1 MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLADQVWTLTQ 68
NTH1/1-312     1 MCS----PQESGMTALSARMLTRSRSLGPGAGPRGCREEPG--PLRRREAAAE--------------- 47

OGG1/1-345    69 TEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQGVRLLRQD 136
NTH1/1-312    48 -ARKSHSPVKRPRKAQRLRVAYEGSDSEKG----EGAEPLKVPVWEPQD--WQQQLVNIRAMRNKKDA 108

OGG1/1-345   137 PIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLR----KLG 200
NTH1/1-312   109 PVDHLGTEHCYDSSAPPKVRRYQVLLSLMLSSQ--TKDQVTAGAMQRLRA-RGLTVDSILQTDDATLG 173

OGG1/1-345   201 -LGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICLMALDKPQAVPV 267
NTH1/1-312   174 KLIYPVGFWRSKVKYIKQTS---AILQQHYGGDIPASVAELVALPGVGPKMAHLAMAVAWGTVSGIAV 238

OGG1/1-345   268 DVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELGNFF-RSLW----GPYAGWAQAVLFSADLRQSRHAQ 330
NTH1/1-312   239 DTHVHRIANR-LRW---TKKATKSPEETRAALEEWLPRELWHEINGLLVGFGQQTCLPVHPRCHACLN 302

OGG1/1-345   331 EPPAKRRKGSKGPEG                                                     345
NTH1/1-312   303 QALCPAAQGL-----                                                     312
```

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure evolution



Pyrobaculum aerophilum AGOG

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

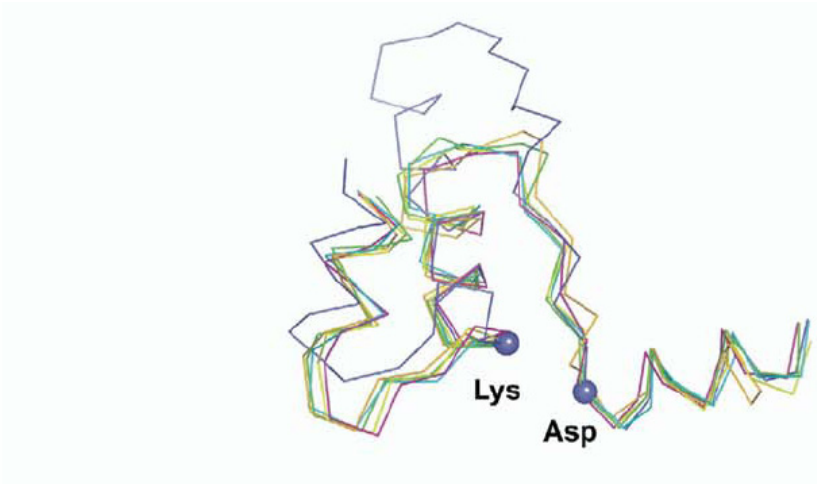
Hardly any detectable sequence similarity to human OGG1, and *E. coli* EndoIII and MutY

Still clearly the same protein fold (overall structure)

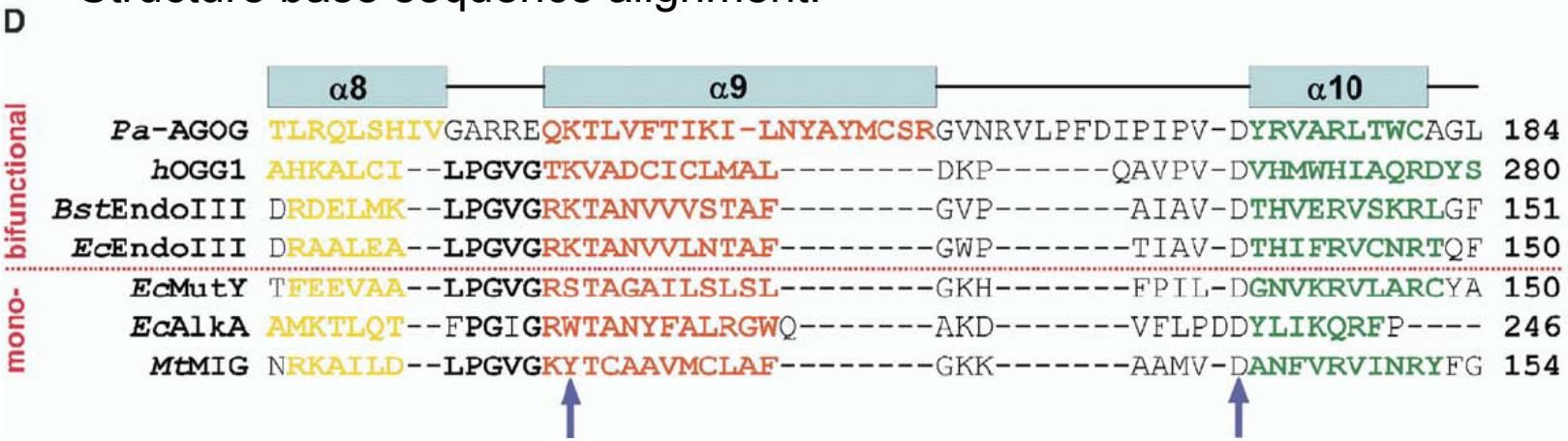
Evolution has “eroded away” sequence similarity but left the structure intact

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics



Structure base 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

Last common ancestor (2 Gyrs?)

Protein structure evolution

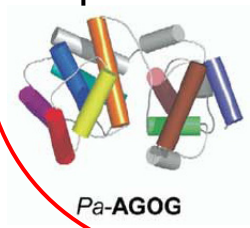
Jon K. Lærdahl,
Structural Bioinformatics

Similar structure
No sequence similarity

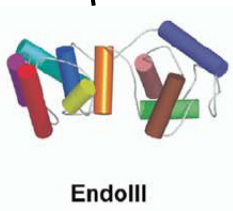
Fairly similar structure
Some sequence similarity

Fairly similar structure
Some sequence similarity

Very similar structure
Significant sequence similarity

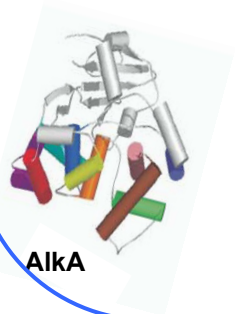


Pa-AGOG

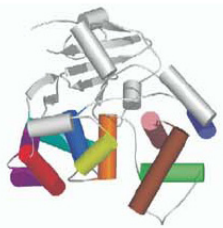


EndoIII

MutY



AlkA



Yeast Ogg1

Human Ogg1

Mouse Ogg1

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure alignments

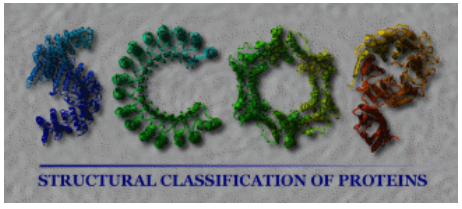
Proteins that fold in the same way, i.e. ”have the same fold” are often homologs.
Structure evolves slower than sequence
Sequence is less conserved than structure

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

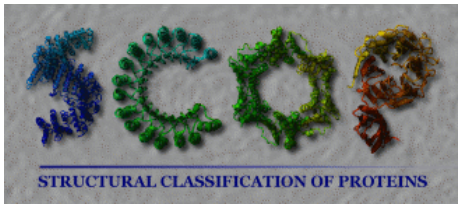
Use structure comparisons to classify, group and cluster proteins. Build protein structure families and hierarchies



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

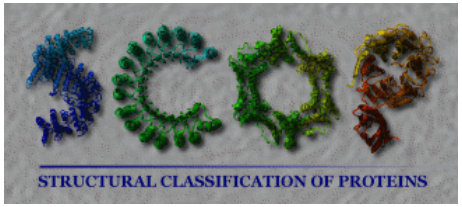
Scop Classification Statistics

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)

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

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

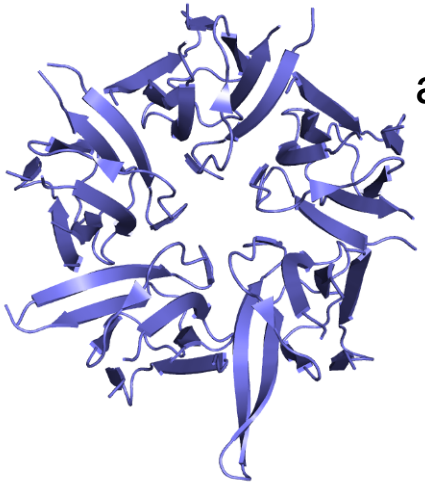
- Classes
  - Upper level of classification (4 major, 3 minor)
  - Contains folds
  - Based on secondary structure composition and “general features”
  - *e.g.* all-α, all-β, ”membrane and cell surface” and ”small proteins”
  - α/β: One β-sheet with strands connected by single α-helices
  - α+β: α-helical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

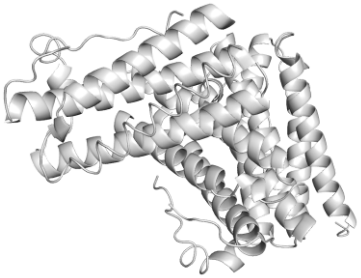
http://scop.mrc-lmb.cam.ac.uk/scop

Jon K. Lærdahl,
Structural Bioinformatics

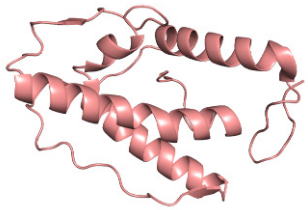
SCOP



all-β class



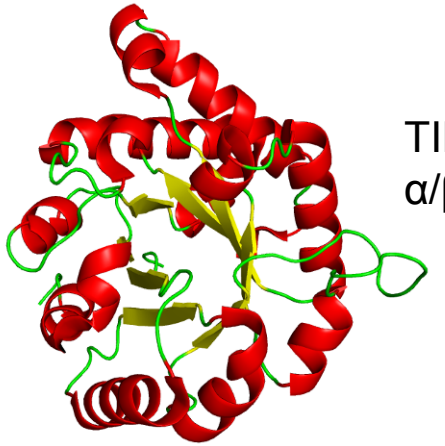
4-helical cytokines



T4 endonuclease V

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

Jon K. Lærdahl,
Structural Bioinformatics

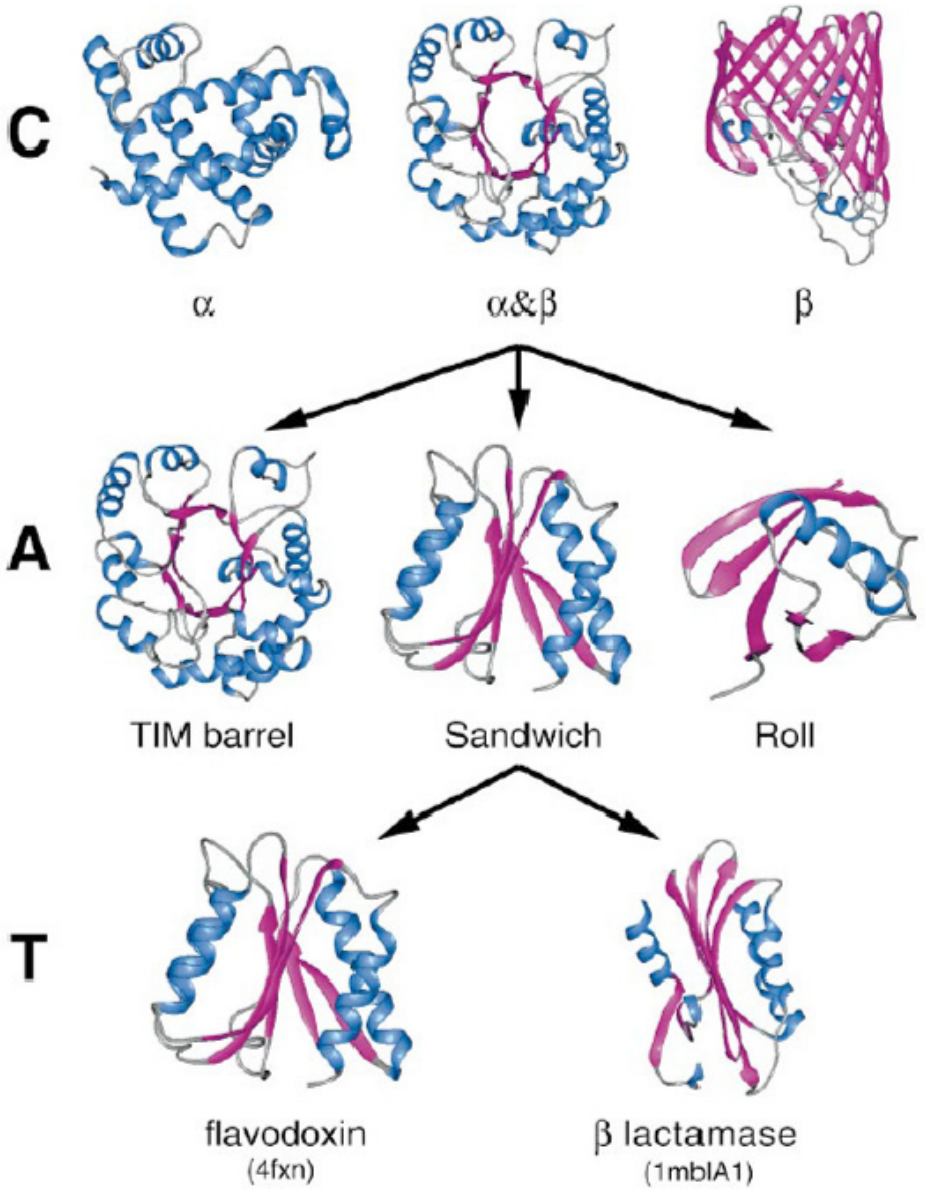
CATH

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)



http://www.cathdb.info

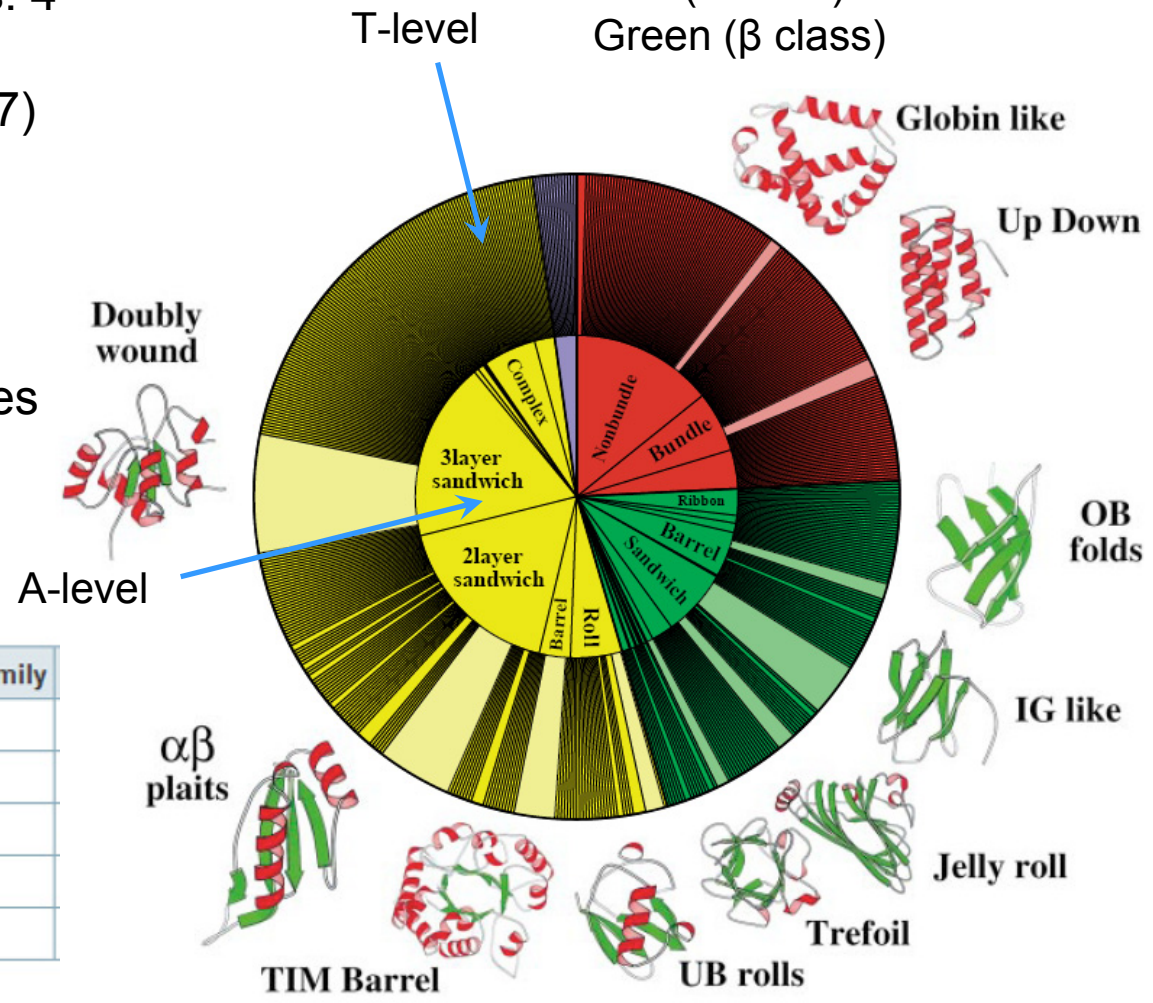
C.A. Orengo *et al. Structure* **5**, 1093 (1997)

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!

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



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)

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!

Here: *secondary structure prediction*, but similar or related methods/algorithms are used in most predictors

Jon K. Lærdahl,
Structural Bioinformatics

Secondary structure prediction

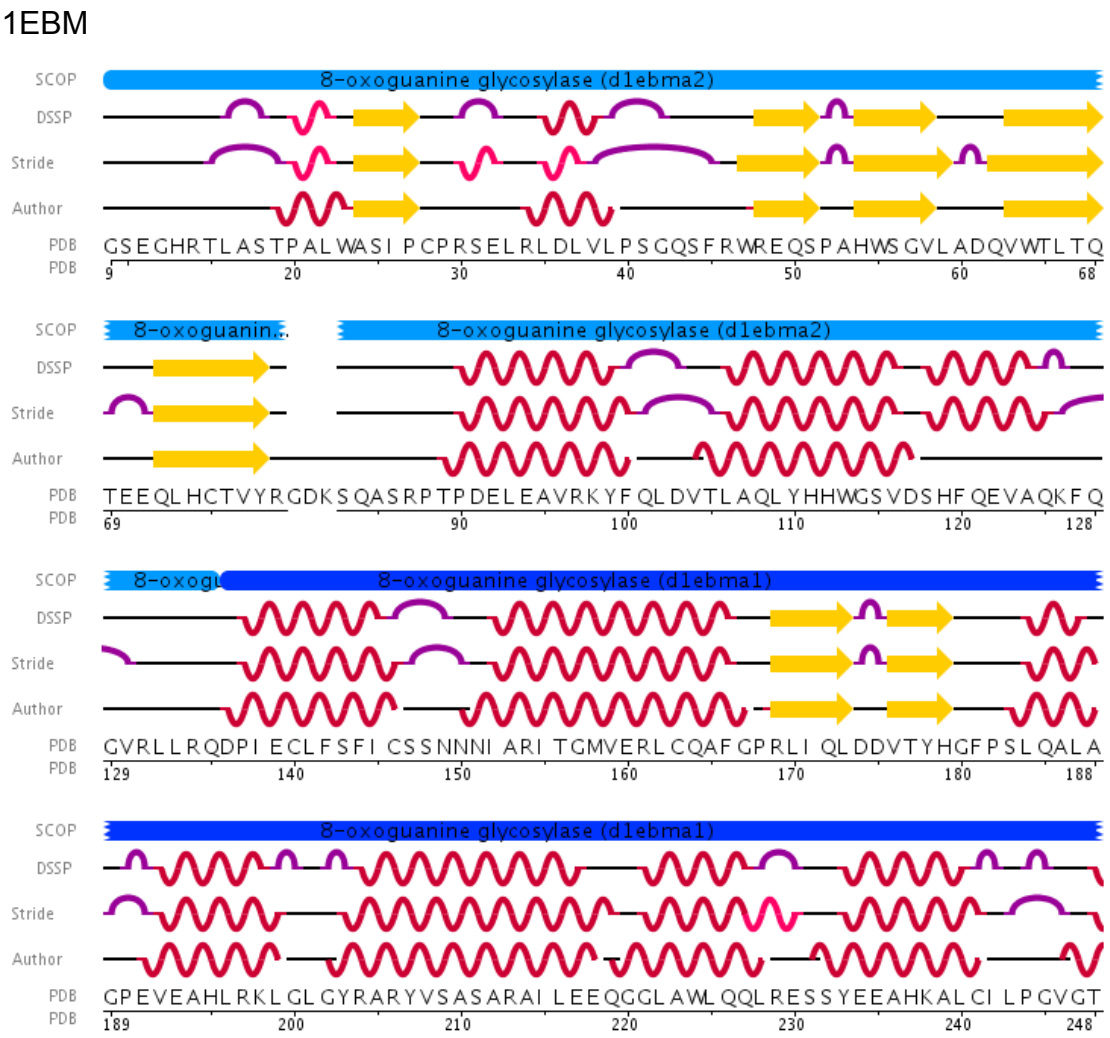
Important: Assigning secondary structure is *not trivial* and there is *no single consensus method* even when 3D structure is known

- 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

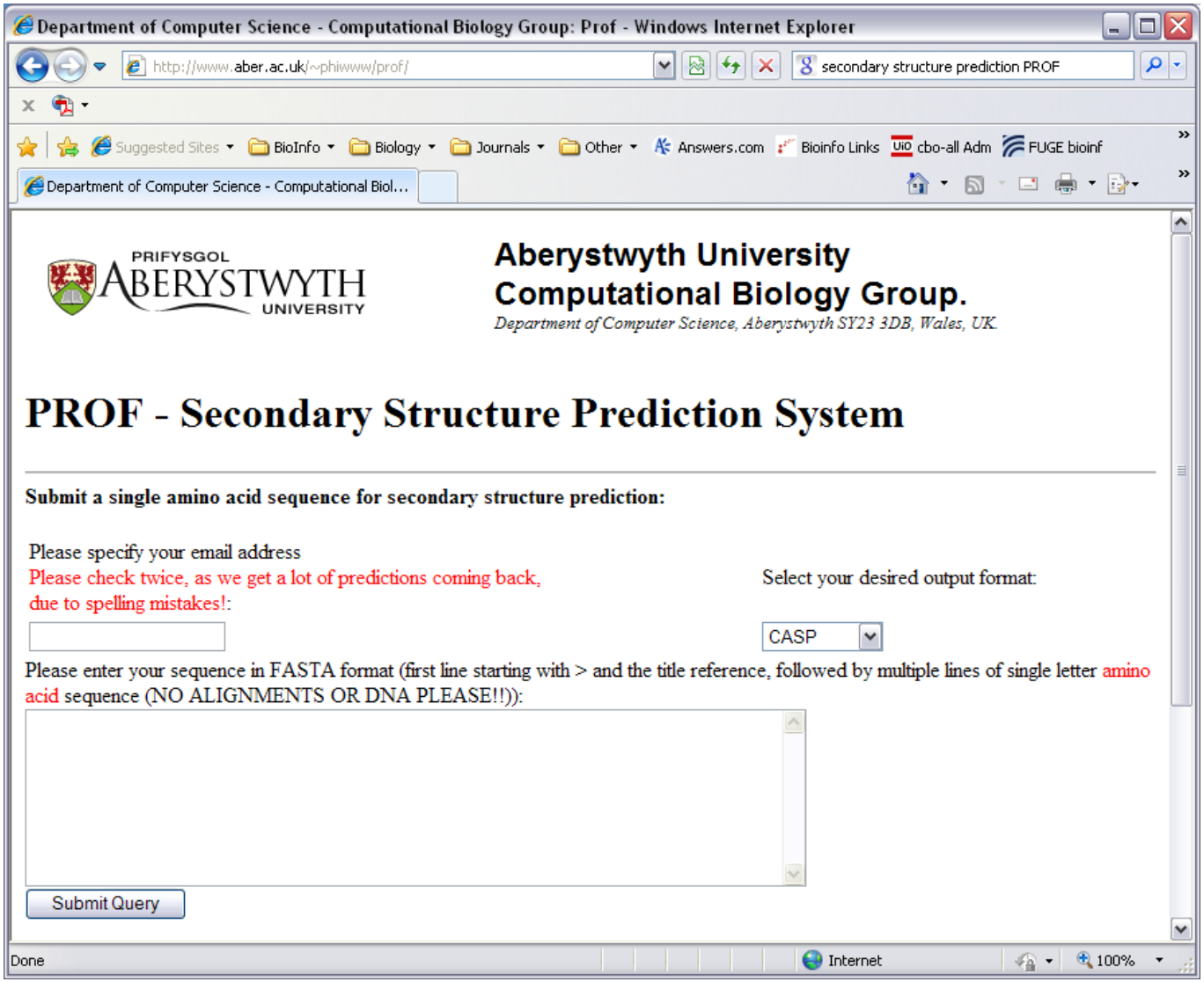
Everything else loop/coil



Jon K. Lærdahl,
Structural Bioinformatics

Secondary structure prediction

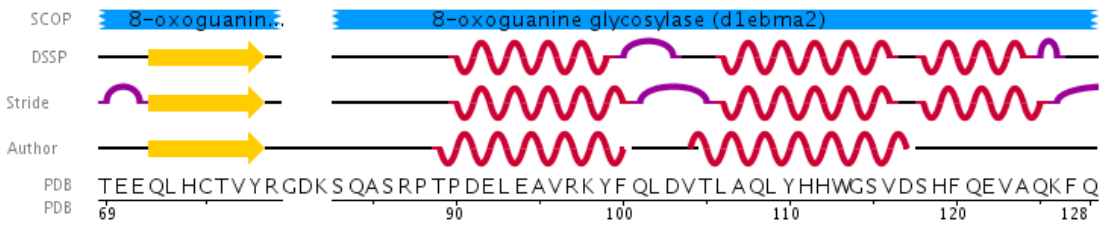
Tools/programs that accept a primary sequence and predicts the secondary structure state (H/helix, E/sheet, or C/Loop&Coil) for each residue



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  CEEEEEEEEECCCCCCCCCCHHHHHHHHHHHHCCCCHHHHHHCCCCCCHHHHHHHHHHHCC
```

