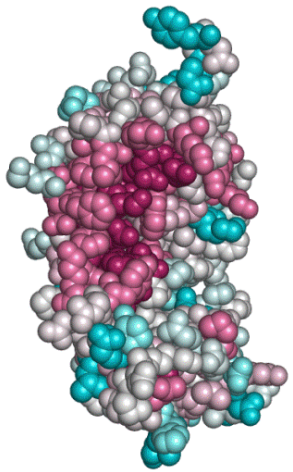
Bioinformatics for molecular biology

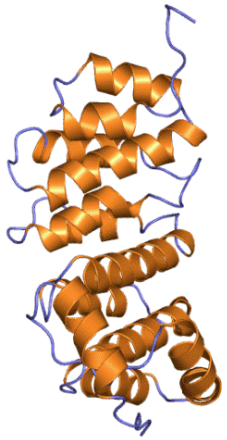
Structural bioinformatics tools, predictors, and 3D modeling – Structural Bioinformatics

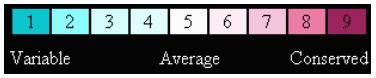
Dr Jon K. Lærdahl, Research Scientist

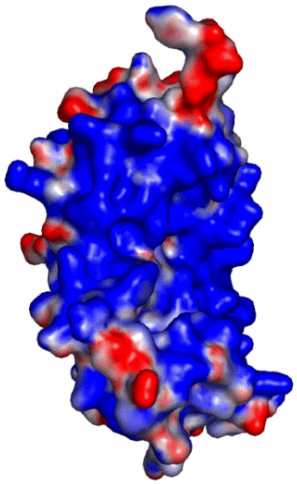
Department of Microbiology,
Oslo University Hospital - Rikshospitalet &
Bioinformatics Core Facility/CLS initiative,
University of Oslo

E-mail: jonkl@medisin.uio Phone: +47 22844784
Group: Torbjørn Rognes
(http://www.ous-research.no/rognes)
CF: Bioinformatics services
(http://core.rr-research.no/bioinformatics)
CLS: Bioinformatics education
(http://www.mn.uio.no/ifi/english/research/networks/clsi)
Main research area:
Structural and Applied Bioinformatics









Oslo universitetssykehus

Overview

Jon K. Lærdahl,
Structural Bioinformatics

Earlier…

- Protein Structure Review
  - Amino acids, polypeptides, secondary structure elements, visualization, structure determination by X-ray crystallography and NMR methods, PDB

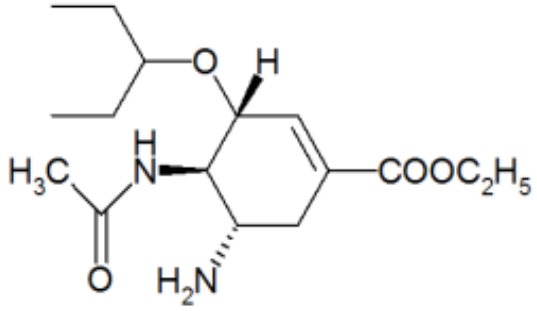
Now

- Structure comparison and classification (CASP & SCOP)
- Predictors
- 3D structure modeling
  - *Ab initio*
  - Threading/fold recognition
  - Homology modeling
- Practical exercises
  - PyMOL & visualization
- Practical Exercises
  - Homology modeling of influenza neuraminidase (Tamiflu resistance?)
  - Other homology modeling
  - Threading
  - Your own project?

Stop me and ask questions!!

Structural bioinformatics

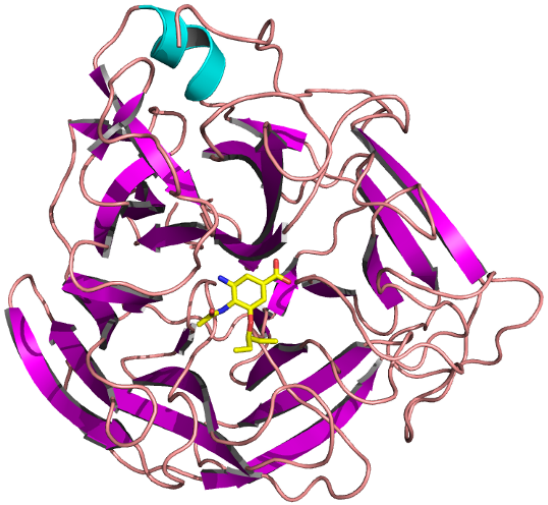
Jon K. Lærdahl,
Structural Bioinformatics

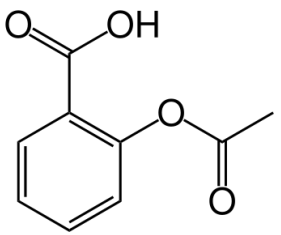


To understand what is really going on in biology you need the 3D structure of the macromolecules, *i.e.* the proteins in particular!

ACACACTGG CTTGGACTCAACCTGATGGGCTTCTGGGCCCAGCCCCAGACAAACCCCCGGCAAACGTC
CCATTCCGAGGA GCATGAGCAGATGGAGTATGGAAGAAATGCCCAAGACGGCAGGCAGCAGCTGTGGC
GGCCGGCGGGACGACAATCCGAGGAGAGGCCTCTGATGTCCTGAGGTCTCAGAGGACGCCTAAAGGCCTT
GAATGGGACAAGCTTAGCGGGCGGGCGCAGAAGAGAATAATACTCTGGAGACACTTCCCGAGGGCTCTGG
GGCCGGAGCTGTGTTCGCTCCGGTTCTTGGTGAAGACAGGGTTCGTGGGAGGCGGCCCAAGGAGGGCGAA
CGCCTAAGACTGCAAAGGCTCGGGGGAGAACGGCTCTCGGAGAACGGGCTGGGGAAGGACGTGGCTCTGA
AGACGGACAGCCCTGAGGAACCGCGGGGCGCCCAGATGGAACTCGTTAGCGCCCCGAGTGCAGACAATCC
CGGAGGGGGAAAGGCGAGCAGCTGGCAGAGAGCCCAGTGCCGC AACCGCGCGAGCGCCTCAGAACGGC

