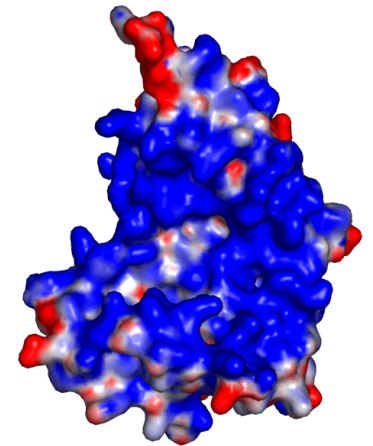
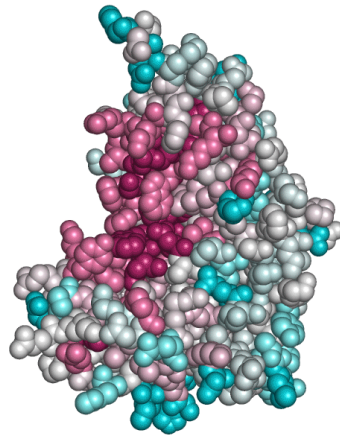
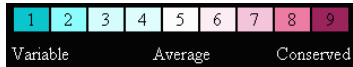


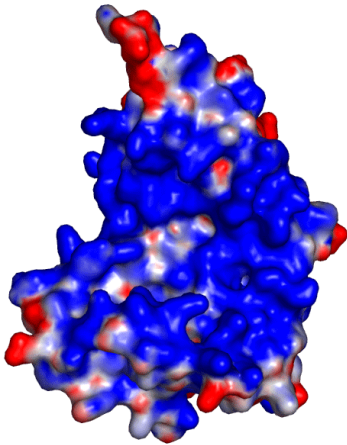
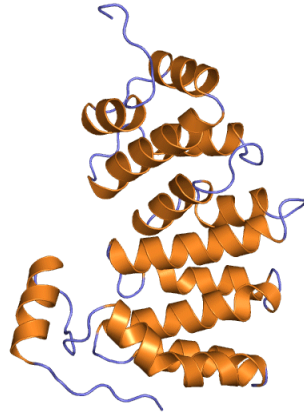
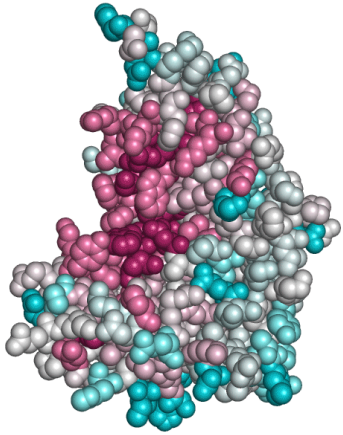
Bioinformatics for molecular biology

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

Structural bioinformatics



Approx.
3 days

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

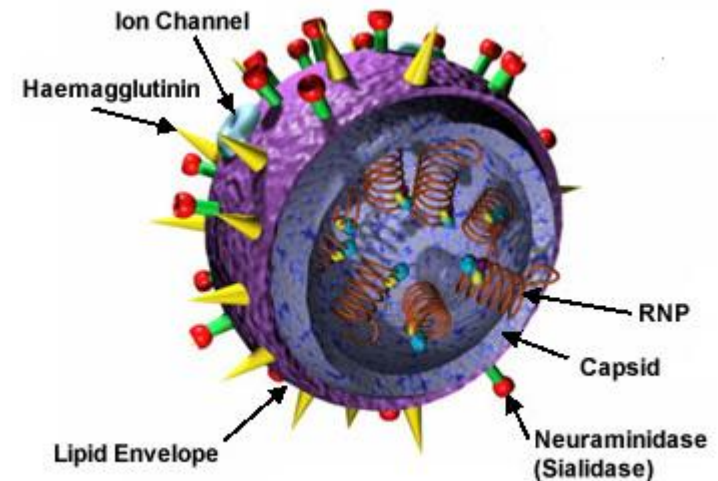
Very expensive (time and money) and difficult to determine 3D structure by experimental methods

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

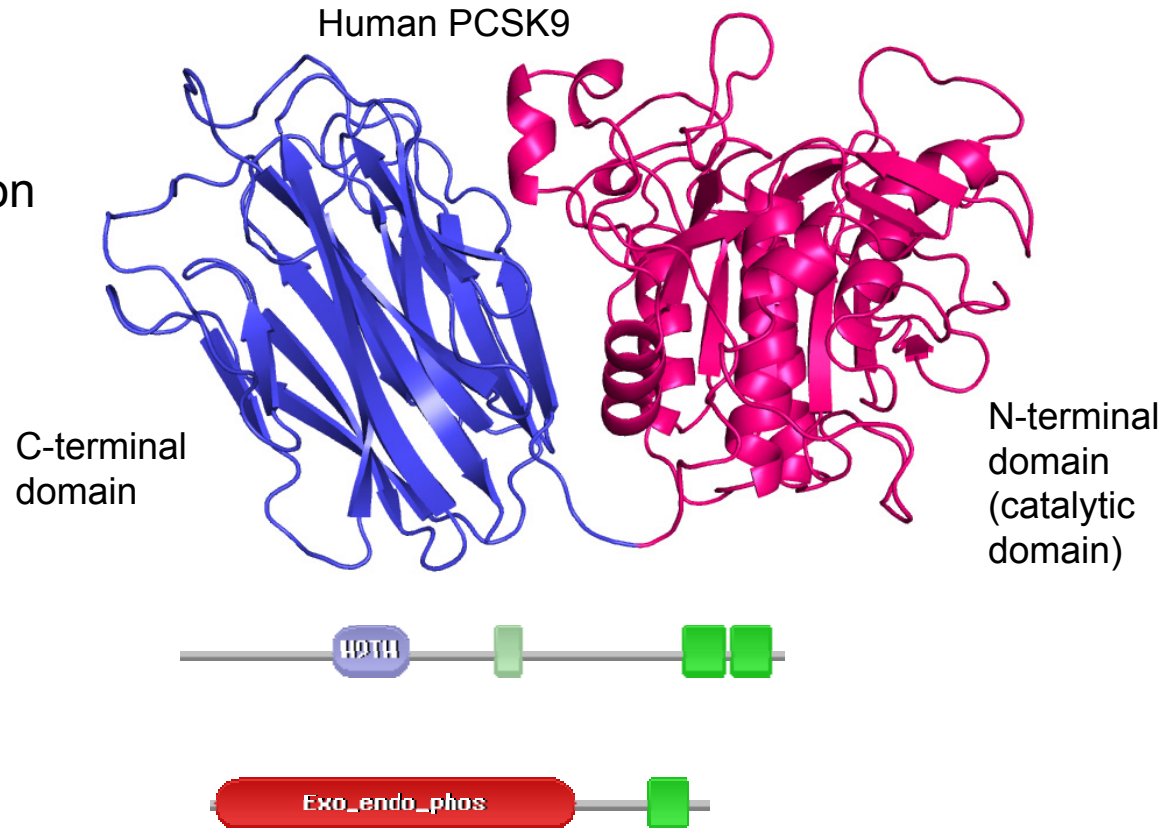
Domains are often separate functional units that may be studied separately

- Structure
- Function

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

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

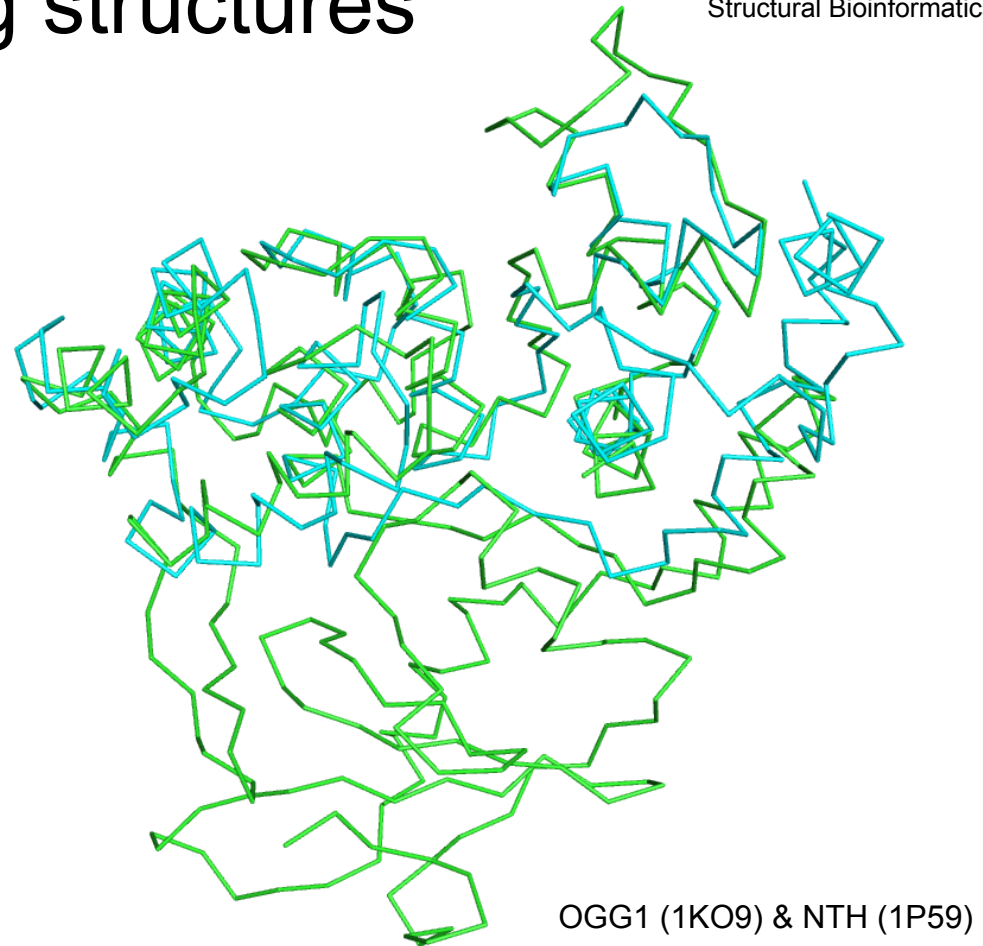
Superimposition/Alignment of 3D structures in space