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

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

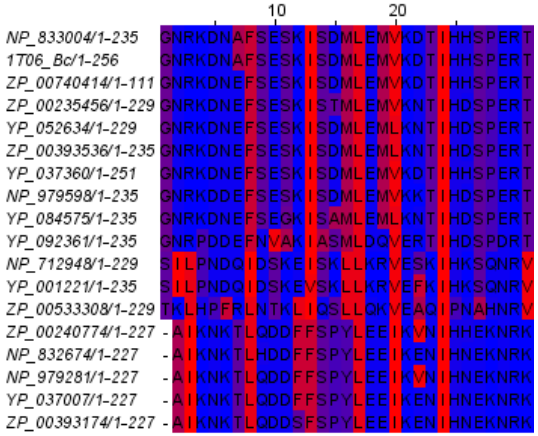
- 2nd generation prediction (until mid 1990's) ~60%
    - Proper inclusion of propensities for neighboring residues
    - Larger experimental data set

- 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:

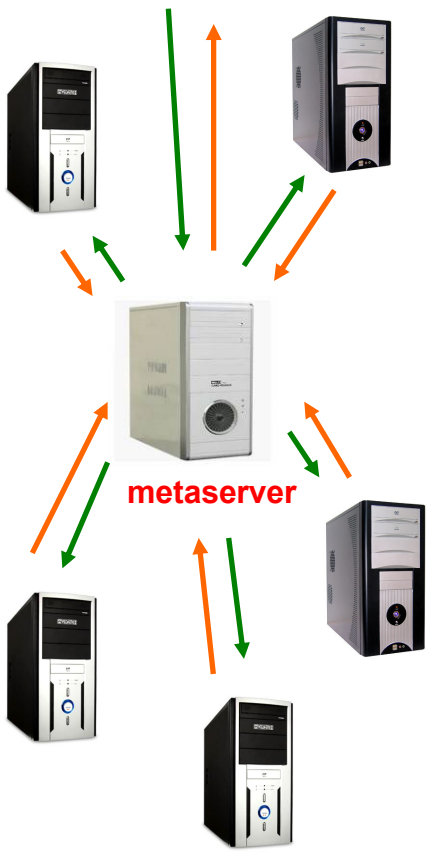
- **PHD** – BLASTP to find homologs, MSA of homologs, neural networks used for prediction, web server
- **PSIPRED** – PSI-BLAST for homologs, MSA generated, neural network prediction, filtering, web server
- **PROF** – PSI-BLAST, MSA, neural network

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

NB!

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)

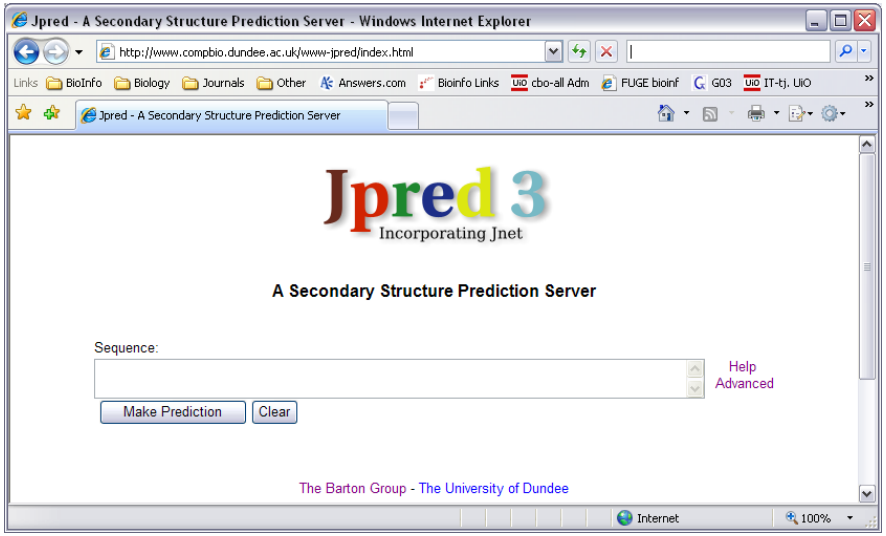


Secondary structure prediction — consensus-based

Jon K. Lærdahl,
Structural Bioinformatics

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

jalign          : ------------------H------EEEEE--------HHHHHHHHHH----H-HHHHHHHHHH--------HHHHHHHHHHHHH----- : jalign
jfreq           : ---------------HHHHHHHH---EEEE-----HHH-HHHHHHHHH-----HHHHHHHHHHH-------HHHHHHHHHHHHHHHH-- : 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------HHHHHHHHHHHHH----EE- : jpssm

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

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



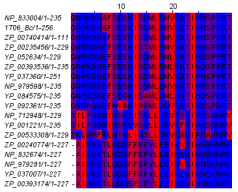
C. Cole *et al. Nucleic Acids Res.* ***36***, W197 (2008)

Predictors – *common features*

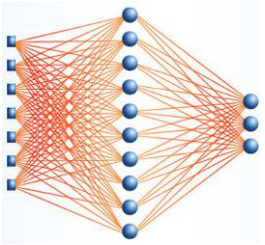
Jon K. Lærdahl,
Structural Bioinformatics

• 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



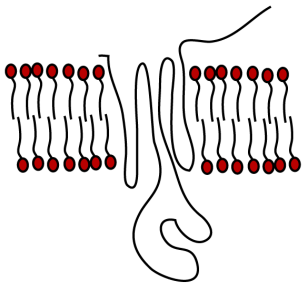
• Use neural networks or similar methods in predictions