Neuraminidase is a glycoside hydrolase enzyme found on the surface of the influenza virus





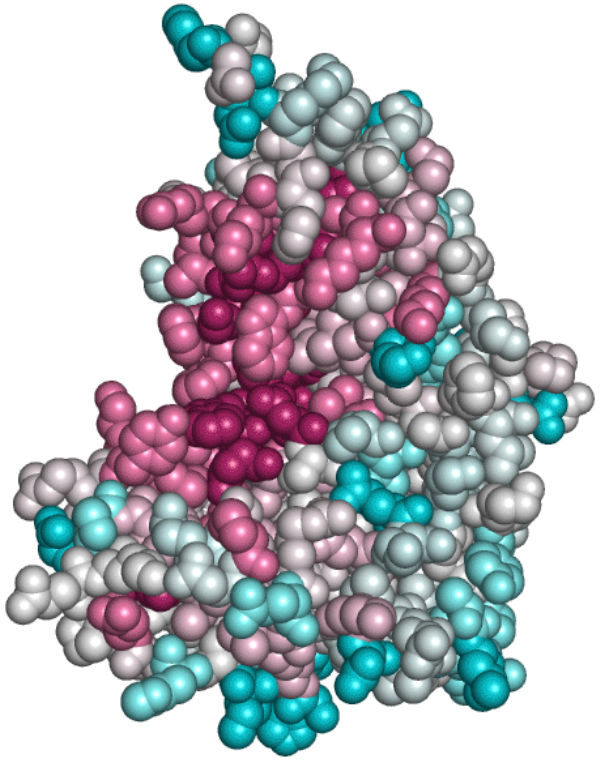
Structural bioinformatics

Experimental methods for determining protein 3D structure are very expensive in terms of money and time

Alternative: Use computational methods to determine protein structure

- Determine 3D structure with computers
- Determine secondary structure, structural disorder, domain boundaries, sites of post-translational modifications (PTMs), etc.
- Understand structure through computations
- Work with 3D structures, compare, classify, etc.

- **Goal is to get biological insight!**



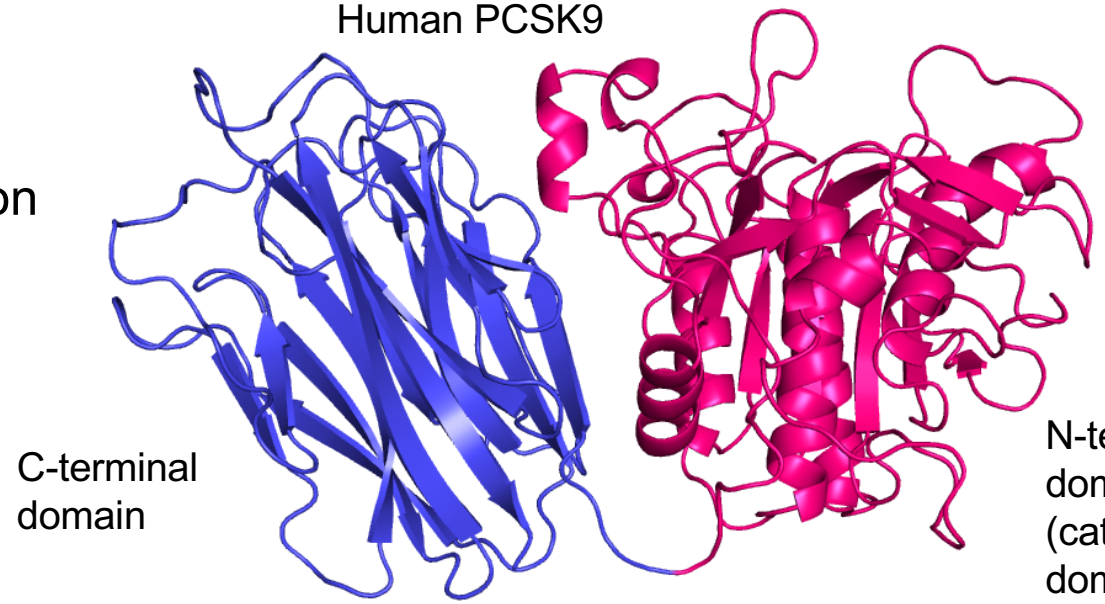
Protein domains

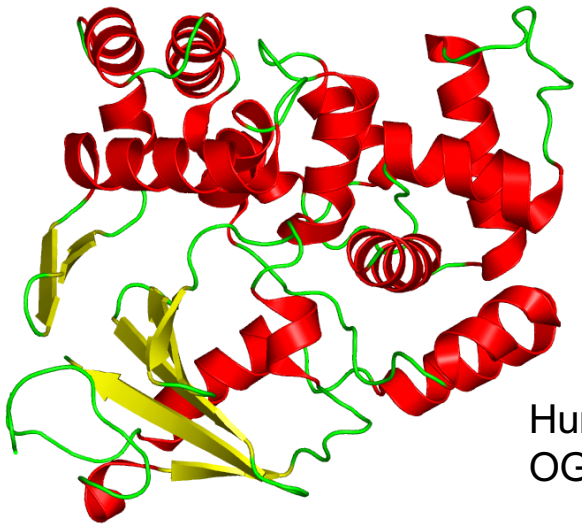
Jon K. Lærdahl,
Structural Bioinformatics