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 C α atoms!)

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



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

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

Calculate RMSD by:

Loop over equivalent positions i

Get coordinates for both C α s

Calculate distance between C α s, δ_i

Square δ_i and add to sum

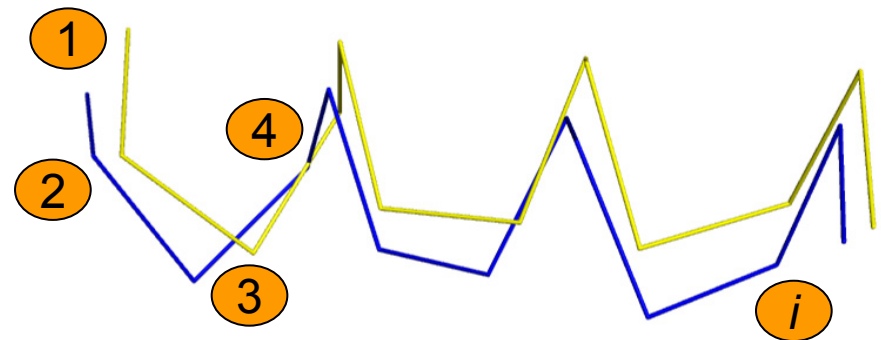
End loop

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

Comparison of protein structures

Human NEIL1



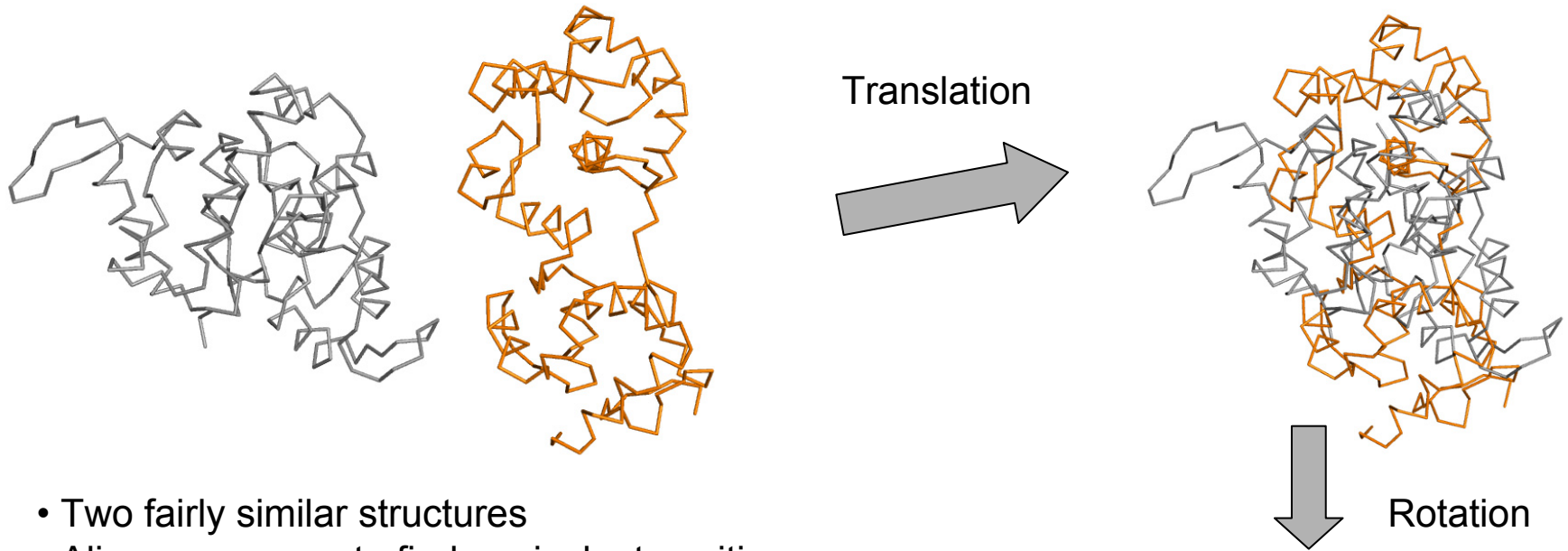
Aligned with RMSD = 1.41 Å



E. coli endonuclease VIII

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

Comparing structures - Intermolecular method



- 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 C α atoms only)

ASTPALWAS I PCRSELRLDLVLP SGQSFWR EQS PAHWSGVLADQWV
SARMLTRSRLGPGAGPRGCREEPG - - PLRRREAAAE - - - - -

ASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEV AQKFQGVRLI
RLRVAYEGSDSEK G - - - - EGAEP LKVP WVEPQD - - WQQQLVNI RAMRNI

ARITGMVERLCQAFGPRLIQLD DVTYHGFP SLQALAGPEVEAHLR - - -
PKVRRYQVLLSLMLSSQ - - TKDQVTAGAMQRLRA - RGLTVDSILQTDD,

LEEQGLAWLQQLRESSYEEAHKALC I L PGVGT K VADC I CLMALDK PQ,
KQTS - - - A I L QCHYGGD I PASVAELVALPGVGPKMAH LAMAVAVGT VSI

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



Comparing structures - Intermolecular method

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

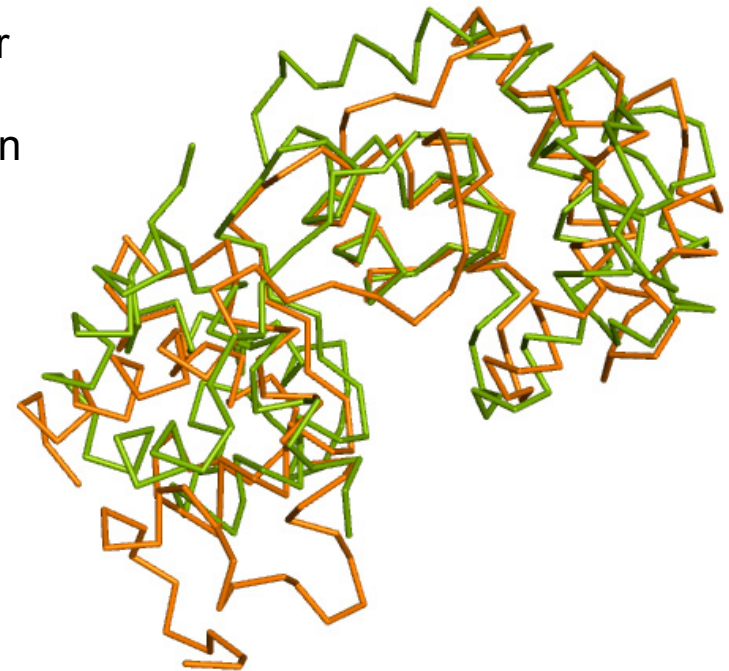


Important!

```
GGLAWLQQLRESSYEEAHKALC ILPGVGTKVADC ICLMALDKPQ  
---AILQOHYGGDIPASVAELVALPGVGPKMAHLAMAVANGT VSI
```

Comparing structures - Intramolecular method

- 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 C α 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

The screenshot shows the Dali server website. At the top, there is a yellow header with the text "Dali server" on the left and the "Institute of Biotechnology" logo on the right. Below the header is a blue navigation bar with five tabs: "SERVICES & TOOLS", "GROUP MEMBERS", "NEWS & VACANCIES", "RESEARCH", and "PUBLICATIONS". The main content area has a yellow background and is titled "Protein Structure Database Searching by DaliLite v. 3". It contains several paragraphs of text explaining the service, including instructions on how to submit requests and find structural neighbors. At the bottom, there is a form with the following fields: "Upload a structure:" with a text input and a "Browse..." button; "Or enter PDB identifier:" with a text input and "chain:" with a text input (optional); "Job name:" with a text input (optional); and "Enter email address for notification:" with a text input (recommended). At the very bottom of the form are two buttons: "submit" and "clear".