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

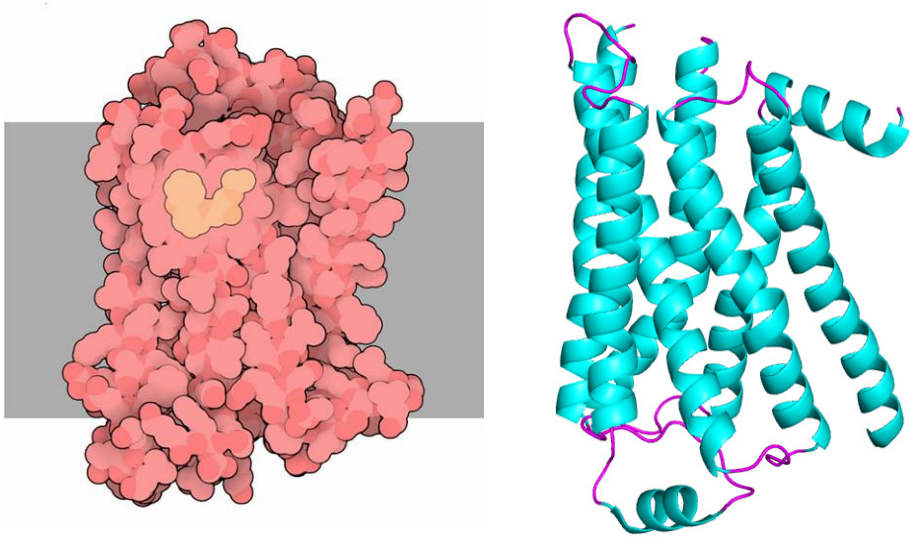
Jon K. Lærdahl,
Structural Bioinformatics

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

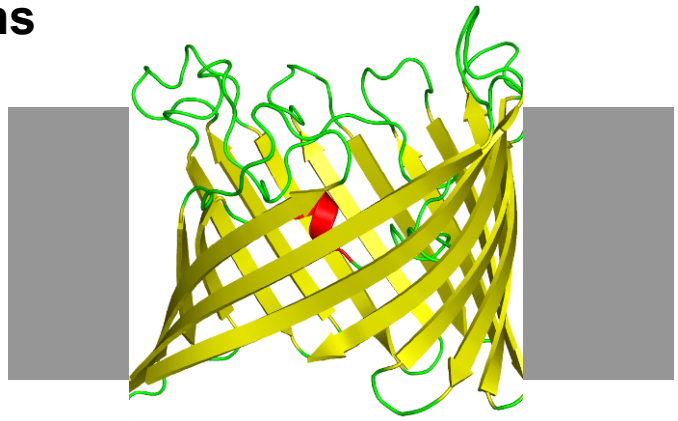


Can usually NOT use the same predictors for secondary structure and other properties as for globular proteins



PDB Apr. 08 “Molecule of the Month”

2RH1, Human adrenergic receptor



Porin

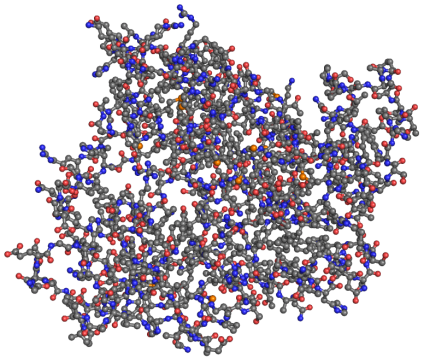
Jon K. Lærdahl,
Structural Bioinformatics

3D Structure Modeling

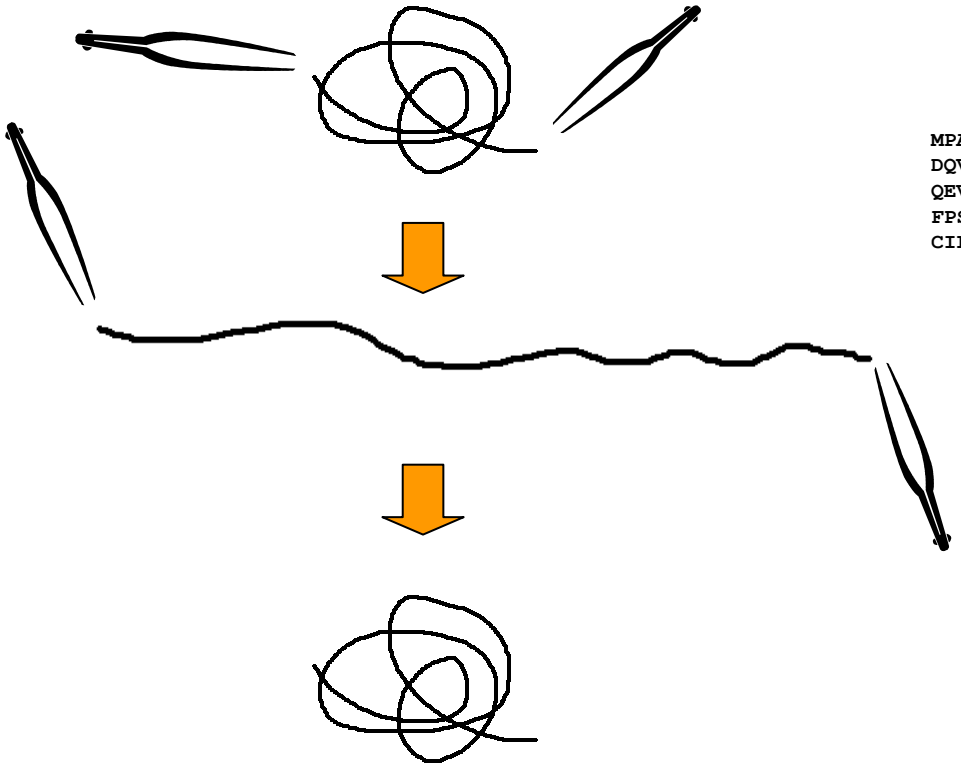
Modeling of 3D structure

Jon K. Lærdahl,
Structural Bioinformatics

- ~83,000,000 sequence records in the traditional GenBank divisions (Feb 2008)
  - Several orders of magnitude more sequences in other public databases
  - Next Generation Sequencing generates ~20 Gb in *a single run*
- ~55,000 3D structures in the PDB (*i.e.* all published structures)
  - (Structures: All data in one place!)
  - 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

Jon K. Lærdahl,
Structural Bioinformatics

Protein folding

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

Folding is spontaneous in the cell

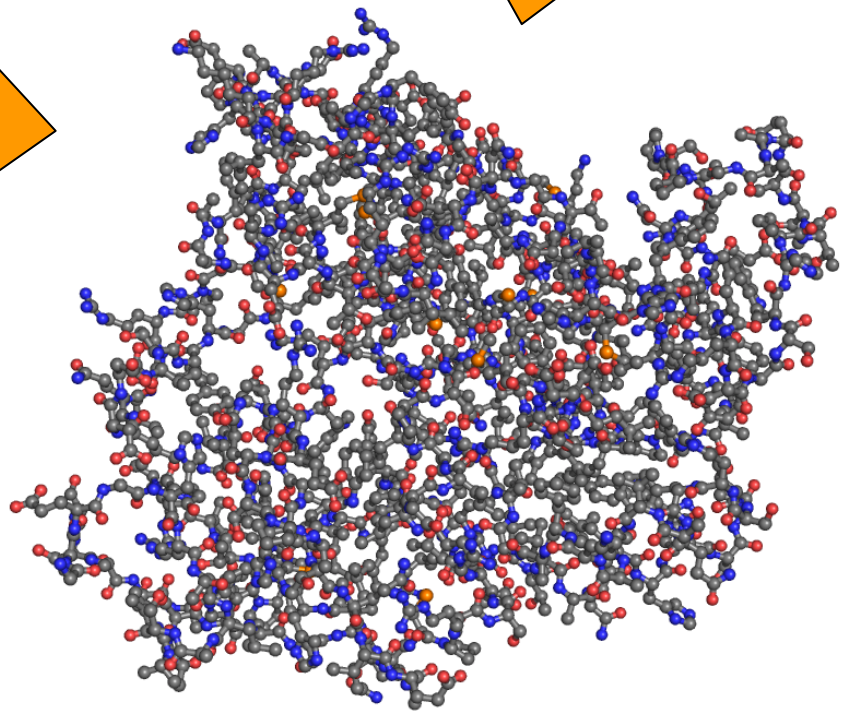
In the cell

The sequence determines the 3D structure!

In the computer

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

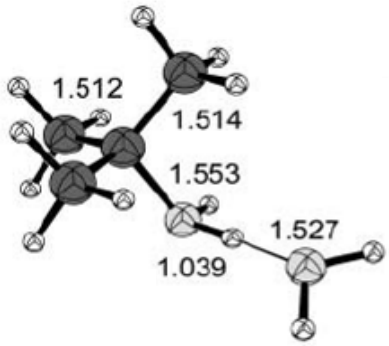
- Based on physical/chemical laws and not already published experimental structures

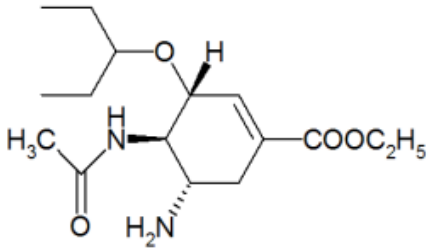


Jon K. Lærdahl,
Structural Bioinformatics

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?)

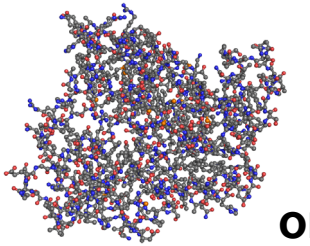




Ab initio structural prediction

Jon K. Lærdahl,
Structural Bioinformatics

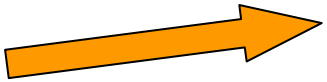
- Does *not* use prior knowledge of structure from the PDB
  - That is why they are known as *ab initio*
- Still, some programs known as *ab initio* protein modeling programs also use *some* information from the PDB, for example structures for small fragments
- At least in some respects based on the "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



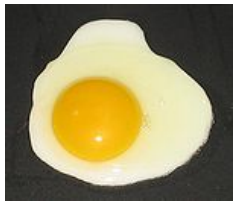
or

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

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSGQSFRWREQSPAHWSGVLA
DQVWTLTQTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGGLAWLQQLRESSYEEAHKAL
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSPQTNKELG
```



?



Structural bioinformatics

IBM BlueGene/L used for *ab initio* protein folding

TOP500 SUPERCOMPUTER SITES

PROJECT | LISTS | STATISTICS | RESOURCES | NEWS | CONTACT | SUBMISSIONS | LINKS | HOME

Home › Lists › June 2008

TOP500 List - June 2008 (1-100)

R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

Power data in KW for entire system

next

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2008 IBM	122400	1026.00	1375.78	2345.50
2	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution / 2007 IBM	212992	478.20	596.38	2329.60
3	Argonne National Laboratory United States	Blue Gene/P Solution / 2007 IBM	163840	450.30	557.06	1260.00
4	Texas Advanced Computing Center/Univ. of Texas United States	Ranger - SunBlade x6420, Opteron Quad 2Ghz, Infiniband / 2008 Sun Microsystems	62976	326.00	503.81	2000.00
5	DOE/Oak Ridge National Laboratory United States	Jaguar - Cray XT4 QuadCore 2.1 GHz / 2008 Cray Inc.	30976	205.00	260.20	1580.71
6	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2007 IBM	65536	180.00	222.82	504.00
7	New Mexico Computing Applications Center (NMCAC) United States	Encanto - SGI Altix ICE 8200, Xeon quad core 3.0 GHz / 2007 SGI	14336	133.20	172.03	861.63
8	Computational Research Laboratories, TATA SONS India	EKA - Cluster Platform 3000 BL460c, Xeon 53xx 3GHz, Infiniband / 2008 Hewlett-Packard	14384	132.80	172.61	786.00
9	IDRIS France	Blue Gene/P Solution / 2008 IBM	40960	112.50	139.26	315.00

Statistics

Top500 List: 06/2008

Statistics Type: Vendors

Generate

Charts

Top500 List: 06/2008

Chart Type: Vendors

Generate

Development

Advance Online Publication|doi:10.1038/nature05990|Published online 14 October 2007

nature

NEWS & VIEWS

COMPUTATIONAL BIOLOGY

Protein predictions

Eleanor J. Dodson

Predicting the three-dimensional structure of a protein from its amino-acid sequence is a dauntingly complex task. But with colossal computer power and knowledge of other structures, it can be done.

de novo/ab initio methods

Fifty years have passed since the Nobel-prizewinning discovery that the amino-acid sequence of a protein determines its three-dimensional structure[1] — yet computational biologists are still unable to predict the shape of a protein from its sequence. Given that there are many more protein sequences available than structures, and that protein shape is crucial for understanding cellular and physiological processes, a method for predicting such structures is vital. The paper by Qian *et al.*[2], which appears on *Nature*'s website today, and in which the structure of a protein containing 112 amino acids is accurately predicted, thus represents a real breakthrough. The authors' model was sufficiently accurate to act as the starting point in the X-ray structure determination of the protein.

Most structural information on proteins is derived from X-ray and nuclear magnetic resonance (NMR) experiments. These have revealed the general characteristics of proteins — for example, sequence motifs that form secondary structural elements such as helices and sheets. Such elements are organized to generate the overall protein architecture, mainly as a result of internal interactions between hydrophobic amino-acid side chains buried within the structure.

The shape of a protein corresponds to the lowest-energy conformation of that mol-

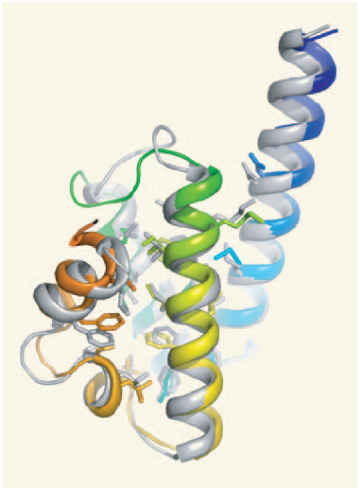


Figure 1 | Model test. Qian *et al.*[2] have developed a computational method for predicting the three-dimensional structure of a protein from its amino-acid sequence. Here, their predicted structure (grey) of a protein is overlaid with the experimentally determined crystal structure (shown in colour) of that protein. The agreement between the two is excellent, with the amino-acid side chains overlapping particularly well.

The main goal of the CASP network is "to obtain an in-depth and objective assessment of our current abilities and inabilities in the area of protein structure prediction"[4]. Every two years the organizers provide the amino-acid sequences of a set of proteins for which undisclosed crystal structures exist. Modellers are challenged to predict structures for the proteins, and these are then assessed against the crystallographic results. The assessors use various scoring systems, but the most rigorous test is the one used by Qian *et al.*[2] to test their own models — can the prediction be used in 'molecular replacement' searches[5] that allow the raw data from X-ray diffraction studies to be related to the structure of the compound being investigated? Normally, a previously determined structure of a protein with a similar amino-acid sequence is used for this purpose.

Qian and colleagues' models[2] passed the molecular-replacement test with flying colours (Fig. 1): one of the authors' *ab initio* predictions was used successfully as a molecular-replacement model. Furthermore, the authors used their method to refine ten NMR models of protein structures, yielding results that were in better agreement with X-ray data than the original models. And finally, they were able to improve the molecular-replacement scores of several models that started from pro-

de novo/ab initio methods

Fifty years have passed since the Nobel-prizewinning discovery that the amino-acid sequence of a protein determines its three-dimensional structure[1] — yet computational biologists are still unable to predict the shape of a protein from its sequence. Given that there are many more protein sequences available than structures, and that protein shape is crucial for understanding cellular and physiological processes, a method for predicting such structures is vital. The paper by Qian *et al.*[2], which appears on *Nature*'s website today, and in which the structure of a protein containing 112 amino acids is accurately predicted, thus represents a real breakthrough. The authors' model was sufficiently accurate to act as the starting point in the X-ray structure determination of the protein.

Most structural information on proteins is derived from X-ray and nuclear magnetic resonance (NMR) experiments. These have revealed the general characteristics of proteins — for example, sequence motifs that form secondary structural elements such as helices and sheets. Such elements are organized to generate the overall protein architecture, mainly as a result of internal interactions between hydrophobic amino-acid side chains buried within the structure.

The shape of a protein corresponds to the lowest-energy conformation of that mol-

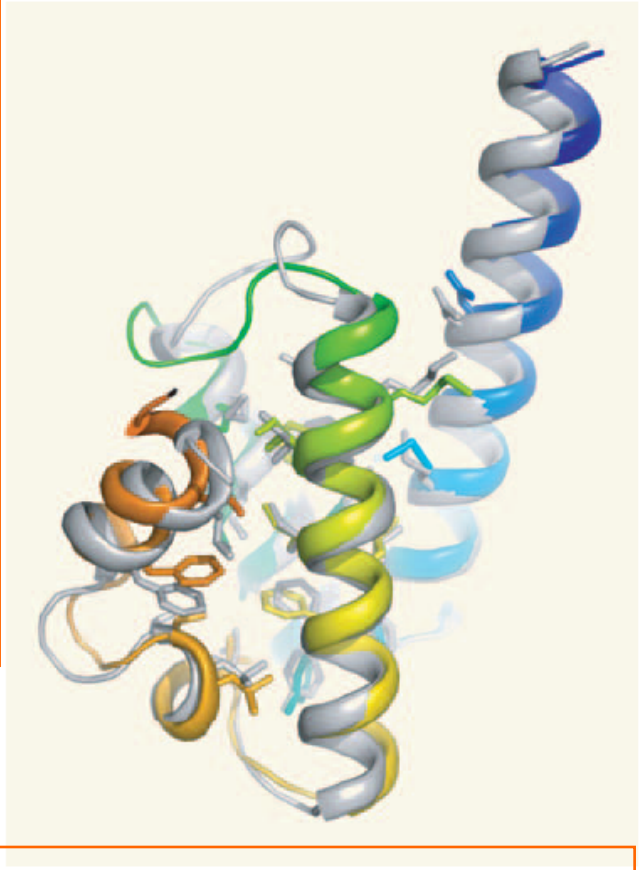


Figure 1 | Model test. Qian *et al.*[2] have developed a computational method for predicting the three-dimensional structure of a protein from its amino-acid sequence. Here, their predicted structure (grey) of a protein is overlaid with the experimentally determined crystal structure (shown in colour) of that protein. The agreement between the two is excellent, with the amino-acid side chains overlapping particularly well.

The main goal of the CASP network is "to obtain an in-depth and objective assessment of our current abilities and inabilities in the area of protein structure prediction"[4]. Every two years the organizers provide the amino-acid sequences of a set of proteins for which undisclosed crystal structures exist. Modellers are challenged to predict structures for the proteins, and these are then assessed against the crystallographic results. The assessors use various scoring systems, but the most rigorous test is the one used by Qian *et al.*[2] to test their own models — can the prediction be used in 'molecular replacement' searches[5] that allow the raw data from X-ray diffraction studies to be related to the structure of the compound being investigated? Normally, a previously determined structure of a protein with a similar amino-acid sequence is used for this purpose.

Qian and colleagues' models[2] passed the molecular-replacement test with flying colours (Fig. 1): one of the authors' *ab initio* predictions was used successfully as a molecular-replacement model. Furthermore, the authors used their method to refine ten NMR models of protein structures, yielding results that were in better agreement with X-ray data than the original models. And finally, they were able to improve the molecular-replacement scores of several models that started from pro-

Jon K. Lærdahl,
Structural Bioinformatics

Rosetta@home

The algorithms used in Rosetta are sophisticated, and the computing resources required to carry out the calculations, to keep track of results and to plan future strategies, are awesome. The authors therefore used a procedure called Rosetta@home[6], which distributes the calculations across a network of home computers — more than 70,000 in June 2006 — whose owners allow the program access to their idle machines.

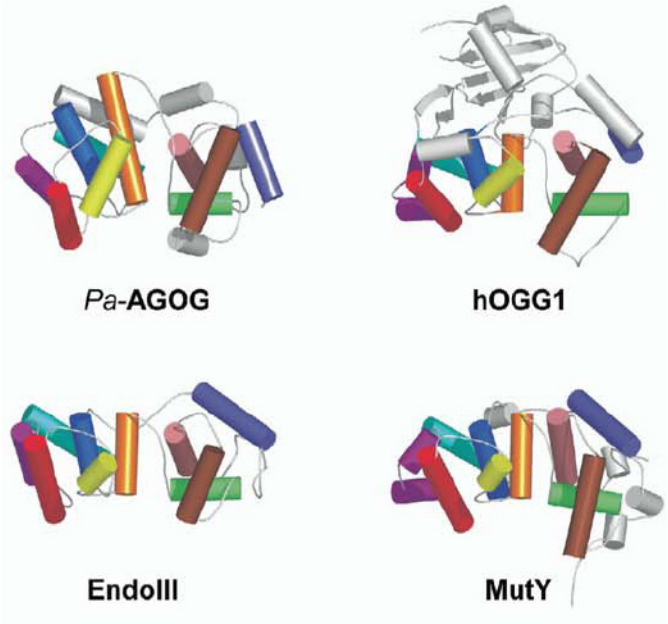
Jon K. Lærdahl,
Structural Bioinformatics

Rosetta@home

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics

```
OGG1_YEAST/1-376    174 SRATEAKLRELGFGYRAKYIIETARKLVNDKAEANITSDTTYLQSICKDAQYEDVREHLMSYNGVGPKVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISAN 276
OGG1_MOUSE/1-345    189 GPEAETHLRKLGLGYRARYVRASAKAILEEQGGP------AWLQQLRV-APYEEAHKALCTLPGVGAKVADCICLMALDKPQAVPVDVHWWQIAHRDYGWHPK 284
OGG1_RAT/1-345      189 GPEVETHLRKLGLGYRARYVCASAKAILEEQGGP------AWLQQLRV-ASYEEAHKALCTLPGVGTKVADCICLMALDKPQAVPVDIHWWQIAHRDYGWQPK 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/folds in Nature?

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**

Last common ancestor
(Long time ago…)

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics

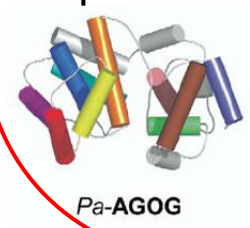
Similar structure
No sequence similarity

Speciation
Gene duplications

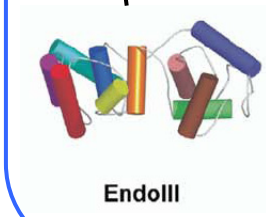
Fairly similar structure
Some sequence similarity

Fairly similar structure
Some sequence similarity

Very similar structure
Significant sequence similarity



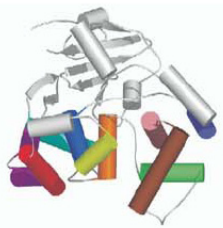
Pa-AGOG



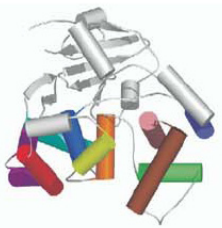
EndoIII

MutY

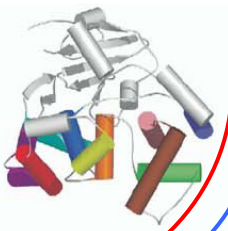
AlkA



Yeast Ogg1



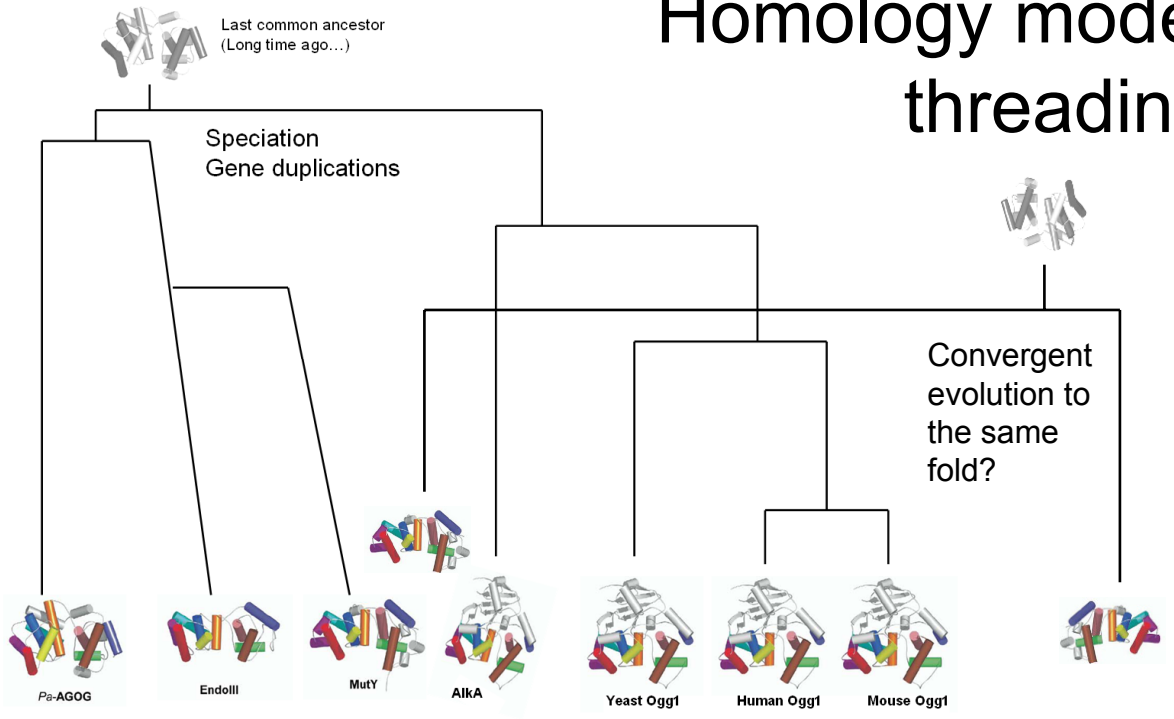
Human Ogg1



Mouse Ogg1

Homology modeling and threading

Jon K. Lærdahl,
Structural Bioinformatics



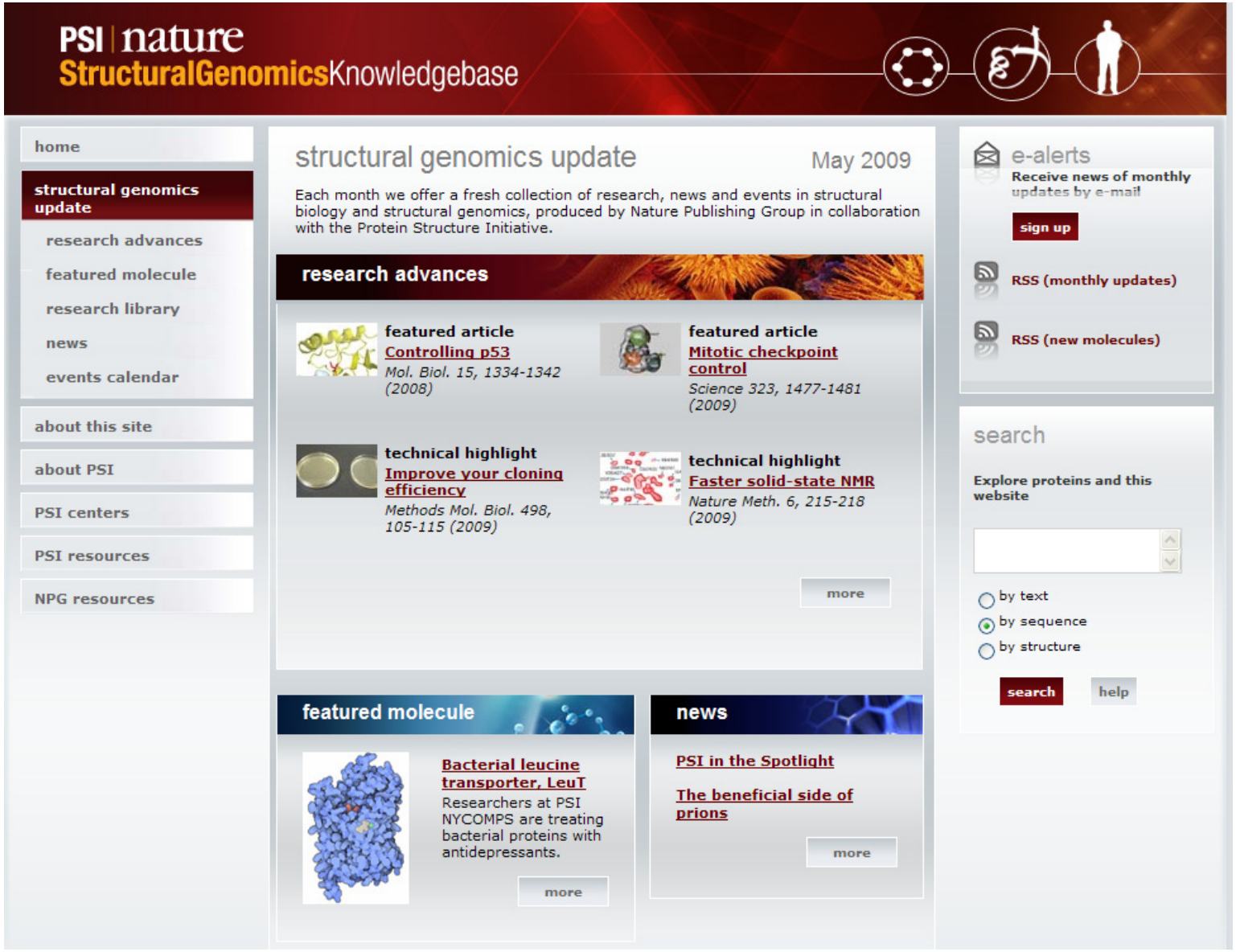
Important goal to have at least one structure in all structural superfamilies!

Structural Genomics Initiatives

- All proteins in a superfamily have the same overall structure/fold
- If we know (from experiment) the structure of 1 protein in a superfamily we may use the information in this structure to model the structure of all other proteins in this superfamily
- Knowledge-based modeling
    - Based on structures in the PDB (*i.e.* they are not *ab initio*)
    - Homology modeling
        - When there is significant sequence identity between the protein you want to model (target) and the known structure (template)
    - Threading
        - When there is no or little sequence identity between target and template

Structural genomics/The Protein Structure Initiative (PSI)

Jon K. Lærdahl,
Structural Bioinformatics



Usually: solve the structure of a protein only after thorough biological analysis (years of research?)

Here: solve structures of lots of proteins with emphasis on those that are likely the have a new fold

Structural genomics/The Protein Structure Initiative (PSI)

Jon K. Lærdahl,
Structural Bioinformatics

Structural genomics/The Protein Structure Initiative (PSI)

Jon K. Lærdahl,
Structural Bioinformatics

Archaeoglobus fulgidus DSM 4304 protein AAB89001.1 has a new fold determined by the MCSG (2PHN/2G9I)

5 yrs ago: “Only” 3D structures for proteins that had been studied a lot

Now: many 3D structures for proteins with unknown function!

Jon K. Lærdahl,
Structural Bioinformatics

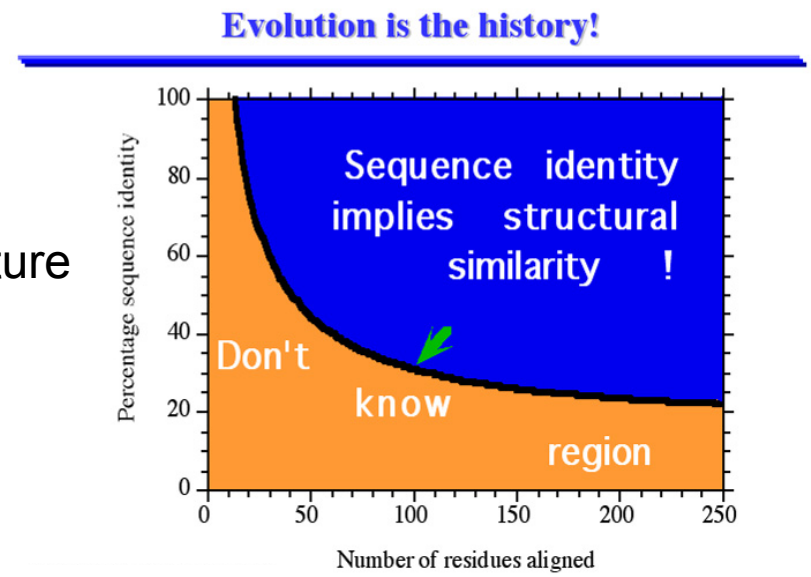
Homology modeling

• Based on: during evolution, structure is more stable and conserved than the associated sequence

• Similar sequences give nearly identical structure
• Distantly related sequences fold into similar structures

• 20-30% identical residues to a known (experimental) structure

→ Might be able to predict the 3D structure with some confidence



B. Rost, *Prot. Engin.* **12**, 85 (1999)

Known (experimental) structure of protein 1 (*template*)
&
Sequence alignment with protein 2 (*target)*

→ Model of protein 2

• 30% sequence identity necessary?
• My experience: Might get reasonable results also at 20% or even below
• Depends on
  • Many indels or not?
  • Length of alignment
  • Automatic or manual modeling?