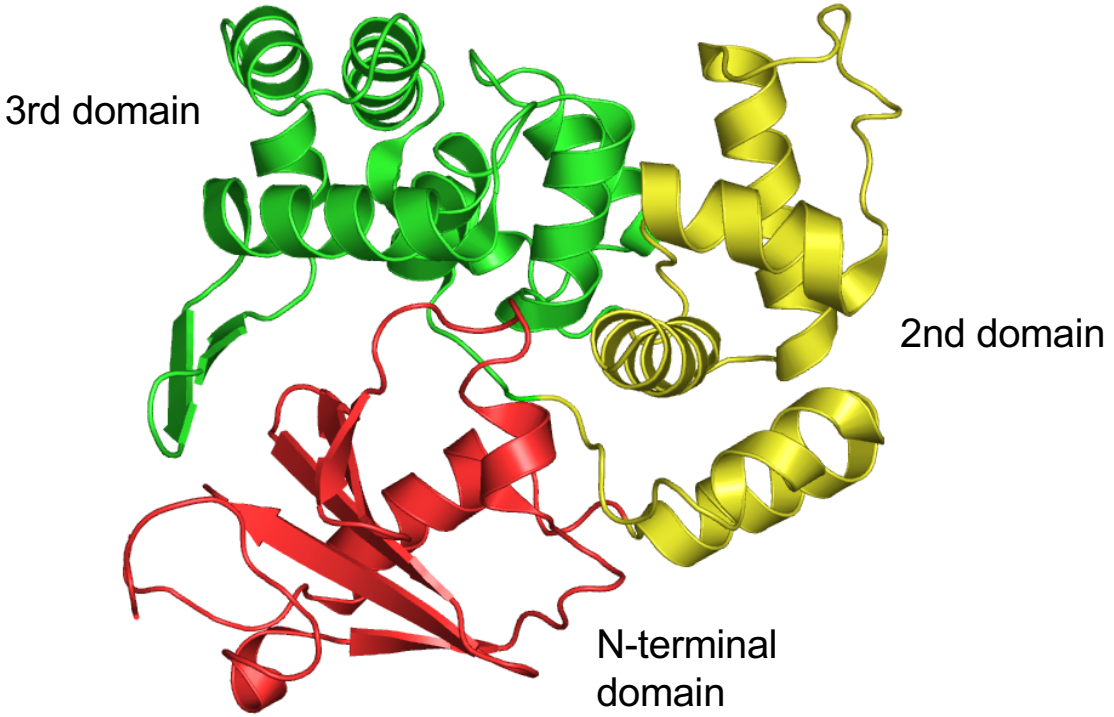
Domain: Compact part of a protein that represents a structurally independent region

Domains are often separate functional units that may be studied separately

Domains fold independently? Not always…



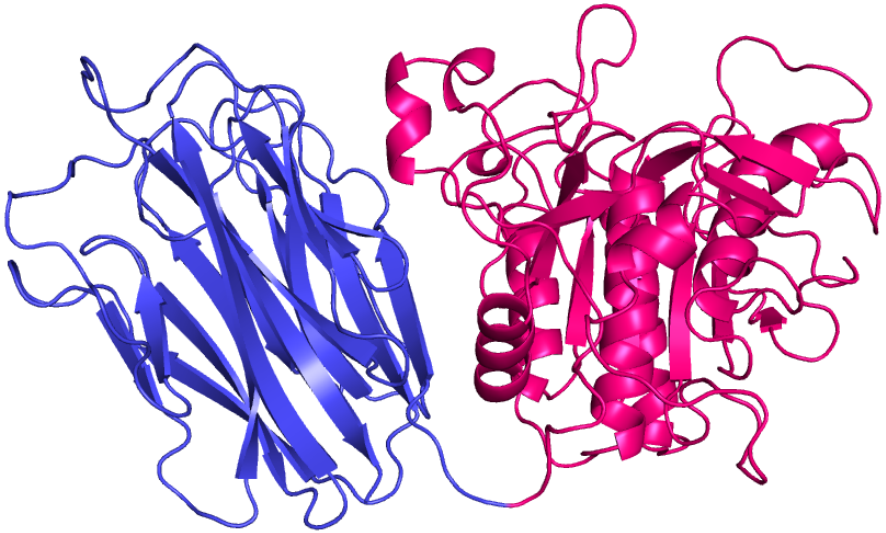




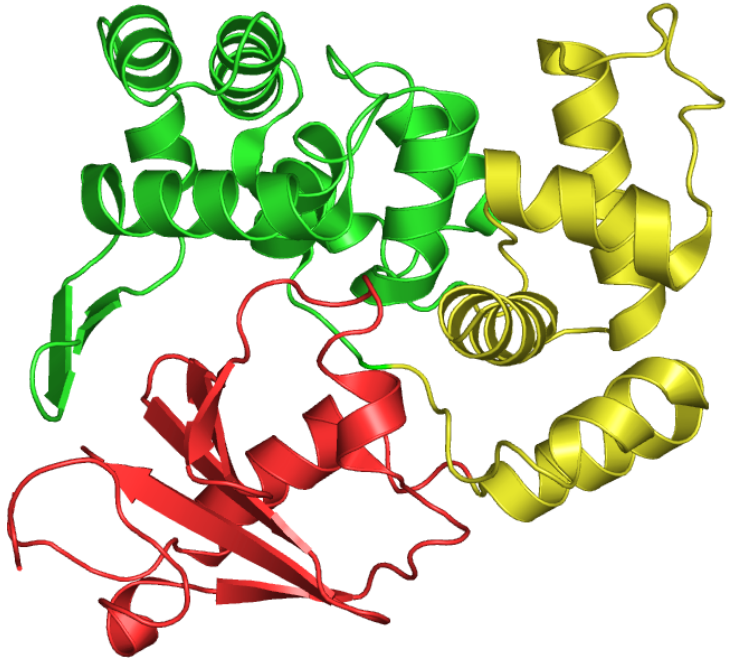
Protein domains

Jon K. Lærdahl,
Structural Bioinformatics

Dividing a protein structure into domains: no "right way to do it" or "correct algorithm", *i.e.* ***a lot of subjectivity involved***



Most people would agree there are two domains here



Three domains?
One domain?
Two?

SCOP vs.
CATH?