Intramolecular
method

Liisa Holm
(Finland)

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

- [Parseable data](#)
- Matches to [PDB90](#)

The match list is truncated at 500 hits.

Dali

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

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.

Expand gaps

Summary

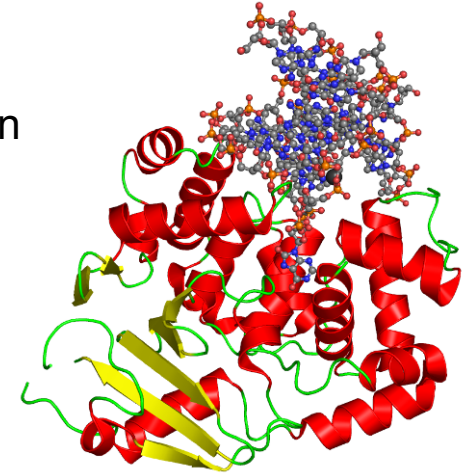
No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
<input type="checkbox"/> 1:	1ebm-A	99.9	0.0	0	314	0	PDB	MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
<input type="checkbox"/> 2:	1n3a-A	51.9	0.0	0	314	0	PDB	MOLECULE: N-GLYCOSYLASE/DNA LYASE;
<input type="checkbox"/> 3:	1m3q-A	51.8	0.0	0	314	0	PDB	MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
<input type="checkbox"/> 4:	1lww-A	51.8	0.0	0	314	0	PDB	MOLECULE: 8-OXOGUANINE DNA GLYCOSYLASE;
<input type="checkbox"/> 13:	2jhj-B	24.3	0.0	0	291	0	PDB	MOLECULE: 3-METHYLADENINE DNA-GLYCOSYLASE;
<input type="checkbox"/> 14:	2jhn-A	23.9	0.0	0	293	0	PDB	MOLECULE: 3-METHYLADENINE DNA-GLYCOSYLASE;
<input type="checkbox"/> 15:	3d4v-D	22.9	0.0	0	281	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
<input type="checkbox"/> 16:	3d4v-B	22.8	0.0	0	282	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
<input type="checkbox"/> 17:	3cwu-D	22.8	0.0	0	282	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
<input type="checkbox"/> 18:	3cwt-D	22.8	0.0	0	281	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
<input type="checkbox"/> 19:	3cvs-D	22.7	0.0	0	282	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSYLASE 2;
<input type="checkbox"/> 38:	1pu8-A	14.7	0.0	0	215	0	PDB	MOLECULE: 3-METHYLADENINE DNA GLYCOSYLASE;
<input type="checkbox"/> 39:	2h56-B	14.1	0.0	0	217	0	PDB	MOLECULE: DNA-3-METHYLADENINE GLYCOSIDASE;
<input type="checkbox"/> 40:	2abk	14.0	0.0	0	211	0	PDB	MOLECULE: ENDONUCLEASE III;
<input type="checkbox"/> 41:	1rrt-A	14.0	0.0	0	346	0	PDB	MOLECULE: MUTY;
<input type="checkbox"/> 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
<input type="checkbox"/> 43:	1kea-A	13.5	0.0	0	217	0	PDB	MOLECULE: POSSIBLE G-T MISMATCHES REPAIR ENZYME;
<input type="checkbox"/> 44:	1kq6-A	13.0	0.0	0	224	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
<input type="checkbox"/> 45:	1kq5-A	12.9	0.0	0	225	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
<input type="checkbox"/> 46:	1kq3-A	12.8	0.0	0	224	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
<input type="checkbox"/> 47:	1kqj-A	12.7	0.0	0	225	0	PDB	MOLECULE: A/G-SPECIFIC ADENINE GLYCOSYLASE;
<input type="checkbox"/> 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...

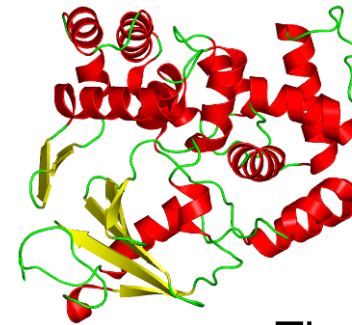
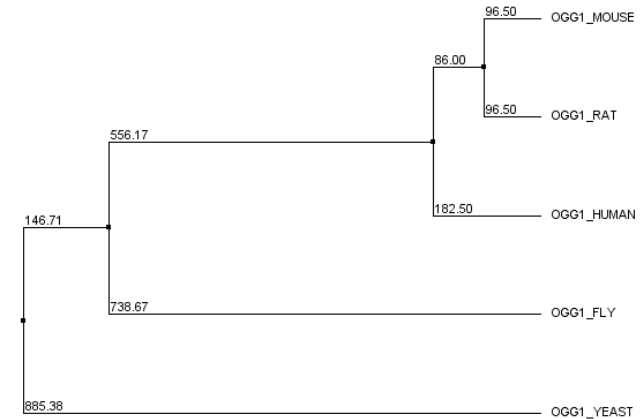
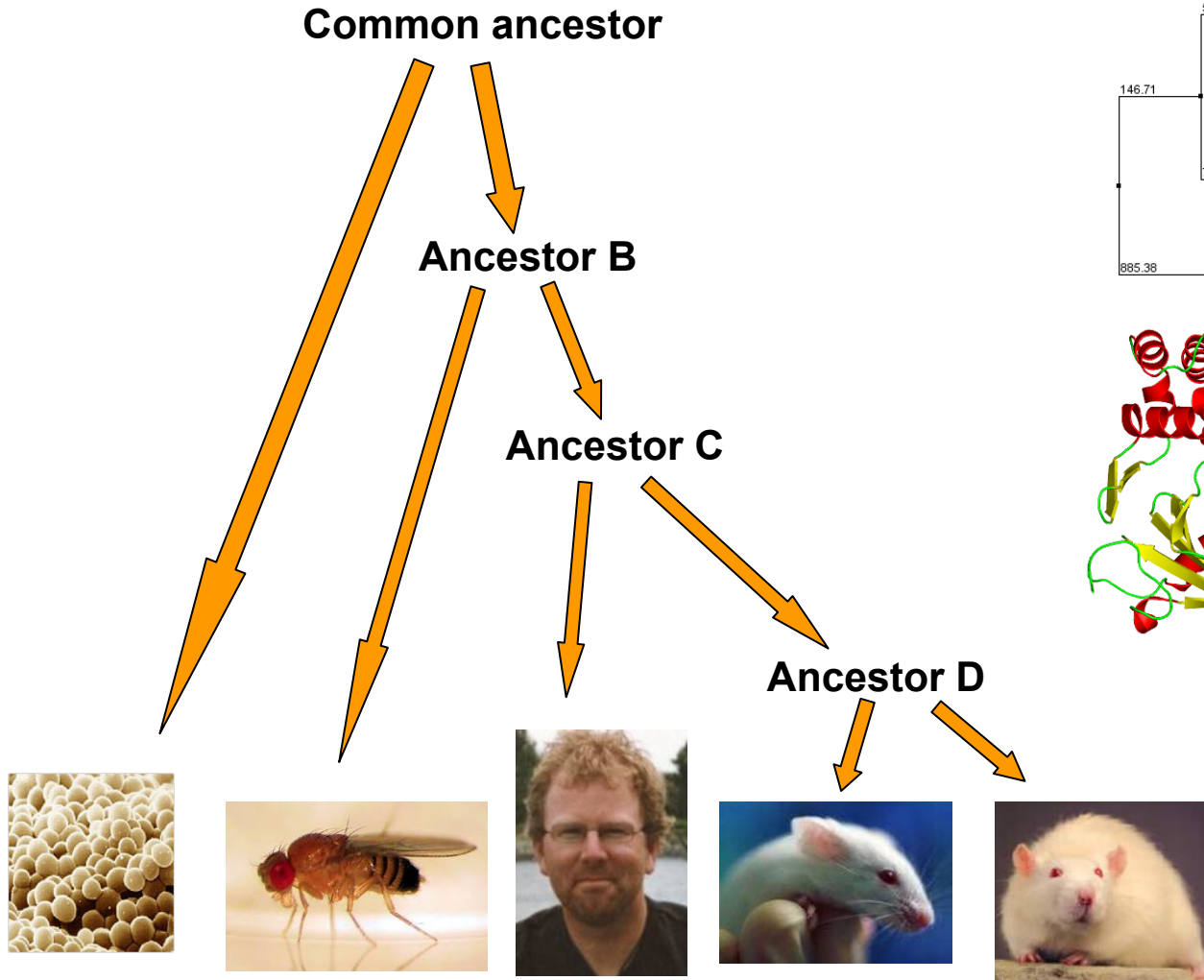
Protein structure evolution