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

1. Template selection:
   - Find template in PDB and align sequences
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

1. Template selection:
   - Find templ
     sequences
2. Correct alignr
   - Use the be
   - Correct pla
     and deletio
3. Backbone mo
4. Model loops a
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

I want to model this!

>gi|84618885|emb|CAJ31885.1| methylpurine-DNA glycosylase [Bacillus cereus]
MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQT
PERRQLLKDVIQIHTLPDPKDFRIIVRELWDLPEREFQA
AALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVP
TFLGNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLK
YKQKMDEELLFWVIGQLHSSKEFFIQKAIGWVLREYAKT
KPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGE
TLS

Homology modeling

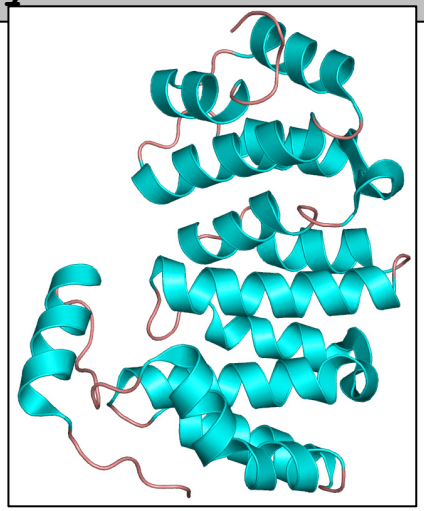
Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

1. Template selection:
   - Find template in PDB and a sequences
2. Correct alignments
   - Use the best MSA program
   - Correct placement of inser and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

Do sequence search in all "PDB sequences"

- **Useful templates have 30% or higher sequence identity to target (but sometimes even lower)**
- **Several templates?**
  - **Resolution?**
  - **Highest sequence identity?**
  - **Cofactors?**
  - **Use the structure that best fits your task**

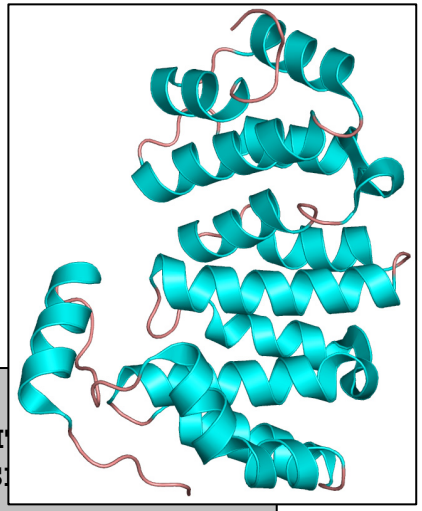


Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

1. Template selection:
   - Find template in PDB and align sequences
2. Correct alignments
3.
4.
5. Refine and optimize model
6. Validate and check model quality!



```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQTPERRQLLKDVIQIH
EF3068    --------MDTLQFQKNPETAAKMSAYMKHQFVFAGIPAPERQALSKQLLKES
                  : :  :****.*  *: ***::*:* ** :***: * *::::       :.
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL  120
EF3068    LCQEIEAYYQKTEREYQYVAIDLALQNVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG  122
          :   :.  :: .***:* .*:*:  :  : :.  .:  ::  :  *:***:**:    *
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF  180
EF3068    SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF  171
          . : *:  **:.  :. : .::*:* :*.*: :**  *:* :::**  .*   ::::***
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS     237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------     217
          ******* **:*:**:*: * * :::  *:**::**. *:: :
```

Alignment of the sequences of *B. cereus* AlkD (target) and *E. faecalis* hypothetical protein EF3068 (template from MCSG).

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (tar

1. Template selection:
   - Find template in PDB and a sequences
2. Correct alignments
   - Use the best MSA program
   - Correct placement of inser eletions
3. Back
4. Mod
   - R
   - L o
5. Refi
6. Valid

Obtaining the correct alignment is *the most important step!!* in homology modeling

FIRST: Align target, template and a large number (50-100?) of homologs with Praline, T-Coffee, Muscle or a different good MSA program

Use target/template alignment from this MSA

SECOND: Look at the template structure and move all indels
- **to loops**
- **out of helices/sheets**

Check indels!

```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARY
EF3068    --------MDTLQFQKNPETAAKMSAY
                  : :  :****.*  *: *
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL  120
EF3068    LCQFIEAYYQKTEREYQVVAIDLALQNVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG  122
          :   :.  :: .***:* .*:*:  :  : :.  .:  ::  :  *:***:**:    *
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF  180
EF3068    SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF  171
          . : *:  **:.  :. : .::*:* :*.*: :**  *:* :::**  .*   ::::***
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS     237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------     217
          ******* **:*:**:*: * * :::  *:**::**. *:: :
```

Alignment of the sequences of *B. cereus* AlkD (target) and *E. faecalis* hypothetical protein EF3068 (template from MCSG).

Homology modeling

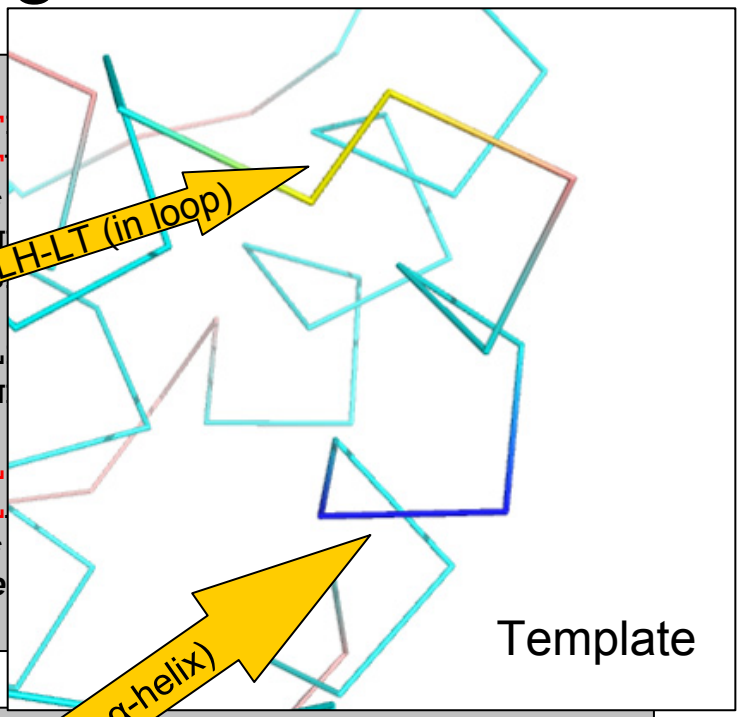
Jon K. Lærdahl,
Structural Bioinformatics

Start with

1. Temp
   - Fir
     se
2. Corre
   - Us
   - Co
     and deletions
3. Back
4. Mod
   - R
   - L
     o
5. Refi
6. Valid

```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHF
EF3068    --------MDTLQFQKNPETAAKMSAYMKHQF
                  : :  :****.*  *: ***::*
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMMQKYKKYIN
EF3068    LCQEIEAYYQKTEREYQYVAIDLALQNF
          :   :.  :: .***:* .*.*: *  : :.
Bc_AlkD   GNIFLQHPELISAYIPKWIASLNIWLQRAAIL
EF3068    SWVALH-LTELPTIFALFYGAFNFWNRRVALN
          . : *:    :.: :. : .::*:* :*.*:
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPL
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPL
          ******* **:*:**:*: * * :::  *:**
```

**CORRECTED Alignment of the sequences of *B. cere*
hypothetical protein EF3068 (template from MCSG).**




```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHF  QTPERRQLLKDVIQIHTLPDPKD   60
EF3068    --------MDTLQFQKNPETAAKMSAYMK  AGIPAPERQALSKQLLKESHTWPKEK   52
                  : :  :****.*  *: *  *:* ** :***: * *::::       :.
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMM  KYINETHIPFLEELIVTKSWWDTVDSIVPTFL  120
EF3068    LCQEIEAYYQKTEREYQYVAI  QNVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG  122
          :   :.  :: .***:* .*:*:  :  : :.  .:  ::  :  *:***:**:    *
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF  180
EF3068    SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF  171
          . : *:  **:.  :. : .::*:* :*.*: :**  *:* :::**  .*   ::::***
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS     237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------     217
          ******* **:*:**:*: * * :::  *:**::**. *:: :
```

Alignment of the sequences of *B. cereus* AlkD (target) and *E. faecalis* hypothetical protein EF3068 (template from MCSG).

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

1. Template selection:
   - Find template in PDB and align sequences
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

The most important step in homology modeling!

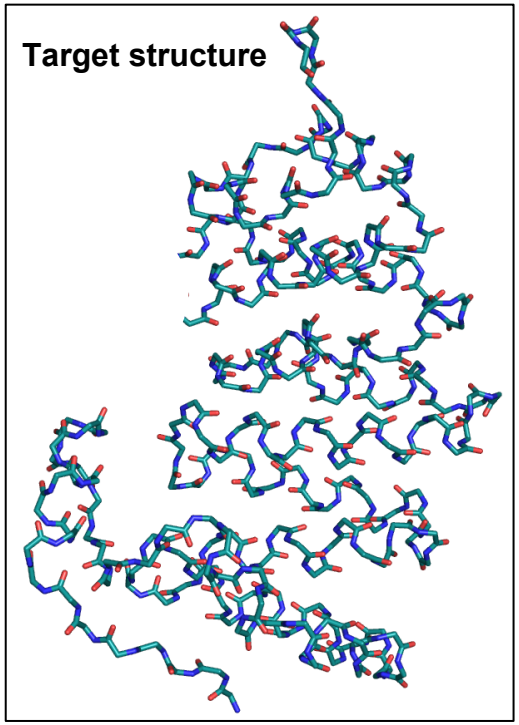
Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (tar

1. Template selection:
   - Find template in PDB and sequences
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

For all aligned residues in template and target:
- **Take coordinates for template backbone atoms and use for target**
- **If residues are identical: Use all atom coordinates from template in target**
- **Indels: Nothing to copy**



Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Sta (tar

1.

nd

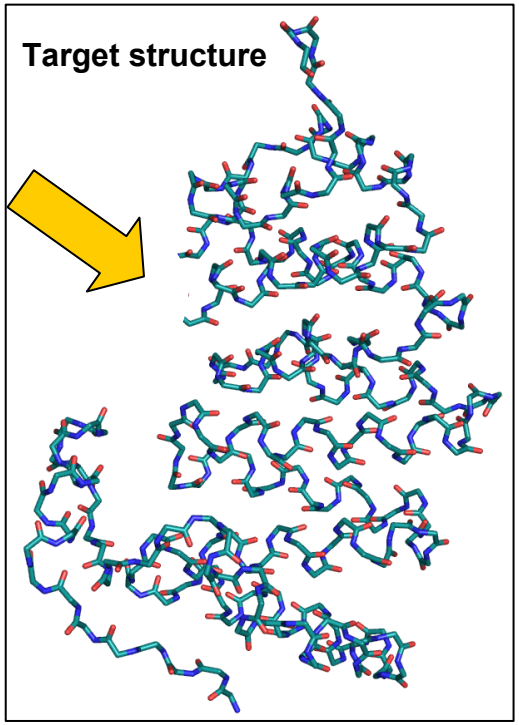
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

Short loops(3-5 residues): Reliable results with both methods

Long loops(more than 10-15 residues): Highly unlikely that you get a correct result!!

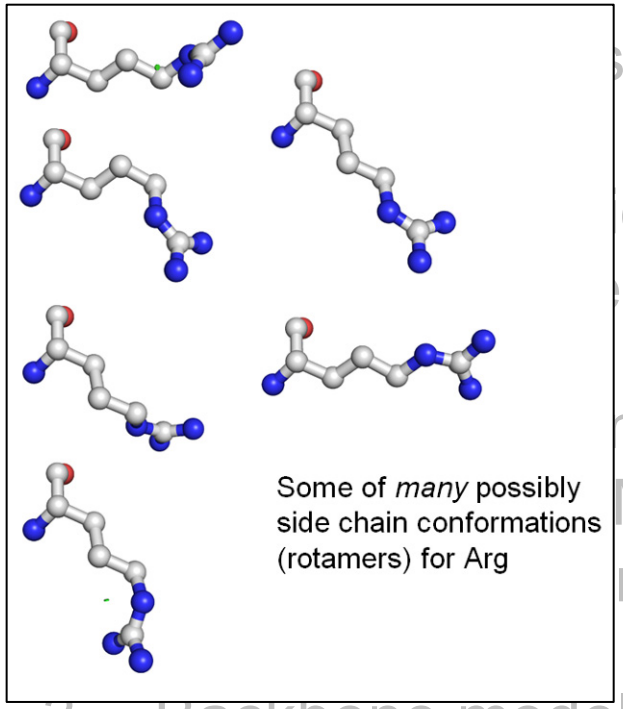
***Ab initio*: Generates random loops and chooses the one with**
- **Lowest energy scores**
- **Ok Ramachandran plot**
- **No clashes**

Database method: Try loops taken from a "loop-library" extracted from the PDB



Homology modeling

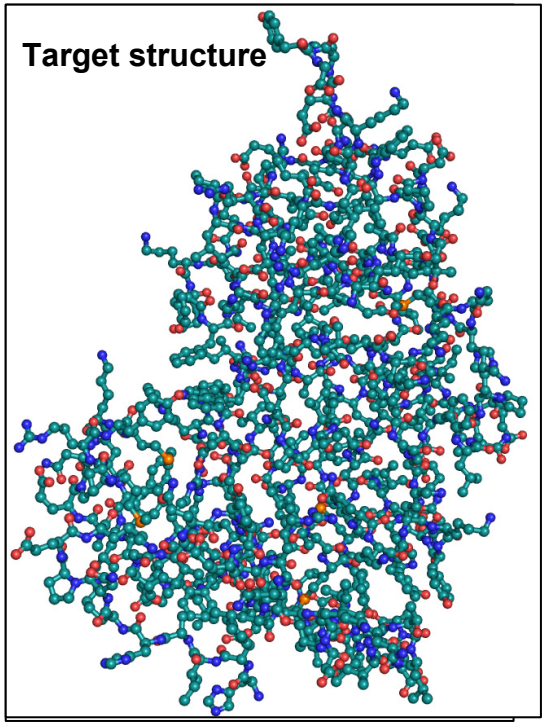
Jon K. Lærdahl,
Structural Bioinformatics



Get side chain conformations from rotamer libraries generated from known structures

Use those that give
- **Lowest energy score**
- **No clashes with backbone/other side chains**

3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!