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

Comparing structures

Jon K. Lærdahl,
Structural Bioinformatics

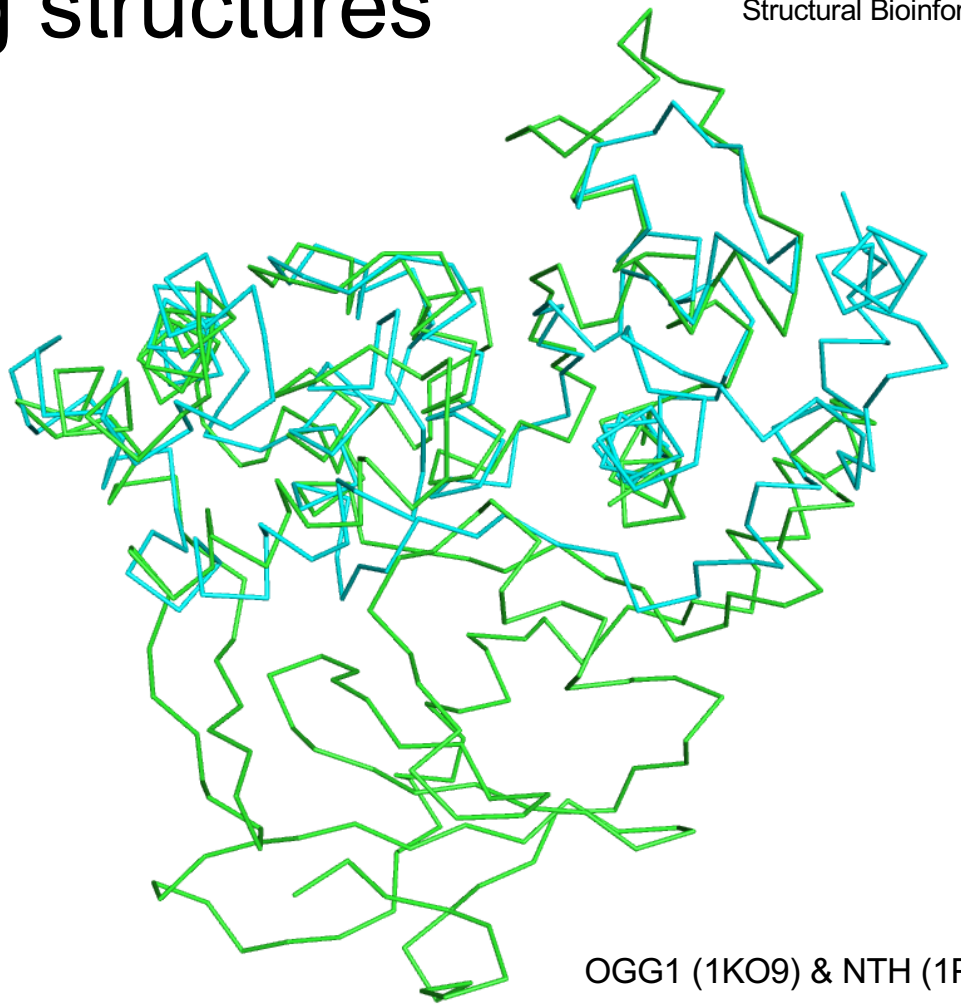
Superimposition/Alignment of 3D structures in space

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)

In many cases we align domains, not full length proteins

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

Comparison of protein structures

Brief demo in PyMOL!

```
>align molecule1, molecule2, object=matches
```

Human NEIL1

E. coli endonuclease VIII

Aligned with RMSD = 1.41 Å

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

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

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

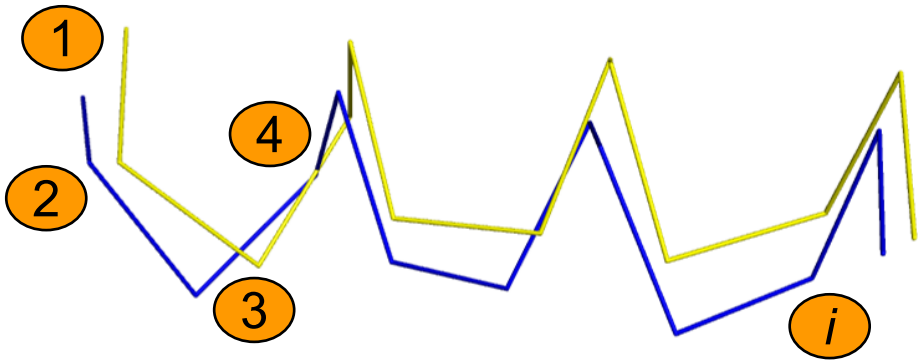
Calculate RMSD by:
Loop over equivalent positions *i*
- Get coordinates for both Cαs
- Calculate distance beetween Cαs, δ_i
- Square δ_i and add to sum

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

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

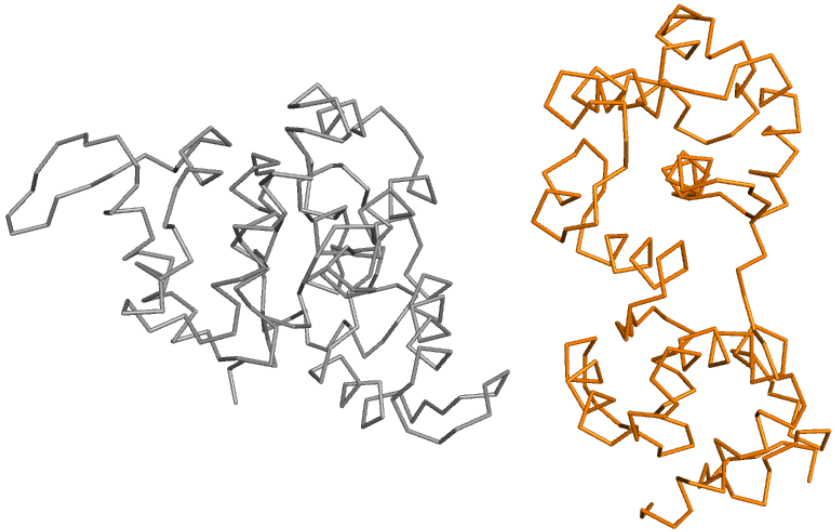
RMSD tells you how similar two structures are