- 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-----FGKLAINKSELCLANVLQAGQSFVW-----IVDEKLNQYSTTMTKIGQQEKYSVVILRQDEENEILEFVAVGDCQNG75
OGG1_MOUSE/1-345	1 -MLFRSWLPSSMRHRTLSSSPALWASIPCPRSELRLDLVLAAGQSFVW-----EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDDSOVSRPTLEEL-----93
OGG1_RAT/1-345	1 -MLFSSLSSSMRHRTLSSSPALWASIPCPRSELRLDLVLAAGQSFVW-----EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDKGOVSRPTLEEL-----93
OGG1_HUMAN/1-345	1 -MPARALLPRMGHRTLASTPALWASIPCPRSELRLDLVLAAGQSFVW-----EQSPAHWSGVLADQWWTLTQTEDQLYCTVYRGDKSQASRPTPEEL-----93
OGG1_FLY/1-343	1 MLAHNLGFHKKRLFNSNMKAVLQDRGVILGLSLEECDLERTLLGGQSFVWRSICDGNRTKYGGVVFNTWWLQQEESFITYEAY-GTSSPLATKDYSSL-----96
OGG1_YEAST/1-376	76 DALKTHLMKYFRLDVSLKHLFDNWVIPSDFKAFKLSR--GGIRILAGEPWETLISFICSSNNNISRITRMNSLCSNFGNLIITIDGVAYHSFPTS---EELT173
OGG1_MOUSE/1-345	94 ---ETLHKYFQLDVSLLAQLYSH-WASVDSHFQVVAQKFOGVRLLRQDPTCELFISFICSSNNN IARITGMVERLCCAFGPRLIQLDVVTYHGFPNL---HALA188
OGG1_RAT/1-345	94 ---ETLHKYFQLDVSLLAQLYSH-WASVDSHFQVVAQKFOGVRLLRQDPTCELFISFICSSNNN IARITGMVERLCCAFGPRLVQLDDVTYHGFPNL---HALA188
OGG1_HUMAN/1-345	94 ---EAVRKYFQLDVTLLAQLYHH-WGSVDSHFQVVAQKFOGVRLLRQDPIECLFSFICSSNNN IARITGMVERLCCAFGPRLIQLDVVTYHGFPNL---QALA188
OGG1_FLY/1-343	97 -----ISDYLRVDFDLKVNQKD-WLSKDDNFVKFLS--KPVRLLSGEPFENIFSFLLCSQNNN IKRISSMIEWFCATFGTKIGHFNGADAYTFPTINRFHDI190
OGG1_YEAST/1-376	174 SRAIEAKLRELGFGRAYKIIEIETARKLVNDKAEANITSDTTYLOSICKDAQYEDVREHLMSYNGVGPKVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISAN276
OGG1_MOUSE/1-345	189 GPEAETHLRKLGGRARYVRSASAKAILEEQGGP-----AWLQQLRV-APYEEAHKALCTLPGVGAQVADCICLMALDKPQAVPVDVHVVQIAHRDYGAPK284
OGG1_RAT/1-345	189 GPEVETHLRKLGGRARYVCSASAKAILEEQGGP-----AWLQQLRV-ASYEEAHKALCTLPGVGTQVADCICLMALDKPQAVPVDIHWVQIAHRDYGAPK284
OGG1_HUMAN/1-345	189 GPEVEAHLRKLGLGRARYVSASARAILLEEQGGL-----AWLQQLRE-SSYEEAHKALCILPGVGTQVADCICLMALDKPQAVPVDVHMAHIAQRDYSVHPT284
OGG1_FLY/1-343	191 CEDLNAQLRAAKFGYRAKFAIQTLEIQKKGQ-----NWFISLKS-MPFEKAREELTLLPGIGYKQVADCICLMSMHLESVPVDIHYRIAQNYLPHLT285
OGG1_YEAST/1-376	277 -KNHLRELRTKYNALPISRKKINLELDHIRLMLFKKWSYAGWAQGVLFSSKEIGGTSGSTTTGTIKKRKWDMIKETEAVITVKQMKLKVLSDLHIKEAKID376
OGG1_MOUSE/1-345	285 -TSQAKGPS-----PLANKELG---NFFRNLD---WGPYAGWAQAVLFSADLROPS-LSREPPAKRK-----KGSKRPEG---345
OGG1_RAT/1-345	285 -TSQTKGPS-----PLANKELG---NFFRNLD---WGPYAGWAQAVLFSADLROQN-LSREPPAKRK-----KGSKKTGEG---345
OGG1_HUMAN/1-345	285 -TSQAKGPS-----PQTNKELG---NFFRSLD---WGPYAGWAQAVLFSADLROSR-HAQEPPAKRR-----KGSKGPEG---345
OGG1_FLY/1-343	286 GQKNVY-----KKIYEEVS---KHFOKL---HGKYAGWAQAILFSADLSDFQ-NITSTVACKKK-----SNKPKK---343

Protein structure evolution



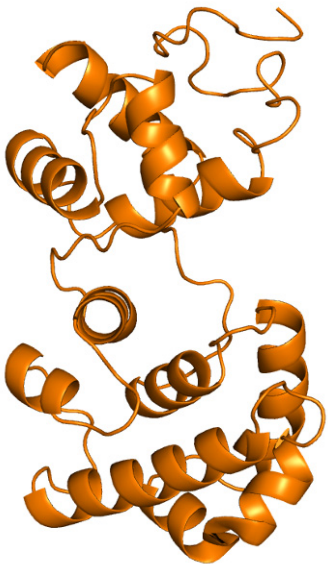
The overall structure of this protein is the same in all these organisms – i.e. many mutations does not change the structure and/or function

Protein structure evolution

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

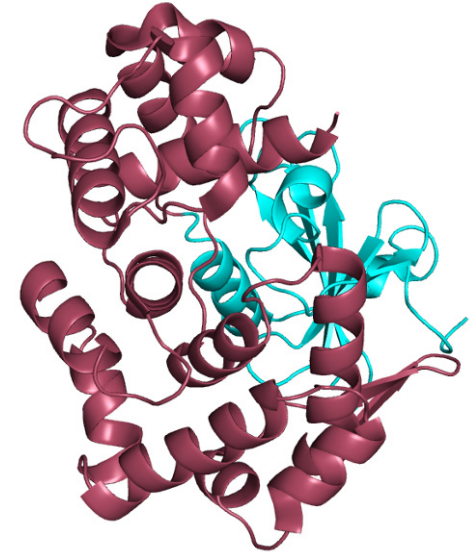
Structure evolves slower than sequence

Sequence is less conserved than structure



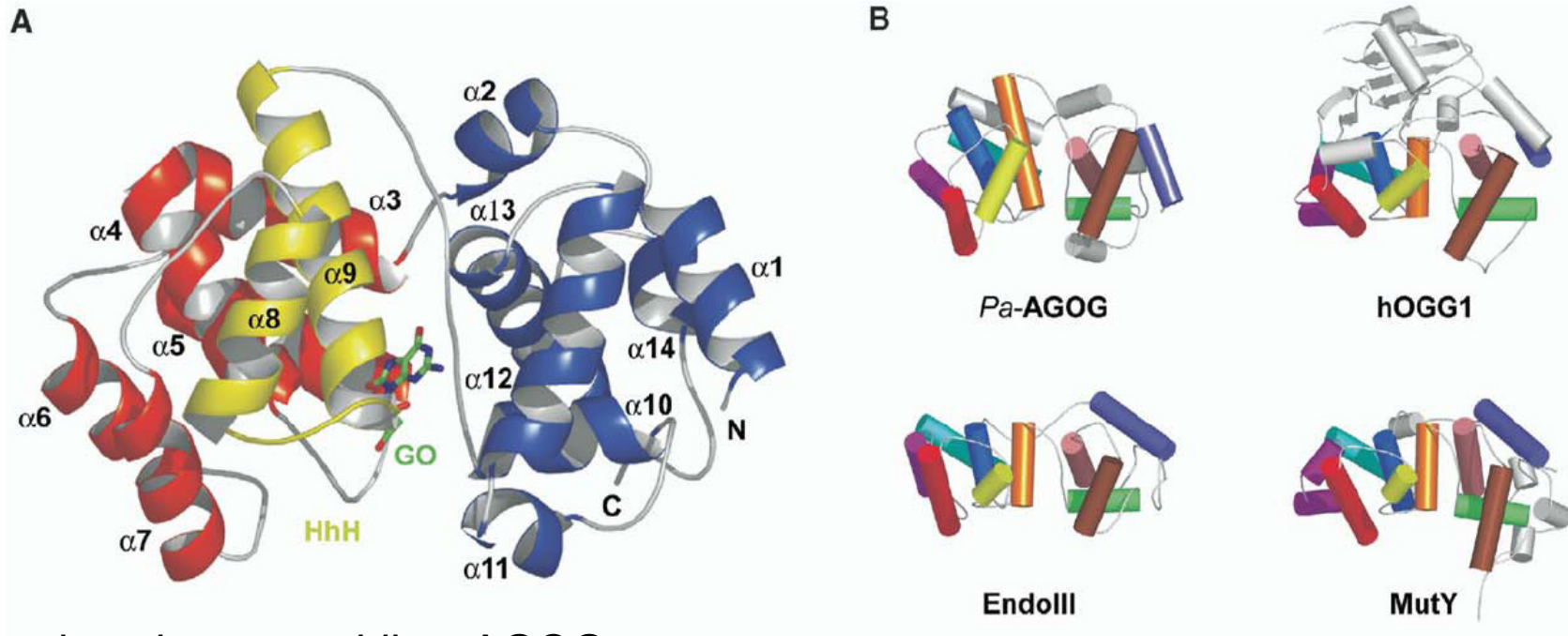
EndoIII (2ABK):