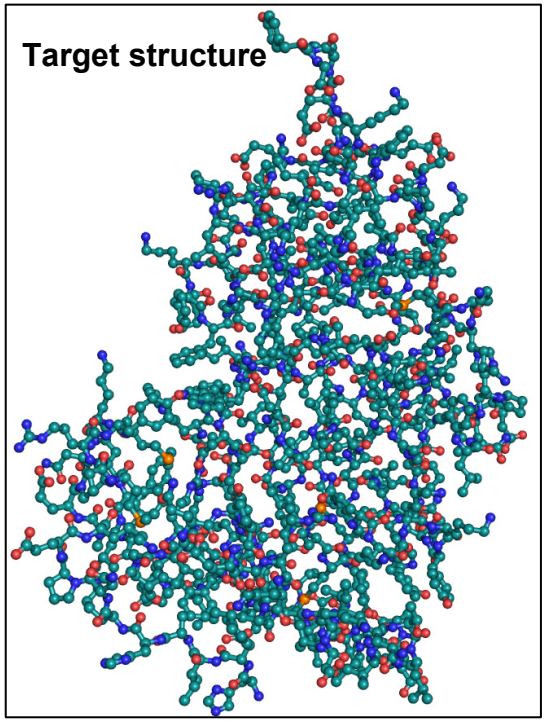
Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (tar

1. Template selection:
   - Find template in PDB and sequences
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

Do a few hundred iterations of energy minimization?
- **Will hopefully remove clashes and very unfavorable conformations**
- **Too many iterations will most likely destroy structure**
- **Not always necessary (depends on the program)**



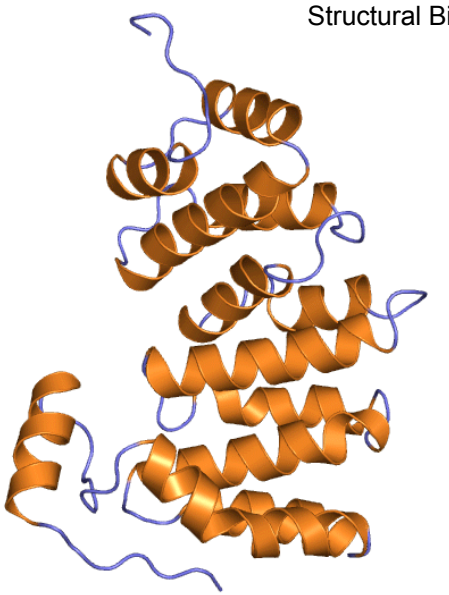
Jon K. Lærdahl,
Structural Bioinformatics

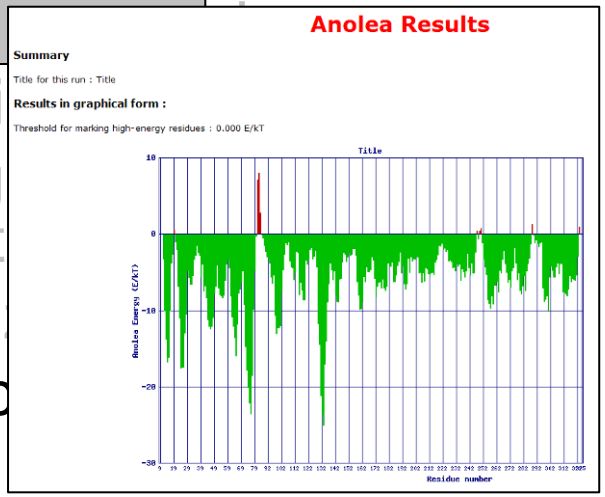
Homology modeling

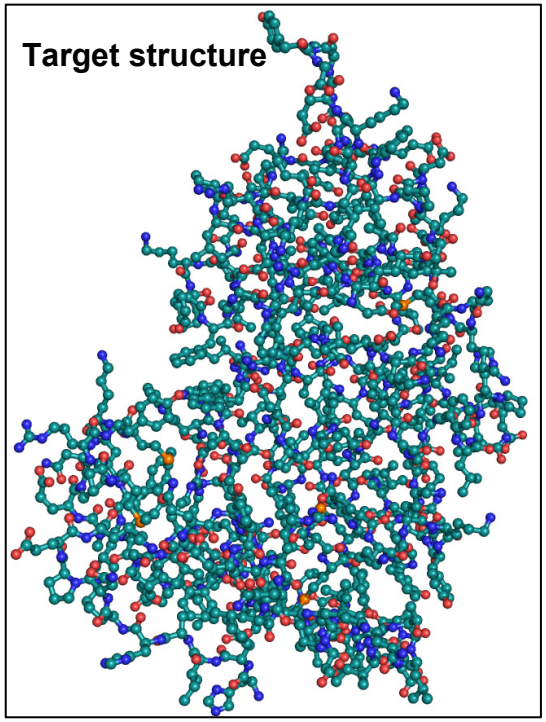
Check if model makes sense?
- **Ramachandran plot ok?**
- **No clashes?**
- **No funny bond lengths/angles/conformations?**
- **Use programs such as:**
  - **Procheck**
  - **WHAT IF**
  - **ANOLEA**
  - **Verify3D**
- **These can only check if the chemical/physical properties are ok**
- **The model might still be 100% meaningless biologically and completely wrong!**

– Rotamer librari
– Loop modeling
or *ab initio* met

5. Refine and optimi
6. Validate and chec

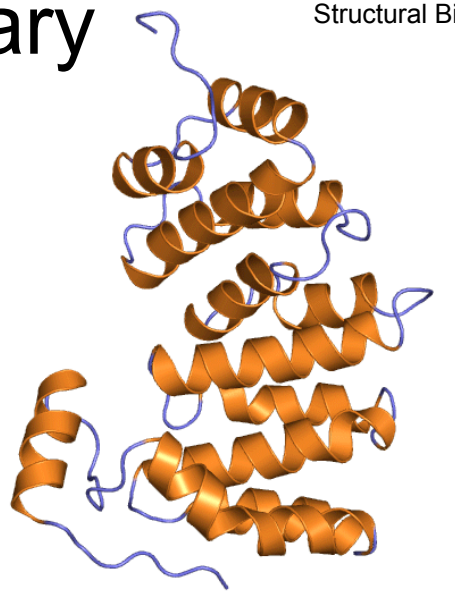






Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling summary



1. Template selection:
   - Find template in PDB and align sequences
2. Correct alignments
   - IMPORTANT!
3. Backbone model building
4. Model loops and side-chains
5. Refine and optimize model(?)
6. Validate and check model quality!

Automatic models usually less accurate that manually generated models (if the modeler knows what she is doing…)

Tools:
- Modeller
- Swiss-Model
- 3D-JIGSAW

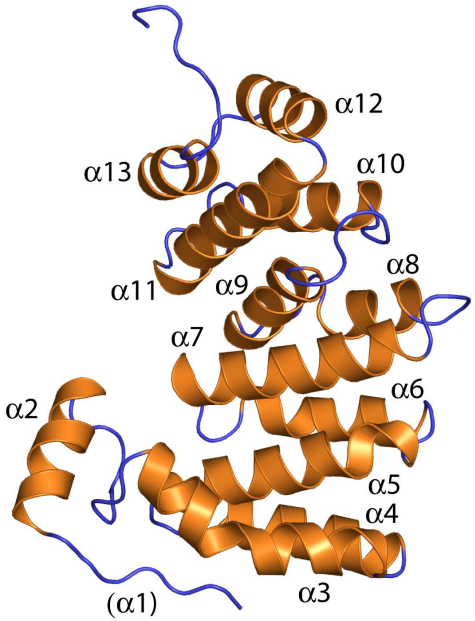
Homology model databases:
- Modbase (automatic modeling with Modeller)
- SWISS-MODEL Repository (automatic modeling with Swiss-Model)

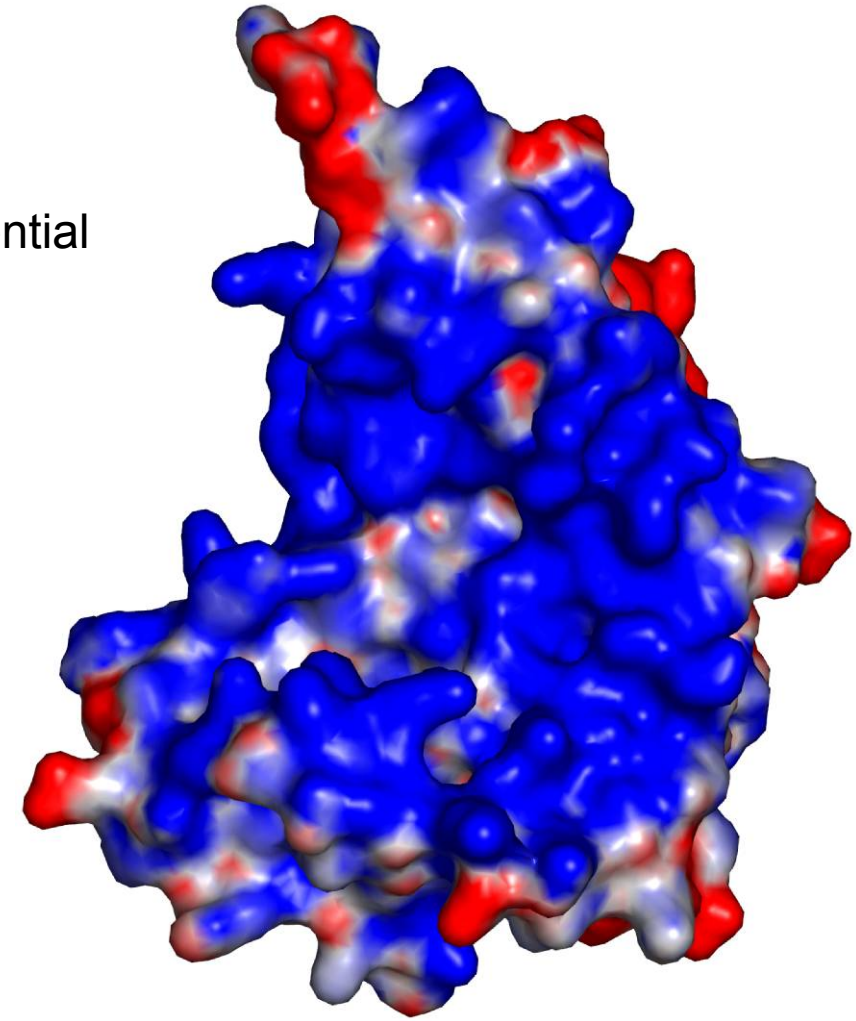
Structural bioinformatics

When the structure (experimental or model) is available, there are many more possibilities to obtain understanding

Some examples:

B. cereus AlkD electrostatic potential

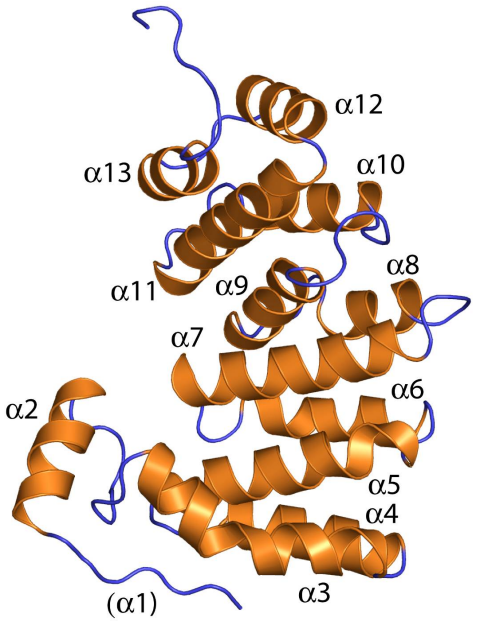


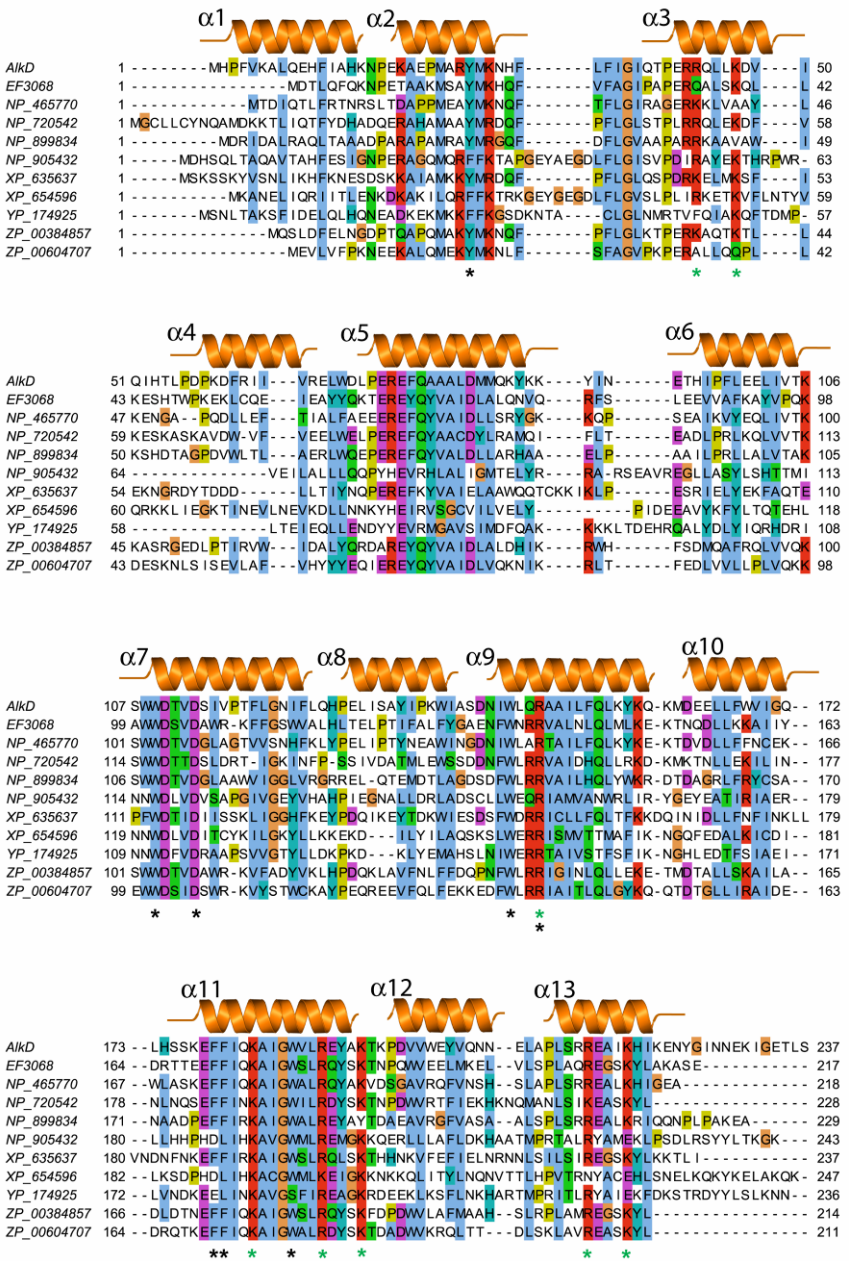


Structural bioinformatics

Jon K. Lærdahl,
Structural Bioinformatics

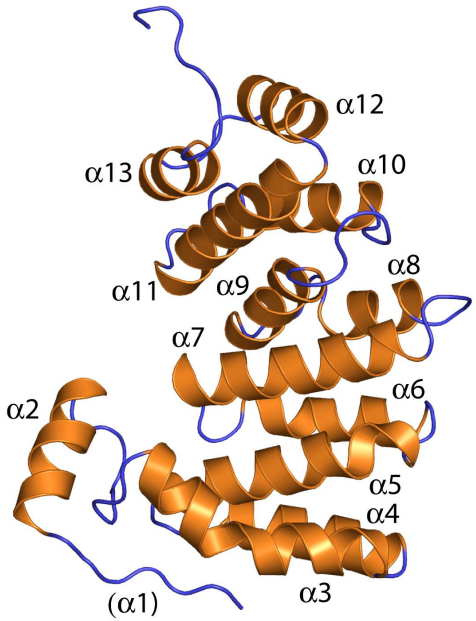
B. cereus AlkD sequence conservation from ConSurf:

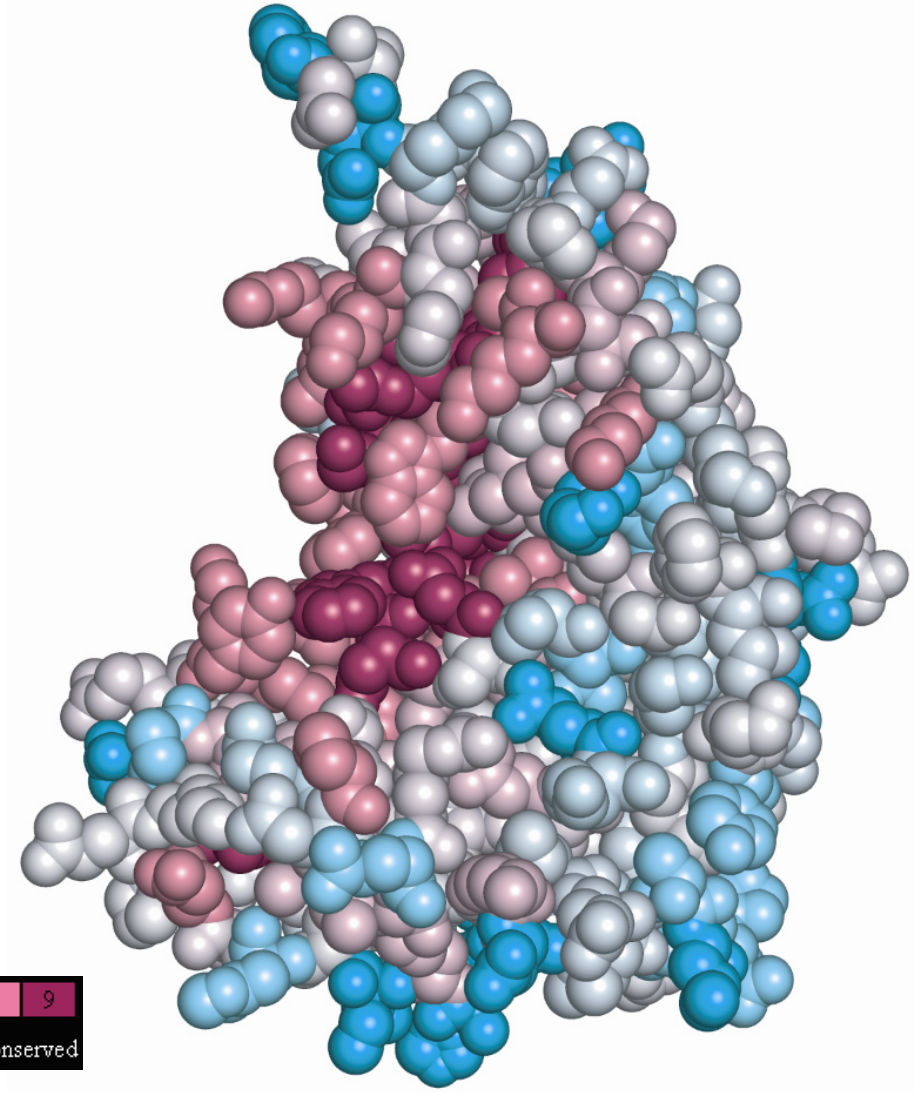


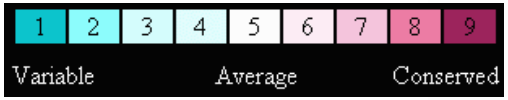


Structural bioinformatics

B. cereus sequence conservation from ConSurf:

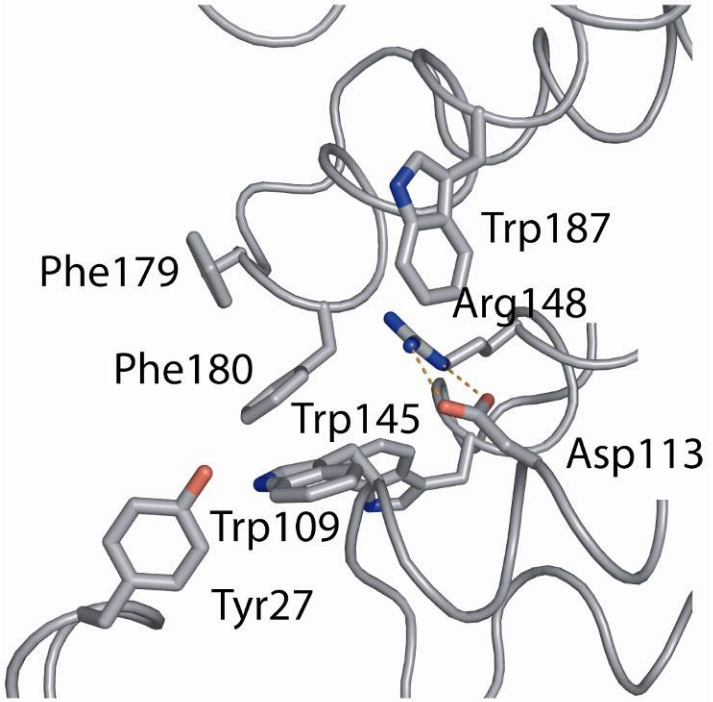


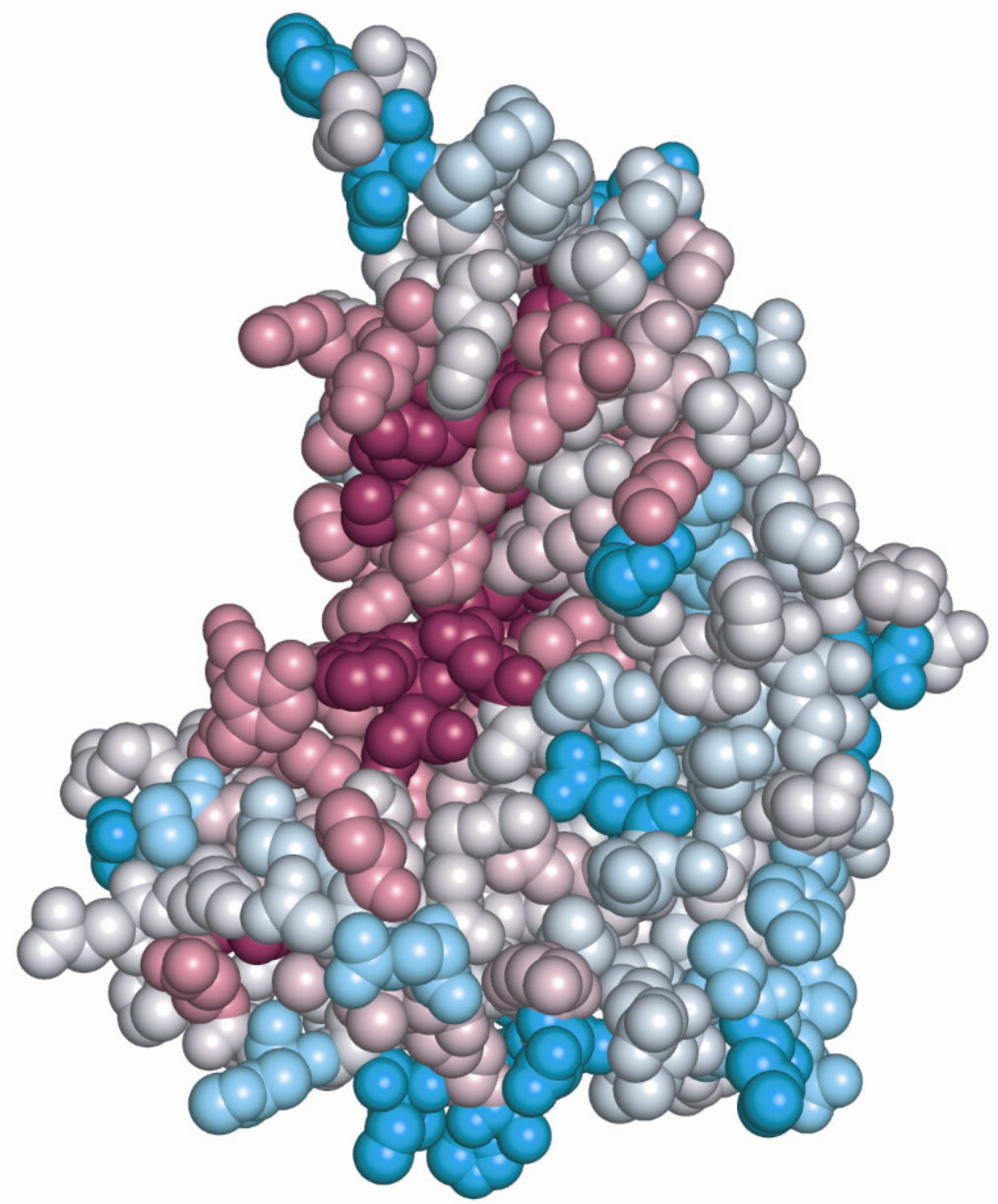




Structural bioinformatics

B. cereus sequence conservation from ConSurf:





B. Dalhus *et al.,* Nucleic Acids Res. **35**, 2451 (2007)

When homology modeling does not work?

Jon K. Lærdahl,
Structural Bioinformatics

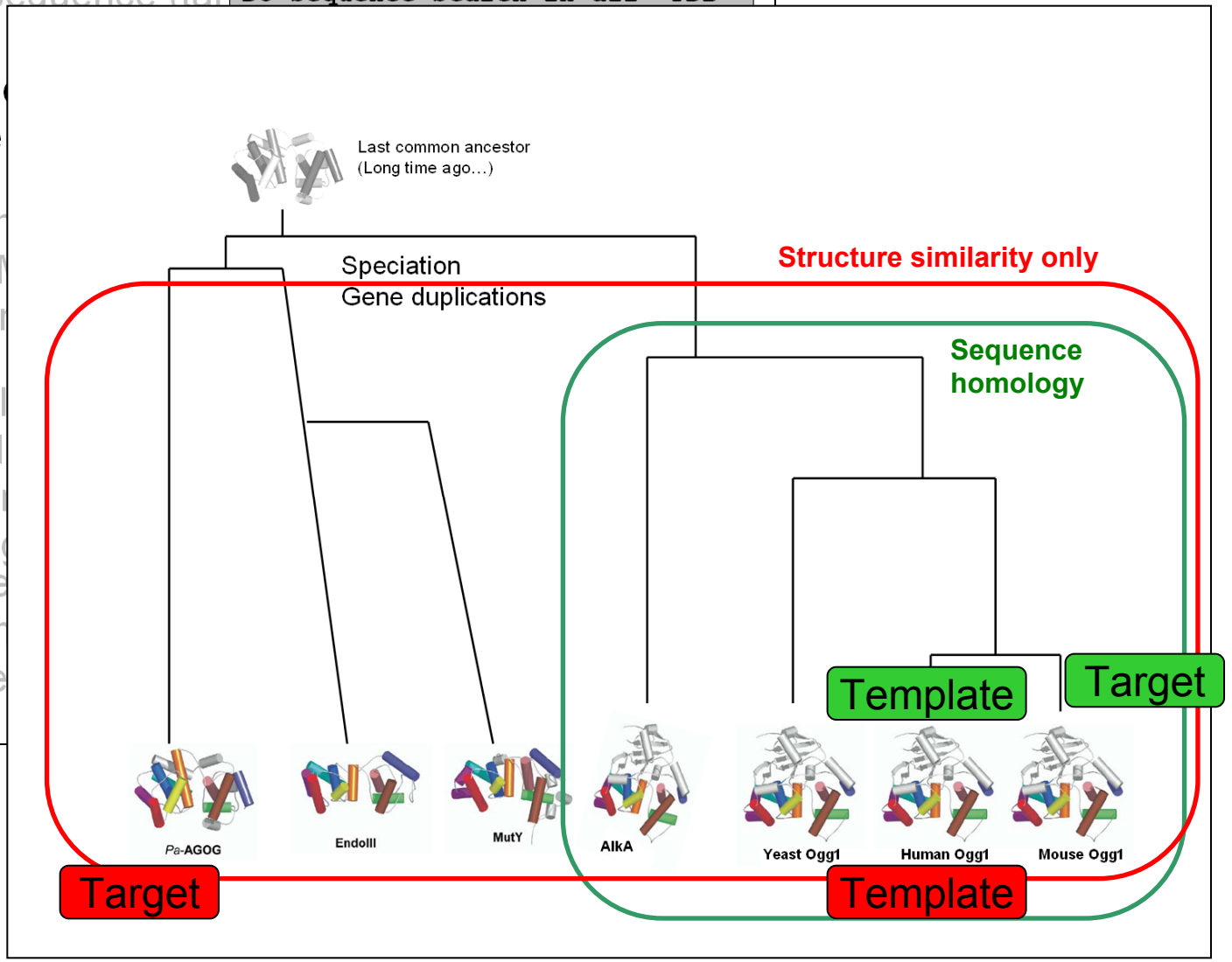
Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

Do sequence search in all "PDB

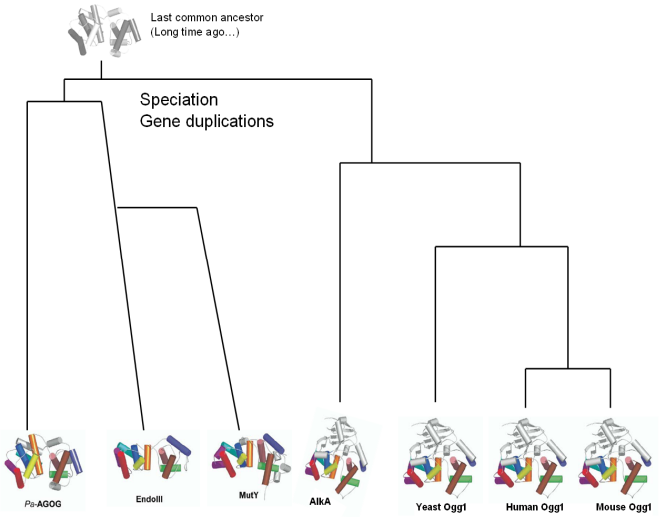
1. Template selecti
   - Find template sequences
2. Correct alignmen
   - Use the best M
   - Correct placer and deletions
3. Backbone model
4. Model loops and
   - Rotamer libra
   - Loop modelin or *ab initio* me
5. Refine and optim
6. Validate and che



Threading

Jon K. Lærdahl,
Structural Bioinformatics

- Same as “fold recognition”
- Prediction of the structural fold of a protein sequence by ”fitting” the sequence onto structures from a structural database
- Secondary structure prediction is important to choose the best template candidates
- Calculate energy and other parameters for all possibilities
- Choose the best fold, for example the one with the lowest energy
- Detects structural similarity in the absence of sequence similarity
    - GenThreader
    - Phyre (new and better than 3D-PSSM)
    - Fugue
- May only be used to generate a *rough model*
- Threading does *not* give accurate models!
- May be used to detect remote homologs
    - No hits with BLAST or PSI-BLAST?
    - *Try threading!*
- “BLAST will give you the protein family”
- “Threading will give you the protein superfamily”
- Threading is more useful for detecting homology than for generating 3D structures?