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



Comparing structures - Intermolecular method

Jon K. Lærdahl,
Structural Bioinformatics



Translation

Rotation

- 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

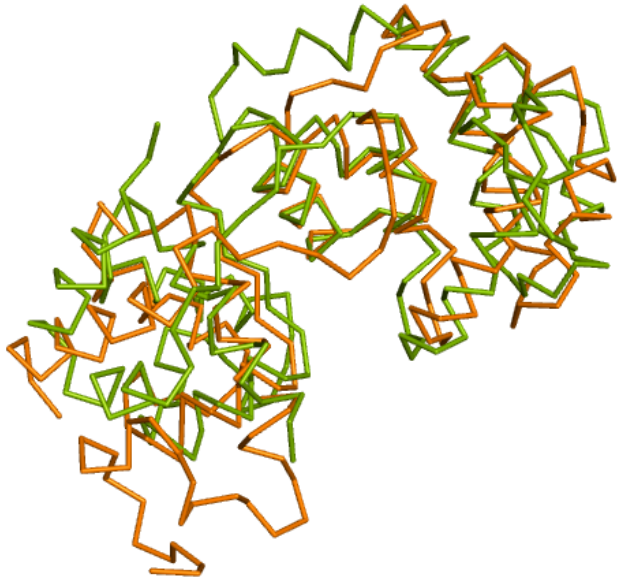
Problems with Intermolecular method:

- 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

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 and 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

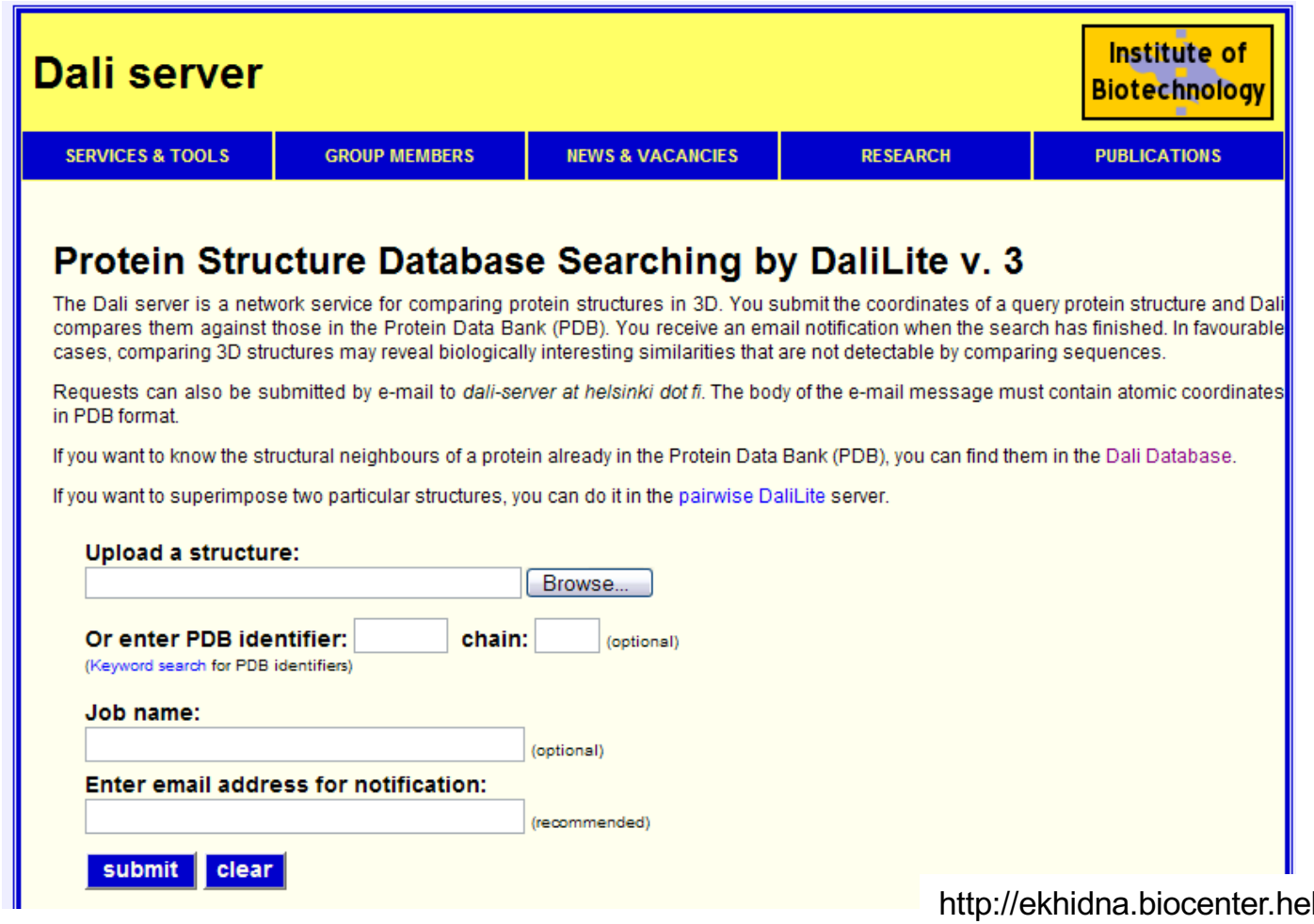
Comparing structures – Some tools:

Jon K. Lærdahl,
Structural Bioinformatics

STAMP (http://www.compbio.dundee.ac.uk/downloads/stamp): 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

Comparing structures – Some tools:

Jon K. Lærdahl,
Structural Bioinformatics

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

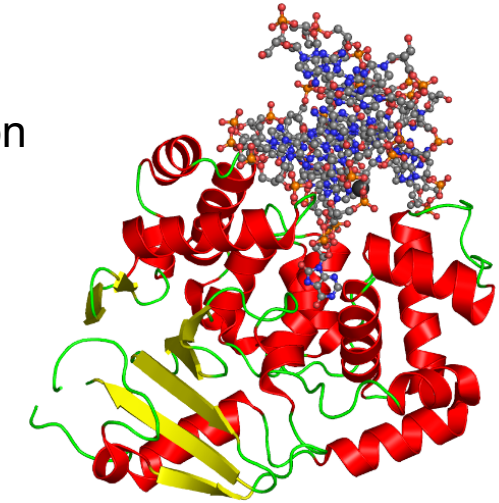
Alternatives:
VAST+ (at NCBI)
CE
SSAP
More…

CEAlign demo

Jon K. Lærdahl,
Structural Bioinformatics

Protein structure evolution

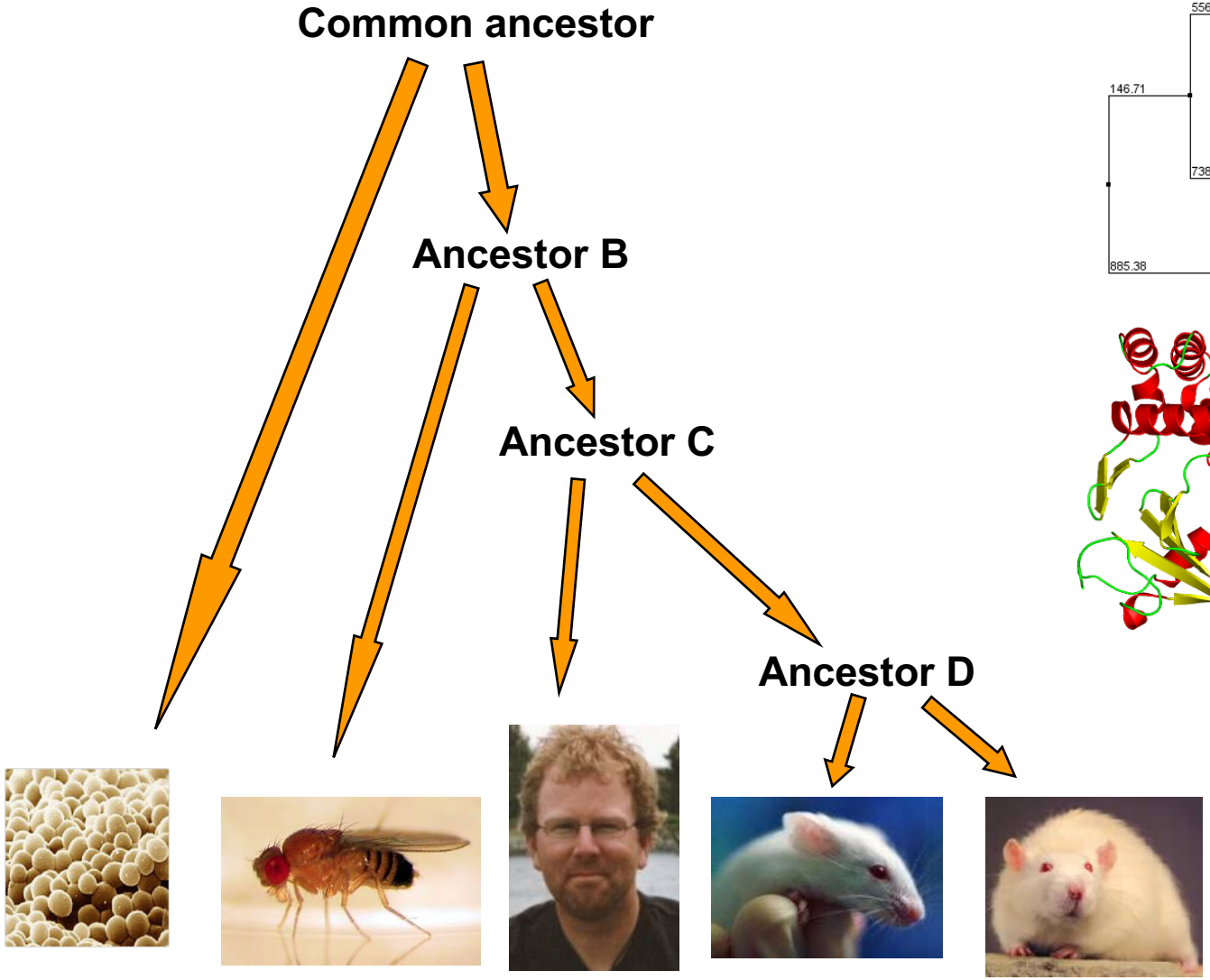
- The origin of this gene/protein is (very likely) before the last common ancestor of *S. cerevisiae* (yeast), human, mouse, rat, and fruit fly
- Some of the amino acids have not mutated in >1 billion years
- Neutral mutation rate in mammals is ~0.01 base pair/5 million yr