OGG1 (1EBM):



OGG1/1-345	1	M	P	A	R	A	L	L	P	R	R	M	G	H	R	T	L	A	S	T	P	A	L	W	A	S	I	P	C	P	R	S	E	L	R	L	D	L	V	L	P	S	G	Q	S	F	R	W	R	E	Q	S	P	A	H	W	S	G	V	L	A	D	Q	W	T	L	T	Q	68
NTH1/1-312	1	M	C	S	-	-	-	-	P	Q	E	S	G	M	T	A	L	S	A	R	M	L	T	R	S	R	S	L	G	P	G	A	G	P	R	G	C	R	E	E	P	G	-	-	P	L	R	R	E	A	A	A	E	-	-	-	-	-	-	-	-	-	-	47					
OGG1/1-345	69	T	E	E	Q	L	H	C	T	V	Y	R	G	D	K	S	G	A	S	R	P	T	P	D	E	L	E	A	V	R	K	Y	F	Q	L	D	V	T	L	A	Q	L	Y	H	H	W	G	S	V	D	S	H	F	Q	E	V	A	Q	K	F	Q	G	V	R	L	L	Q	D	136
NTH1/1-312	48	-	A	R	K	S	H	S	P	V	K	R	P	R	K	A	O	R	L	R	V	A	Y	E	G	S	D	S	E	K	G	-	-	-	-	E	G	A	E	P	L	K	V	P	W	E	P	Q	D	-	-	W	Q	Q	L	V	N	I	R	A	M	R	N	K	K	D	A	108	
OGG1/1-345	137	P	I	E	C	L	F	S	F	I	C	S	S	N	N	I	A	R	I	T	G	M	V	E	R	L	C	Q	A	F	G	P	R	L	I	Q	L	D	D	V	T	Y	H	G	F	P	S	L	Q	A	L	A	G	P	E	V	E	A	H	L	R	-	-	-	-	K	L	G	200
NTH1/1-312	109	P	V	D	H	L	G	T	E	H	C	Y	D	S	S	A	P	P	K	V	R	R	Y	Q	V	L	S	L	M	L	S	S	Q	-	-	T	K	D	Q	V	T	A	G	A	M	Q	R	L	R	A	-	R	G	L	T	V	D	S	I	L	Q	T	D	D	A	T	L	G	173
OGG1/1-345	201	-	L	G	Y	R	A	R	Y	V	S	A	S	A	R	A	I	L	E	Q	G	G	L	A	W	L	Q	L	R	E	S	S	Y	E	E	A	H	K	A	L	C	I	L	P	G	V	G	T	K	V	A	D	C	I	C	L	M	A	L	D	K	P	Q	A	V	P	V	267	
NTH1/1-312	174	K	L	I	Y	P	V	G	F	W	R	S	K	V	K	Y	I	K	Q	T	S	-	-	-	A	I	L	Q	H	Y	G	G	I	P	A	S	V	A	E	L	V	A	L	P	G	V	G	P	K	M	A	H	L	A	M	A	V	A	W	G	T	V	S	G	I	A	V	238	
OGG1/1-345	268	D	V	H	M	H	I	A	Q	R	D	Y	S	W	H	P	T	S	Q	A	K	G	P	S	P	Q	T	N	K	E	L	G	N	F	F	-	R	S	L	W	-	-	-	G	P	Y	A	G	W	A	Q	A	V	L	F	S	A	D	L	R	Q	S	R	H	A	Q	330		
NTH1/1-312	239	D	T	H	V	H	R	I	A	N	R	-	L	R	W	-	-	-	T	K	K	A	T	K	S	P	E	E	T	R	A	A	E	E	W	L	P	R	E	L	W	H	E	I	N	G	L	L	V	G	F	G	Q	Q	T	C	L	P	V	H	P	R	C	H	A	C	L	N	302
OGG1/1-345	331	E	P	P	A	K	R	R	K	G	S	K	G	P	E	G	345																																																				
NTH1/1-312	303	Q	A	L	C	P	A	A	Q	G	L	-	-	-	-	-	312																																																				

Protein structure evolution



Pyrobaculum aerophilum AGOG

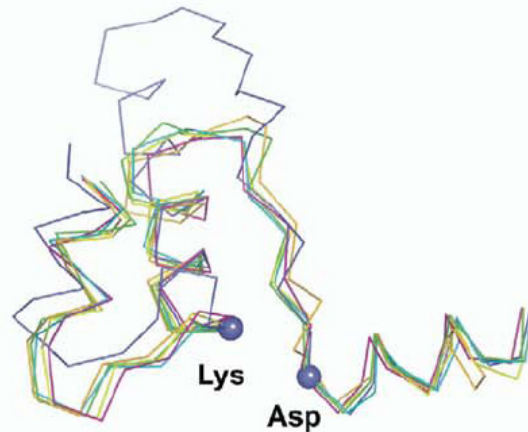
G.M. Lingaraju *et al.* *Structure*
13, 87 (2005)

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



Structure base sequence alignment:

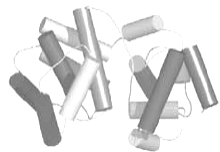
D

	$\alpha 8$	$\alpha 9$	$\alpha 10$	
<i>Pa</i> -AGOG	TLRQLSHIV	GARREQKTLVFTIKI-LNYAYMCSR	GVNRLVLPFDIPIPV-DYRVARLTWCAGL	184
<i>h</i> OGG1	AHKALCI	--LPGVGTKVADCICLMAL	-----DKP-----QAVPV-DVHMWHIAQRDYS	280
<i>Bst</i> EndoIII	DRDELTK	--LPGVGRKTANVVVSTAF	-----GVP-----AIAV-DTHVERVSKRLGF	151
<i>Ec</i> EndoIII	DRAALEA	--LPGVGRKTANVVLNTAF	-----GWP-----TIAV-DTHIFRVCNRTQF	150
<i>Ec</i> MutY	TFEEVAA	--LPGVGRSTAGAILSLSL	-----GKH-----FPIL-DGNVKRVLARCYA	150
<i>Ec</i> AlkA	AMKTLQT	--FPGIGRWTANYFALRGWQ	-----AKD-----VFLPDDYLIKQRFP	246
<i>Mt</i> MIG	NRKAILD	--LPGVGRKYTCAAVMCLAF	-----GKK-----AAMV-DANFVRVINRYFG	154