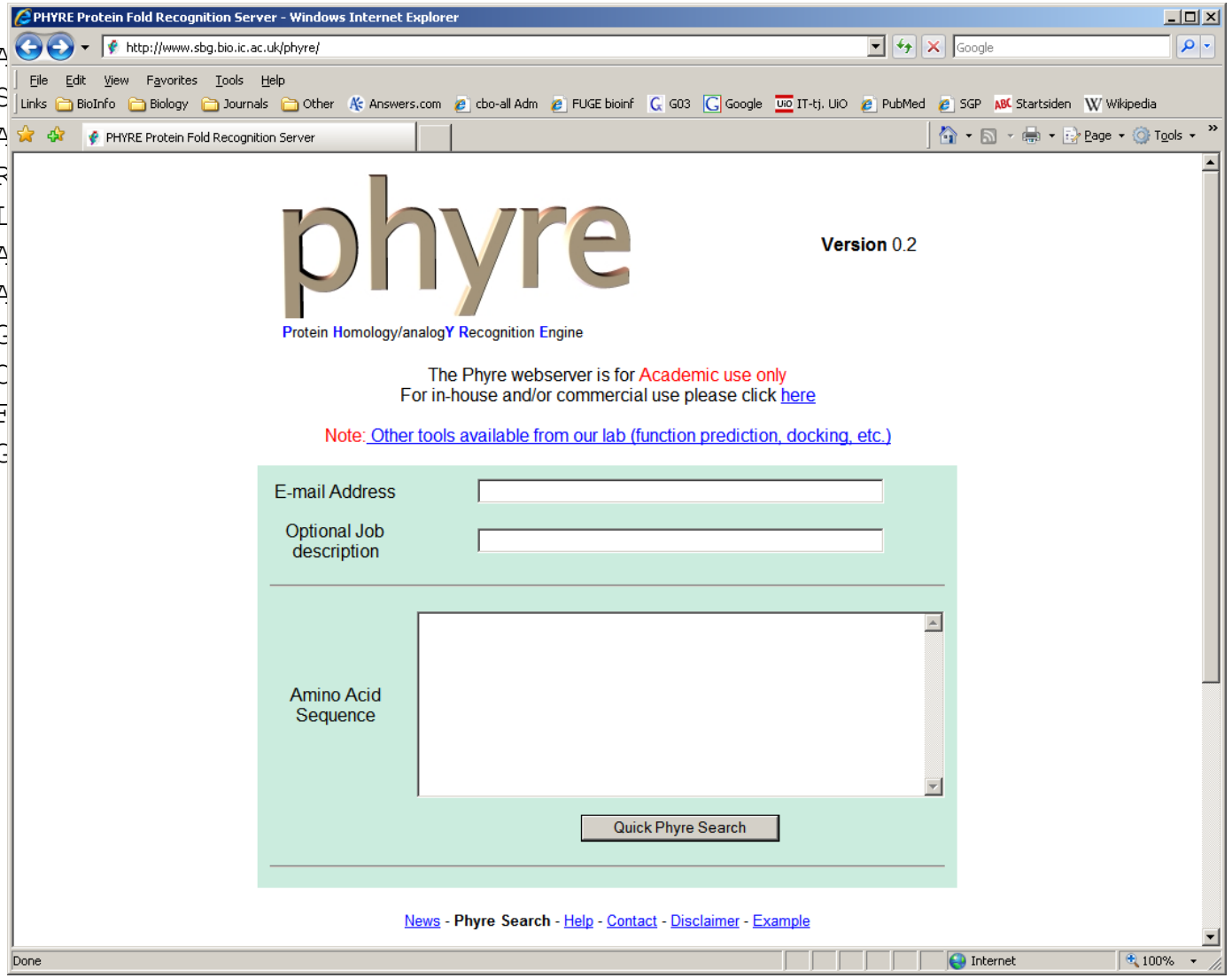


Jon K. Lærdahl,
Structural Bioinformatics

Threading/Fold recognition

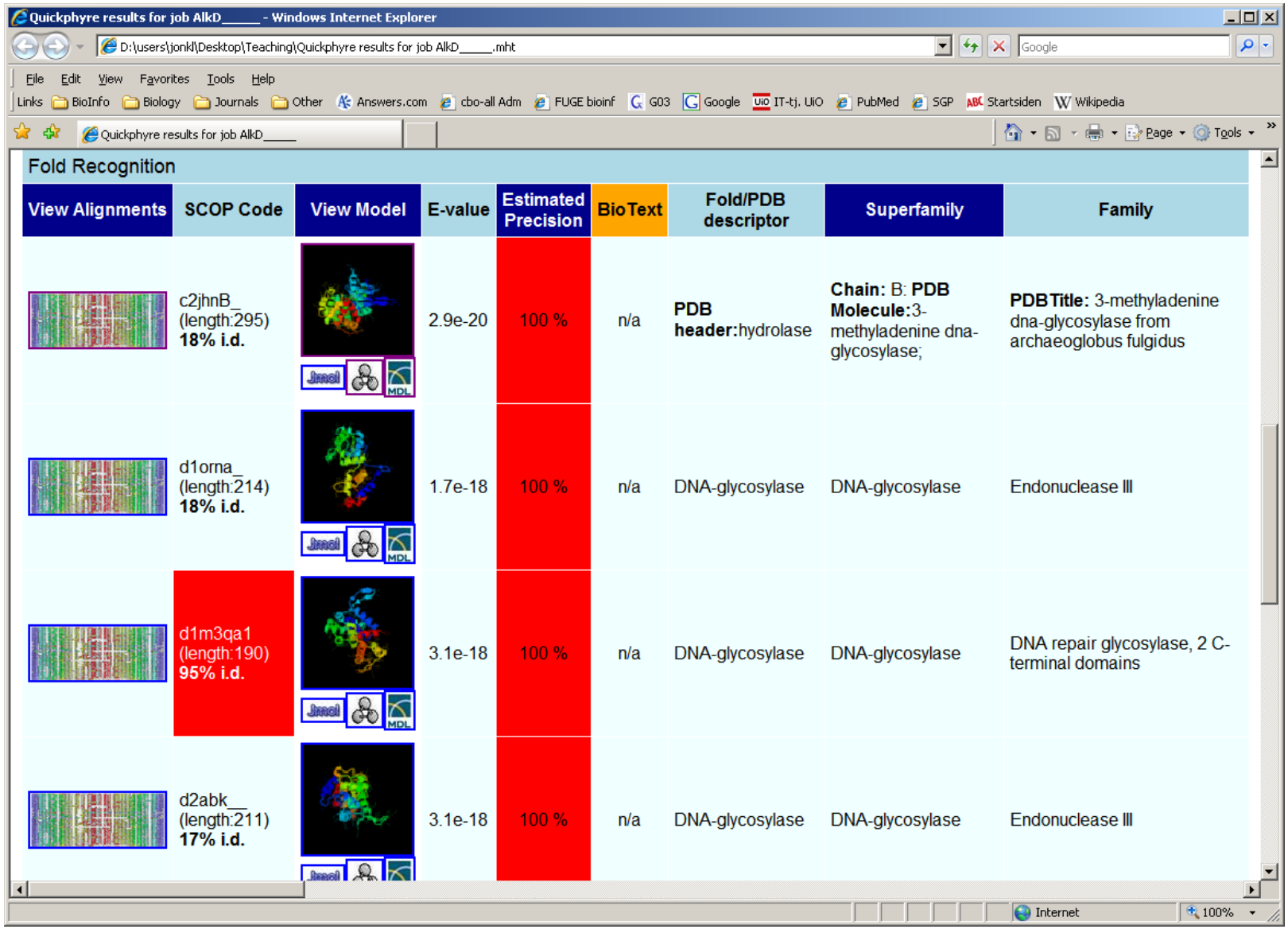
```
>Unknown_protein
MPARALLPRRMGHRTLA
PSGQSFRWREQSPAHWS
RGDKSQASRPTPDELEA
SHFQEVAQKFQGVRLLR
GMVERLCQAFGPRLIQL
LRKLGLGYRARYVSASA
EAHKALCILPGVGTKVA
IAQRDYSWHPTTSQAKG
WAQATPPSYRCCSVPTC
RWGTLDKEIPQAPSPPF
KARHPQIKQSVCTTRWG
```

What is the structure of this?



Jon K. Lærdahl,
Structural Bioinformatics

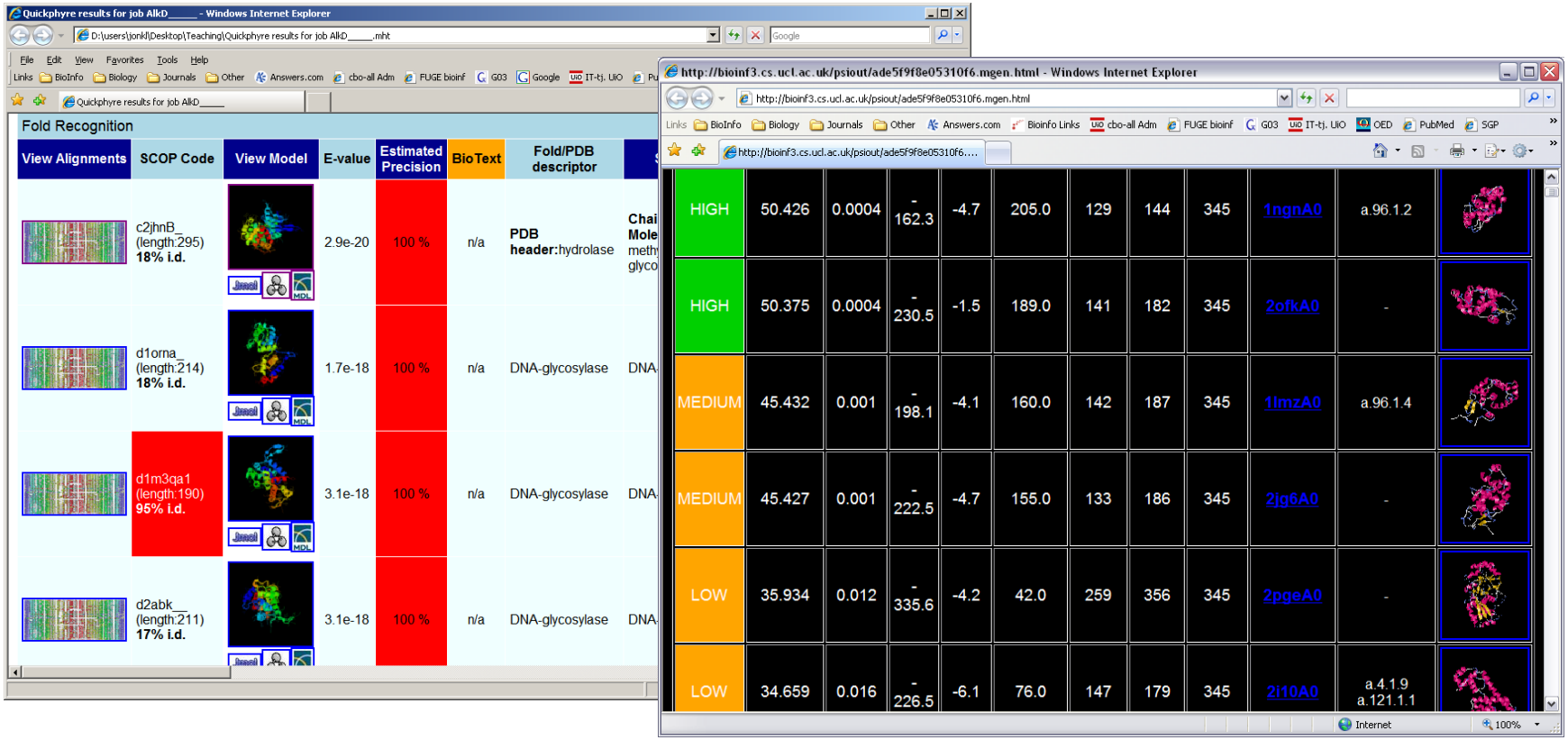
Threading/Fold recognition



View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily	Family
	c2jhnB_ (length:295) **18% i.d.**		2.9e-20	100 %	n/a	**PDB header:**hydrolase	**Chain:** B: **PDB Molecule:**3-methyladenine dna-glycosylase;	**PDB Title:** 3-methyladenine dna-glycosylase from archaeoglobus fulgidus
	d1orna_ (length:214) **18% i.d.**		1.7e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	Endonuclease III
	d1m3qa1 (length:190) **95% i.d.**		3.1e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	DNA repair glycosylase, 2 C-terminal domains
	d2abk__ (length:211) **17% i.d.**		3.1e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	Endonuclease III

Jon K. Lærdahl,
Structural Bioinformatics

Threading/Fold recognition



As for other bioinformatics methods:

- Precision might be overestimated
- The results might be completely wrong
- If several independent tools give the same result it is much more likely to be correct
- *Use several tools!*

3D structure prediction - Summary

Jon K. Lærdahl,
Structural Bioinformatics

Start with target sequence

1. Sequence homology to protein that has structure in PDB (better than 20-30% sequence identity) → Homology modeling
2. No good hit with sequence searching:
   - Fold recognition/threading might give correct fold
3. No results from fold recognition/threading:
   - You *might* try *ab initio* folding, but the result will most likely be very unreliable

Homology models can be of good quality and might be useful for:

- Docking two or more proteins together
- Designing drugs
- Identifying active sites and amino acids for generating mutant proteins, etc.

Fold recognition/threading might give the protein overall fold and possibly indicate function

If the fold is that of a helical cytokine → Your protein is also possibly a helical cytokine

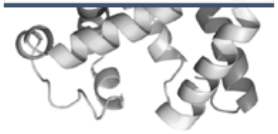
CASP: Critical Assessment of Techniques for Protein Structure Prediction
- Benchmarking of structure prediction tools

Jon K. Lærdahl,
Structural Bioinformatics

Protein Structure Prediction Center

Menu

- Home
- FORCASP Forum
- PC Login
- PC Registration
- CASP Experiments
  - CASP8 (2008)
    - Home
    - My CASP8 profile
    - Targets
    - Predictions
    - CASP8 in numbers
  - CASP7 (2006)
  - CASP6 (2004)
  - CASP5 (2002)
  - CASP4 (2000)
  - CASP3 (1998)
  - CASP2 (1996)
  - CASP1 (1994)
- CASP Initiatives
- Outreach
- Local Services
- Downloads
- Links
- Feedback
- FAQ
- People



Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. In addition to support of the CASP meetings our goal is to promote an evaluation of prediction methods on a continuing basis.

CASP experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. The organizers are thankful to CASP assessors for their valuable contribution to this field.

There have been seven previous CASP experiments.

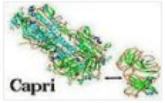
CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008)

Proceedings

Click on the logo below to proceed to the main page of the latest CASP experiment.

FORCASP "no more dead trees"

Discussion Forum
Old Discussion Forum



Prediction of docking interactions

Live Bench

Automated benchmarking of prediction servers (will be again available shortly)

CAFASP

Assessment of automated structure prediction

EVA

Automatic evaluation of prediction servers

Message Board

CASP8 job fair

Dear CASPers, we would like to take advantage of the CASP8 meeting to help young scientists looking for a position in Computational Biology meeting potential group leaders and vice versa. We pla ...

CASP8 junior scientist session

Dear CASP Participants, During the CASP8 meeting we will dedicate a special session to topics suggested by junior scholars. We will also award a poster prize and invite oral presentations. All ...

CASP8 fellowships

We have now completed awarding CASP8 registration fee /lodging /travel fellowships and the recipients have been notified. The CASP8 speaker selection will be completed around Oct. 23. CASP8 speakers w ...

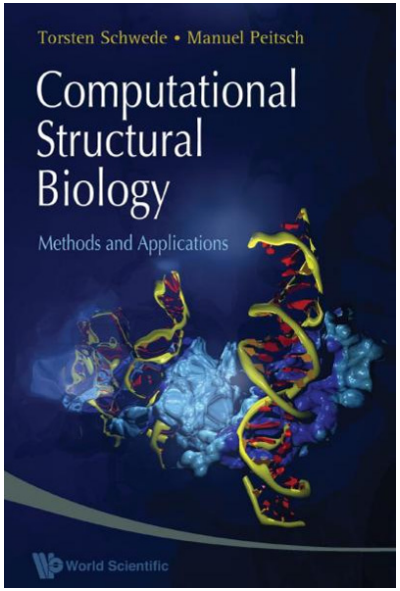
Literature & Useful links

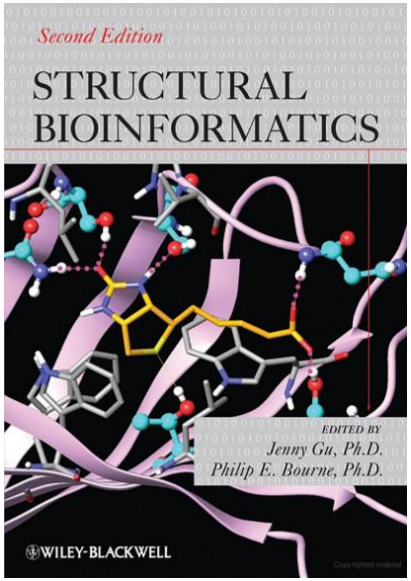
- http://bioinformatics.ca/links_directory
- Homology modeling: L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey & T. Schwede, *“Protein structure homology modeling using SWISS-MODEL workspace”,* Nature. Protoc. **4**, 1 (2009).

- B. Dalhus, I. Høydal Helle, P. H. Backe, I. Alseth, T. Rognes, M. Bjørås, and J. K. Laerdahl, *"Structural insight into repair of alkylated DNA by a new superfamily of DNA glycosylases comprising HEAT-like repeats",* Nucleic Acids Res. **35**, 2451 (2007).
- J. Cameron, Ø. L. Holla, K. E. Berge, M. A. Kulseth, T. Ranheim, T. P. Leren, and J. K. Laerdahl, *"Investigations on the evolutionary conservation of PCSK9 reveal a functionally important protrusion",* FEBS J. **275**, 4121 (2008).

Talk to us!

jonkl@medisin.uio.no





End

jonkl@medisin.uio.no