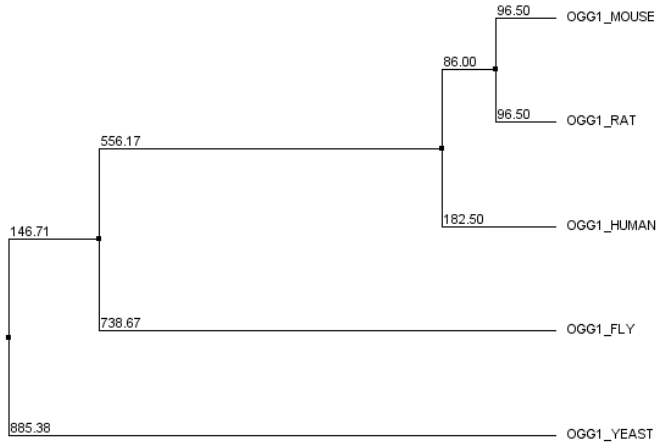


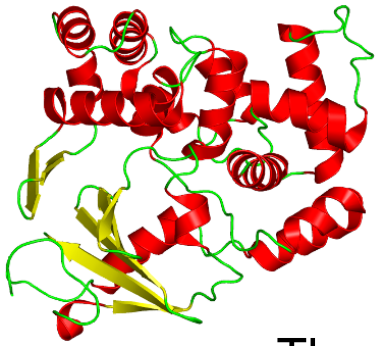
```
OGG1_YEAST/1-376     1 -MSYK------------------FGKLAINKSELCLANVLQAGQSFRW---------IWDEKLNQYSTTMKIGQQEKYSVVILRQDEENEILEFVAVGDCGNQ 75
OGG1_MOUSE/1-345     1 -MLFRSWLPSSMRHRTLSSSPALWASIPCPRSELRLDLVLASGQSFRWK---EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDDSQVSRPTLEEL------ 93
OGG1_RAT/1-345       1 -MLFSSSLSSSMRHRTLTSSPALWASIPCPRSELRLDLVLASGQSFRWR---EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDKGQVGRPTLEEL------ 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-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

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

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics

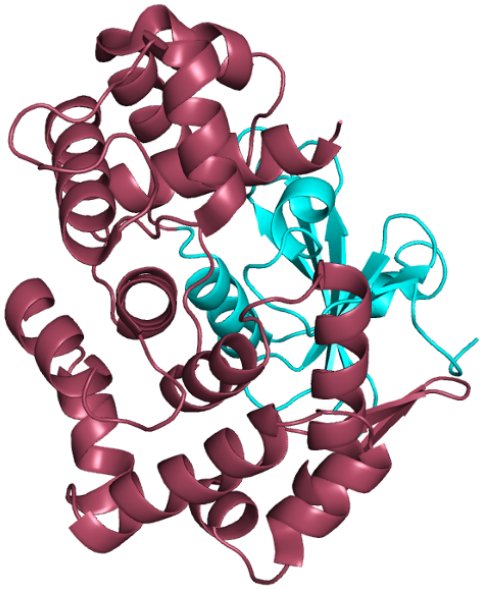
Proteins that fold in the same way, i.e. ”have the same fold” are often homologs.

Structure evolves slower than sequence

Sequence is less conserved than structure

EndoIII (2ABK):

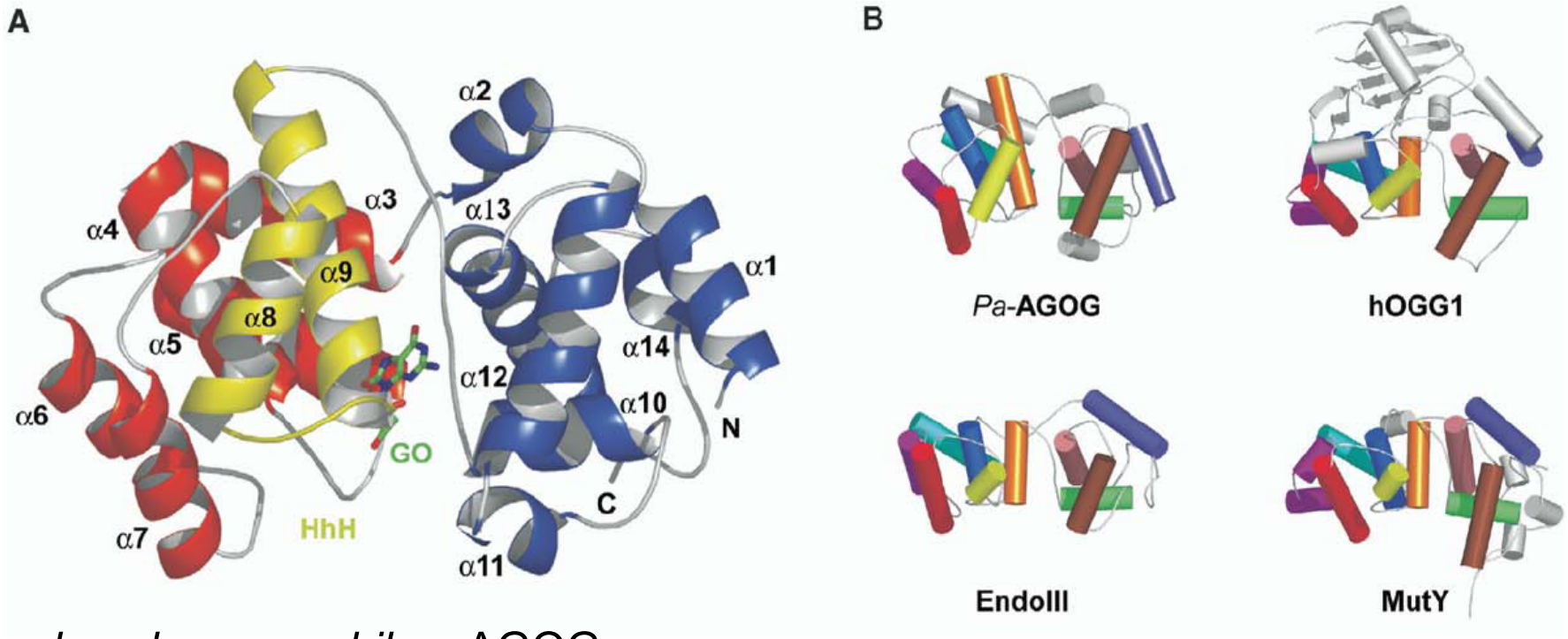
OGG1 (1EBM):



```
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

Last common ancestor (2 Gyrs?)