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

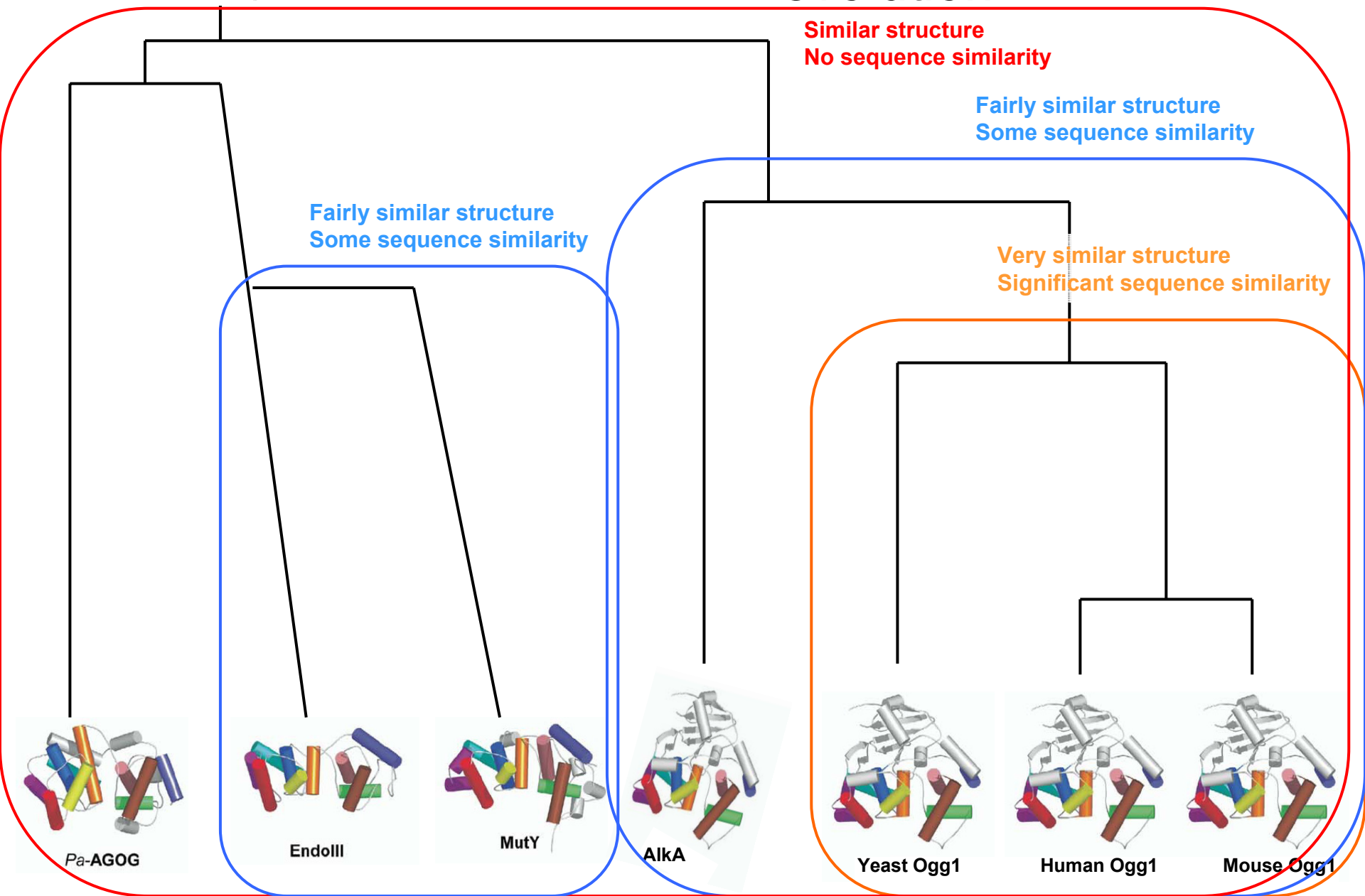
Evolution has “eroded away” sequence similarity but left the
structure intact



Last common ancestor (2 Gyrs?)

Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics



Protein structure alignments

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

Structure evolves slower than sequence

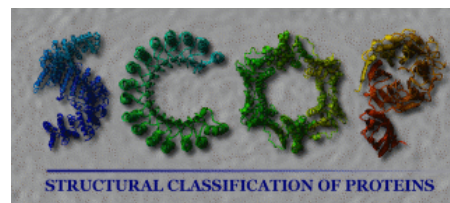
Sequence is less conserved than structure

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

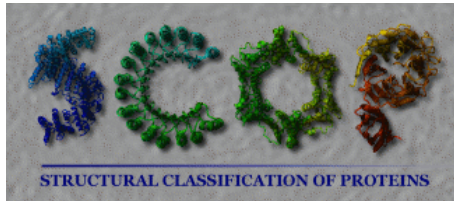
- For example using DALI
- Much more sensitive than sequence search
- Problems
 - Much smaller database (PDB vs. Genbank)
 - Need 3D structure of protein

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



Scop Classification Statistics

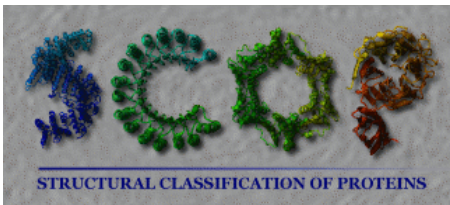
SCOP: Structural Classification of Proteins. 1.73 release
34494 PDB Entries (26 Sep 2007). 97178 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Structural Classification of Proteins

- 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

SCOP

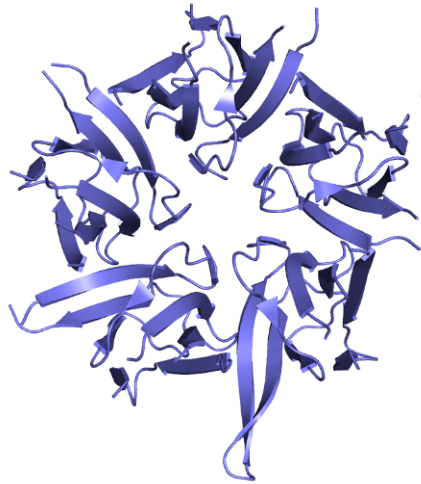
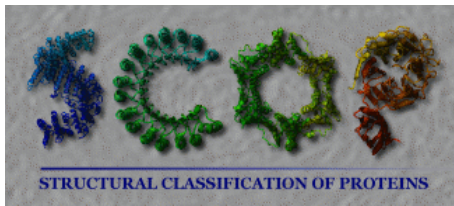


- Families
 - Sequence identity ~30% or higher
 - Very similar structures
 - Clearly homologous proteins
- Superfamilies
 - Contains families
 - May have no or little sequence similarity
 - Common fold
 - Are probably evolutionary related
- Folds
 - Contains superfamilies
 - Difficult level of classification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

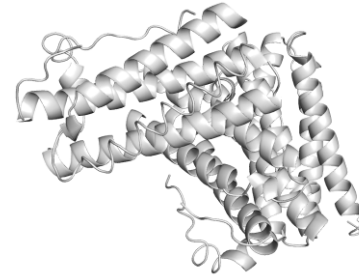
- Classes
 - Upper level of classification (4 major, 3 minor)
 - Contains folds
 - Based on secondary structure composition and “general features”
 - e.g. all- α , all- β , “membrane and cell surface” and “small proteins”
 - α/β : One β -sheet with strands connected by single α -helices
 - $\alpha+\beta$: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

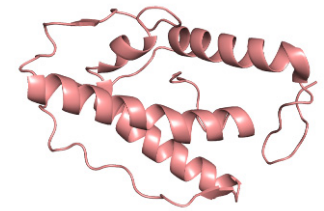
SCOP



all- β class



4-helical cytokines



T4 endonuclease V



Globin-like

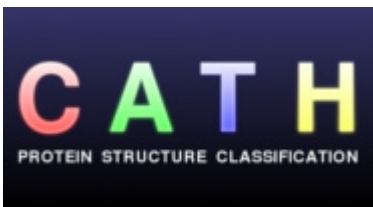
all- α class,
3 different folds



TIM-barrel fold
 α/β class



Profilin-like fold
 $\alpha+\beta$ class



CATH

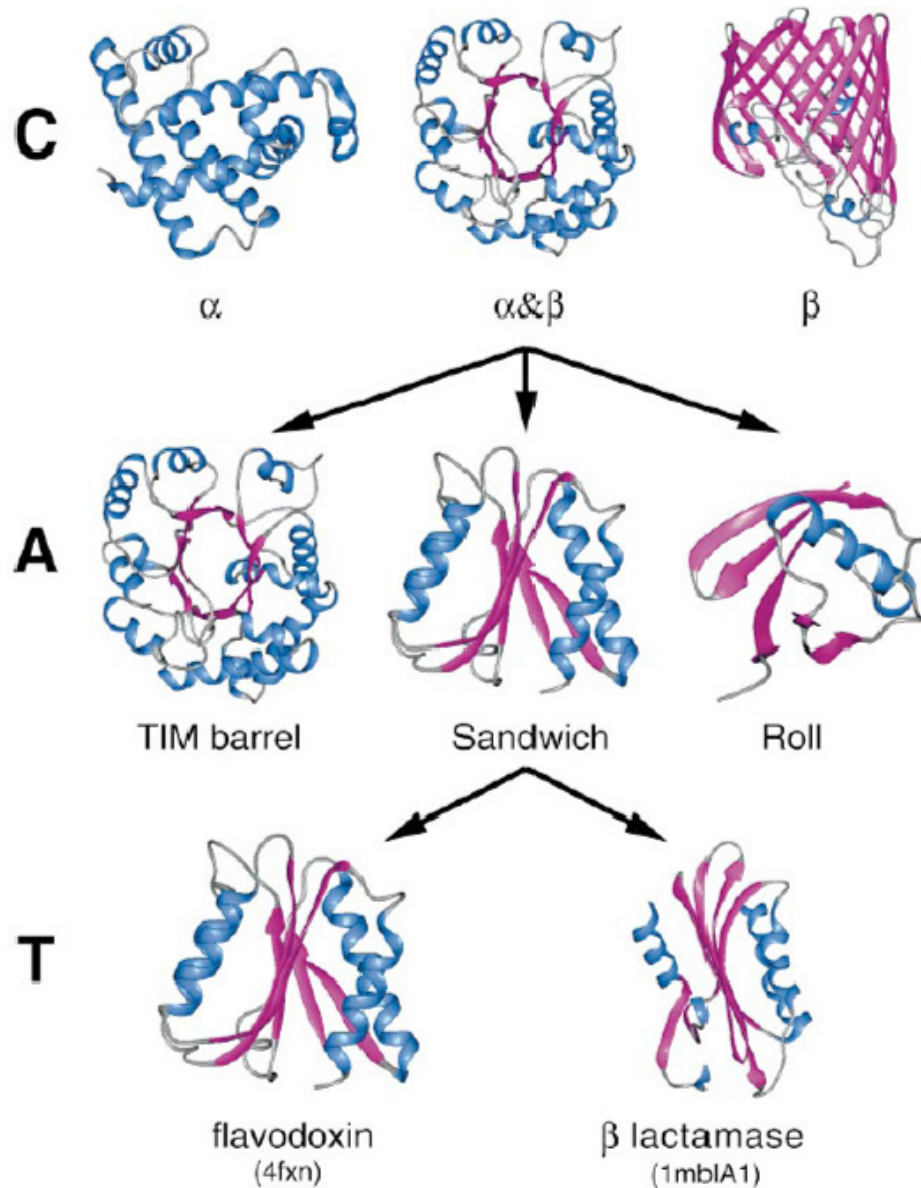
Jon K. Lærdahl,
Structural Bioinformatics

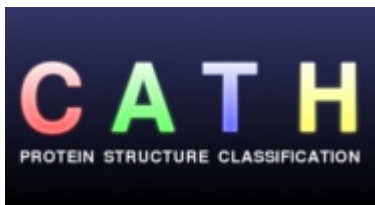
Class, Architecture, Topology
and Homologous

Both manual structural
alignment and automatic
alignment with SSAP

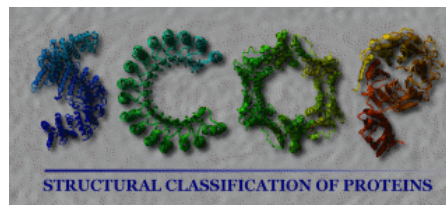
5 levels in hierarchy

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





CATH vs. SCOP

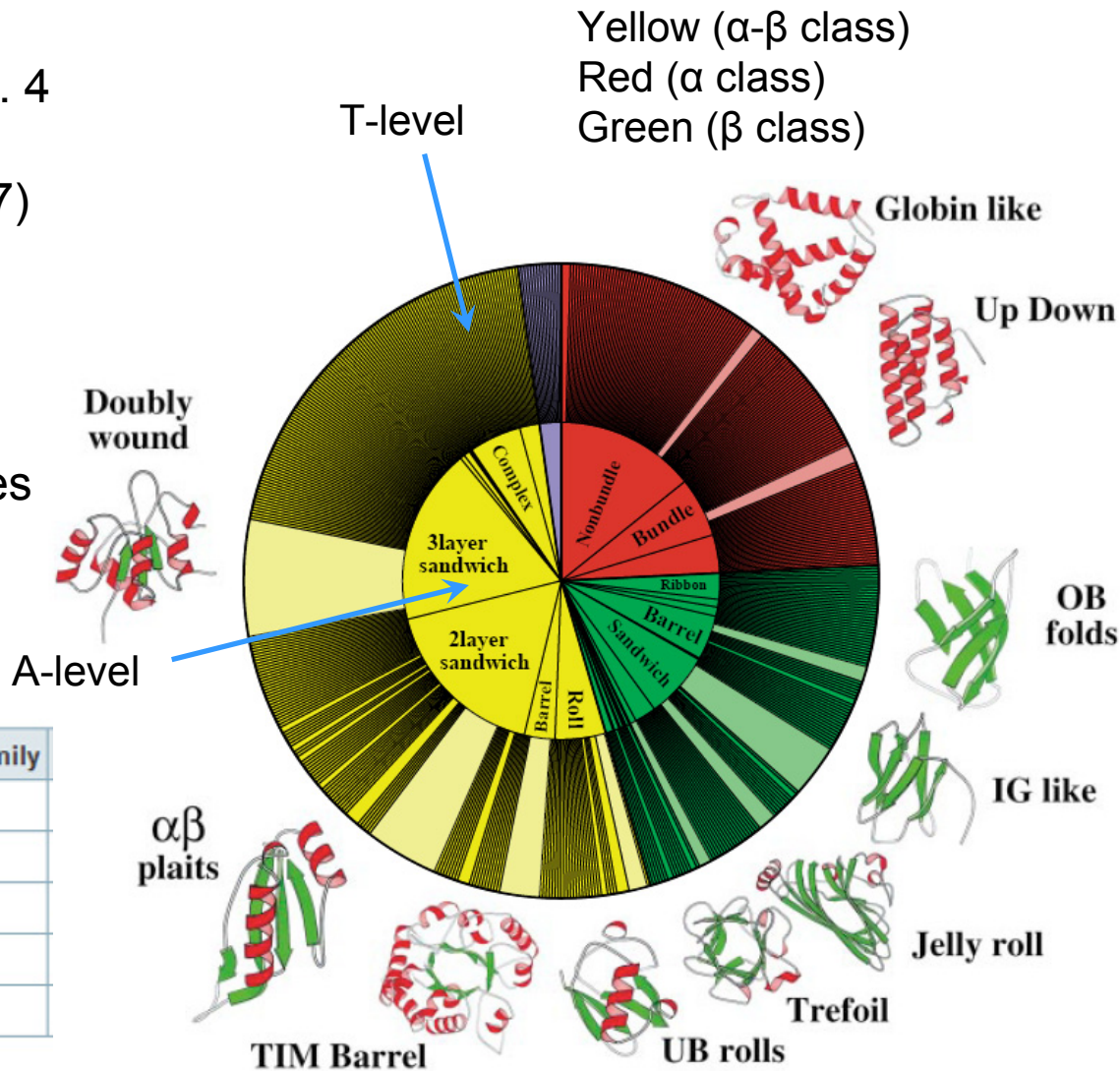


Jon K. Lærdahl,
Structural Bioinformatics

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

CATH Version 3.2

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



New topologies/folds are not found often!

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

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!

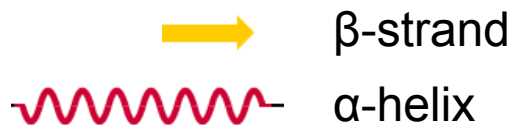
Here: *secondary structure prediction*, but similar or related methods/algorithms are used in most predictors

Secondary structure prediction

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

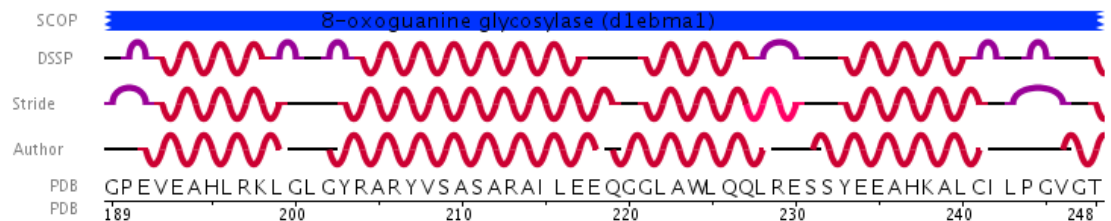
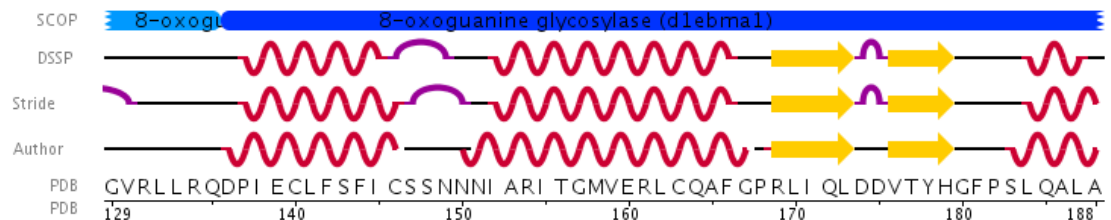
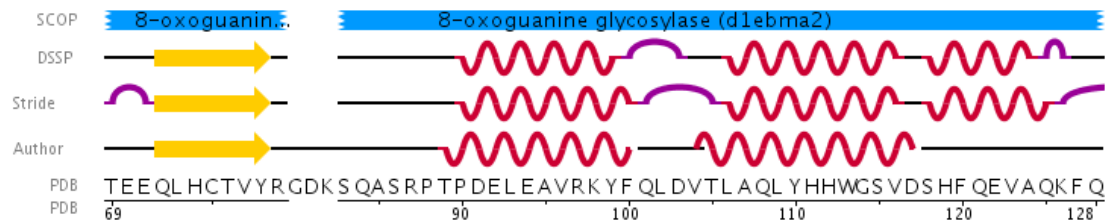
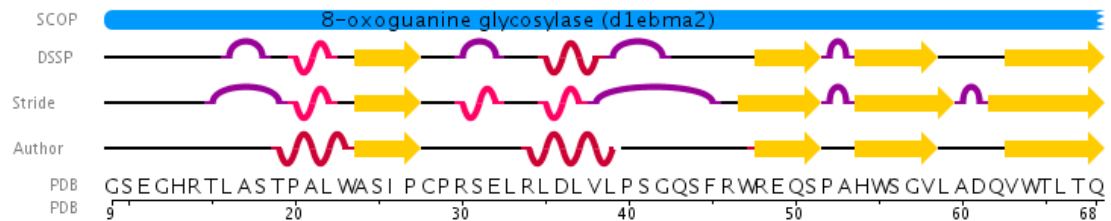
- Secondary structure may be put in manually by the authors behind a PDB-file
- Algorithms based on calculated H-bonds, Ramachandran plot, etc.