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics

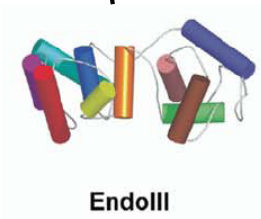
Similar structure
No sequence similarity

Fairly similar structure
Some sequence similarity

Fairly similar structure
Some sequence similarity

Very similar structure
Significant sequence similarity

Pa-AGOG



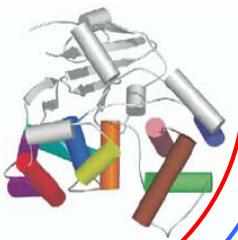
EndoIII

MutY

AlkA

Yeast Ogg1

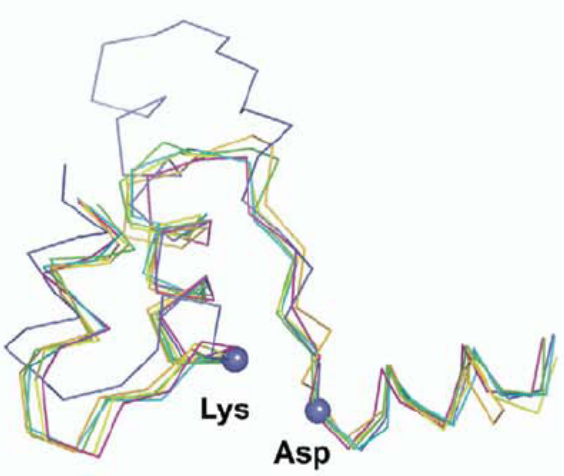
Human Ogg1

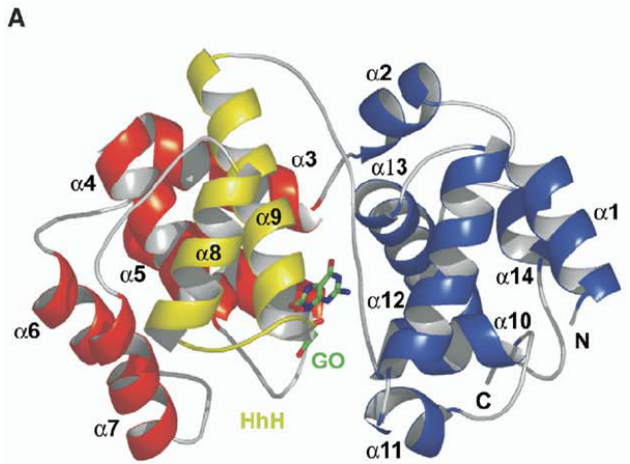


Mouse Ogg1

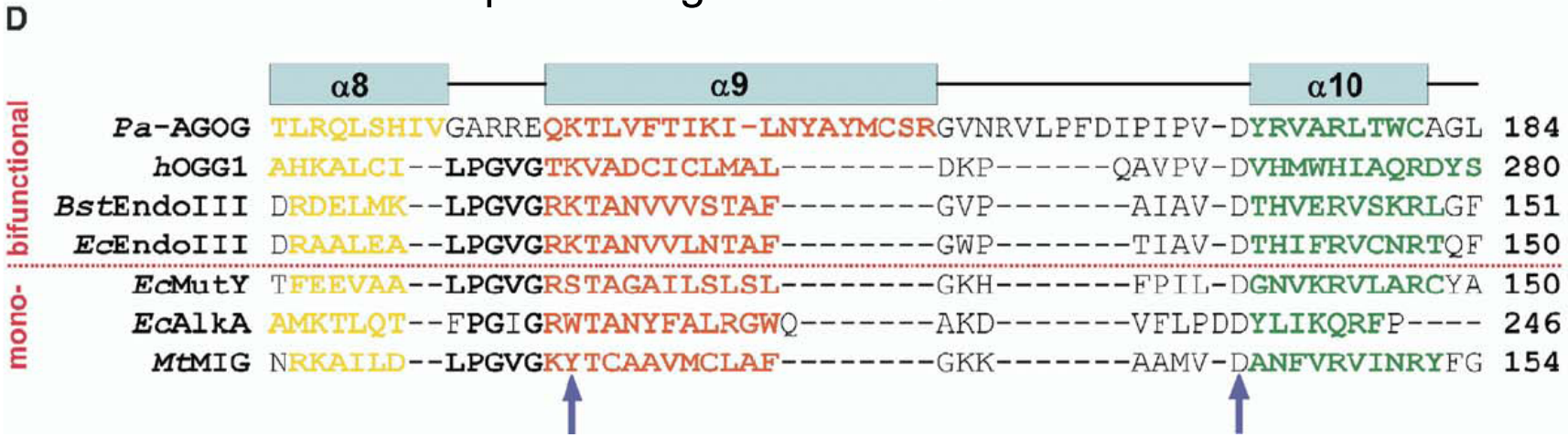
Protein structure evolution

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

Protein structure alignments

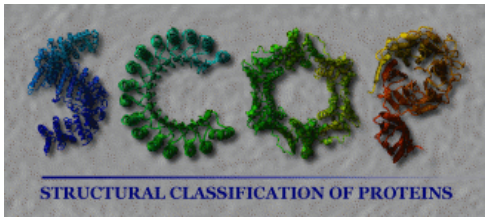
Proteins that fold in the same way, i.e. ”have the same fold” are often homologs.
Structure evolves slower than sequence
Sequence is less conserved than structure

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

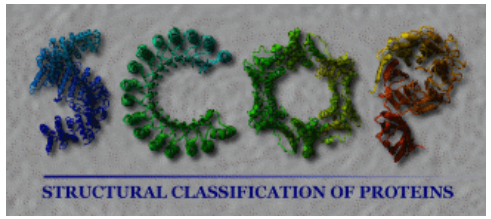
- For example using DALI
- Much more sensitive than sequence search
- Problems
  - Much smaller database (PDB vs. Genbank)
  - Need 3D structure of protein

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



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

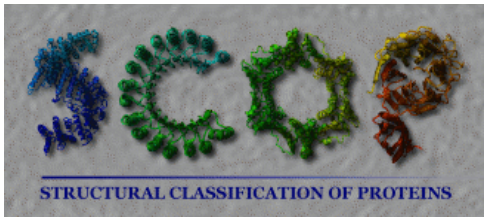
- Almost 100% manually generated
- Proteins grouped into hierarchy of classes, folds, superfamilies and families

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