- DSSP
- STRIDE
- DEFINE



Everything else loop/coil

1EBM



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

The screenshot shows a web browser window titled "Department of Computer Science - Computational Biology Group: Prof - Windows Internet Explorer". The address bar contains "http://www.aber.ac.uk/~phiwww/prof/". The search bar contains "secondary structure prediction PROF". The browser's toolbar shows various icons and a search button. The main content area displays the Aberystwyth University logo and the text "Aberystwyth University Computational Biology Group. Department of Computer Science, Aberystwyth SY23 3DB, Wales, UK." Below this is the title "PROF - Secondary Structure Prediction System". The form includes a section for "Submit a single amino acid sequence for secondary structure prediction:" with a text input field for an email address, a dropdown menu for "Select your desired output format:" set to "CASP", and a large text area for the "Please enter your sequence in FASTA format" instruction. A "Submit Query" button is at the bottom of the form. The browser's status bar at the bottom shows "Done" and "Internet".

Department of Computer Science - Computational Biology Group: Prof - Windows Internet Explorer

http://www.aber.ac.uk/~phiwww/prof/ secondary structure prediction PROF

Suggested Sites BioInfo Biology Journals Other Answers.com Bioinfo Links cbo-all Adm FUGE bioinf

Department of Computer Science - Computational Biol...

PRIFYSGOL ABERYSTWYTH UNIVERSITY

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

PROF - Secondary Structure Prediction System

Submit a single amino acid sequence for secondary structure prediction:

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

Select your desired output format:
CASP

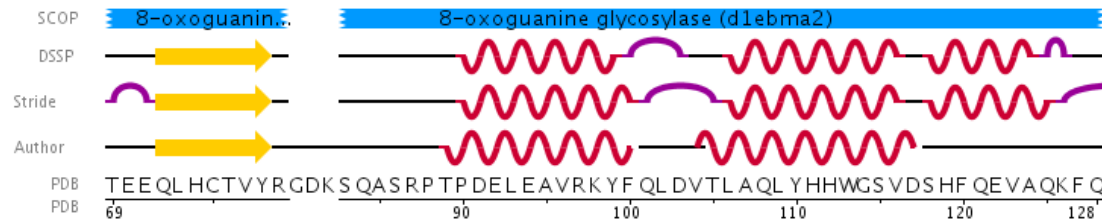
Please enter your sequence in FASTA format (first line starting with > and the title reference, followed by multiple lines of single letter amino acid sequence (NO ALIGNMENTS OR DNA PLEASE!!!)):

Submit Query

Done Internet 100%

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

TEEQLHCTVYRGDKSQASR**PTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHFQEVAQKFQ**

PROF Prediction

CEEEEEEEEECCCCCCCCCHHHHHHHHHHHHCCCCCHHHHHHHCCCCCHHHHHHHHHHHHCC

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

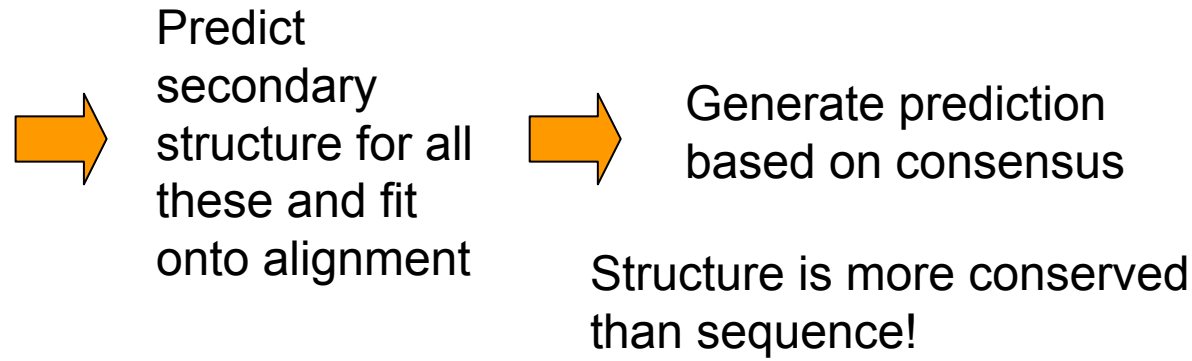
Secondary structure prediction

- 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

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

	10	20
NP_833004/1-235	GNRKDN	AFSESK
1706_Bc/1-256	GNRKDN	AFSESK
ZP_00740414/1-111	GNRKDN	AFSESK
ZP_00235456/1-229	GNRKDNE	FSESK
YP_052634/1-229	GNRKDN	FSESK
ZP_00393536/1-235	GNRKDN	FSESK
YP_037360/1-251	GNRKDN	FSESK
NP_979598/1-235	GNRKDN	FSESK
YP_084575/1-235	GNRKDN	FSESK
YP_092361/1-235	GNRKDN	FSESK
NP_712948/1-229	SILPNDG	IDSKE
YP_001221/1-235	SILPNDG	IDSKE
ZP_00533308/1-229	TKLHPFR	LNTKL
ZP_00240774/1-227	-A	IKNK
NP_832674/1-227	-A	IKNK
NP_979281/1-227	-A	IKNK
YP_037007/1-227	-A	IKNK
ZP_00393174/1-227	-A	IKNK

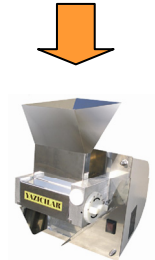


Sequences & known secondary structures from PDB



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:

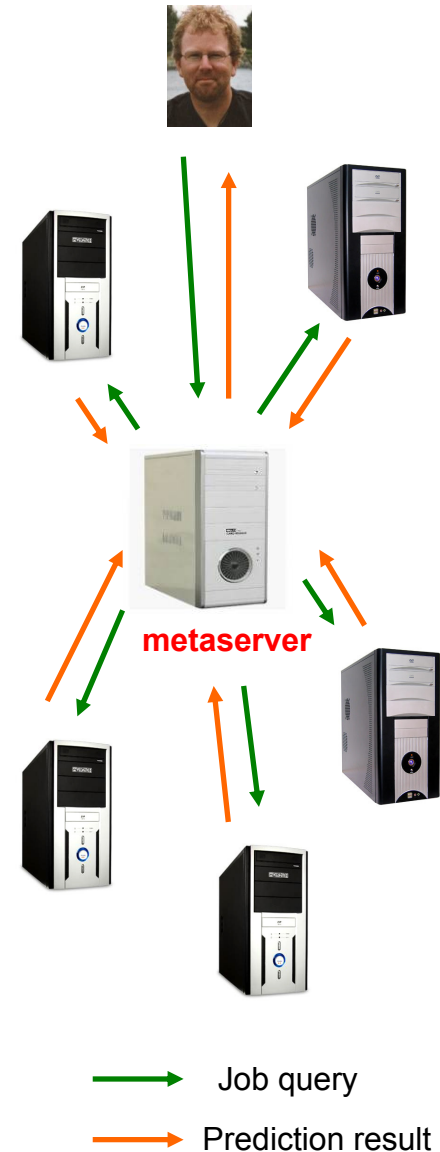
- **PHD** – BLASTP to find homologs, MSA of homologs, neural networks used for prediction, web server
- **PSIPRED** – PSI-BLAST for homologs, MSA generated, neural network prediction, filtering, web server
- **PROF** – PSI-BLAST, MSA, neural network

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

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

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

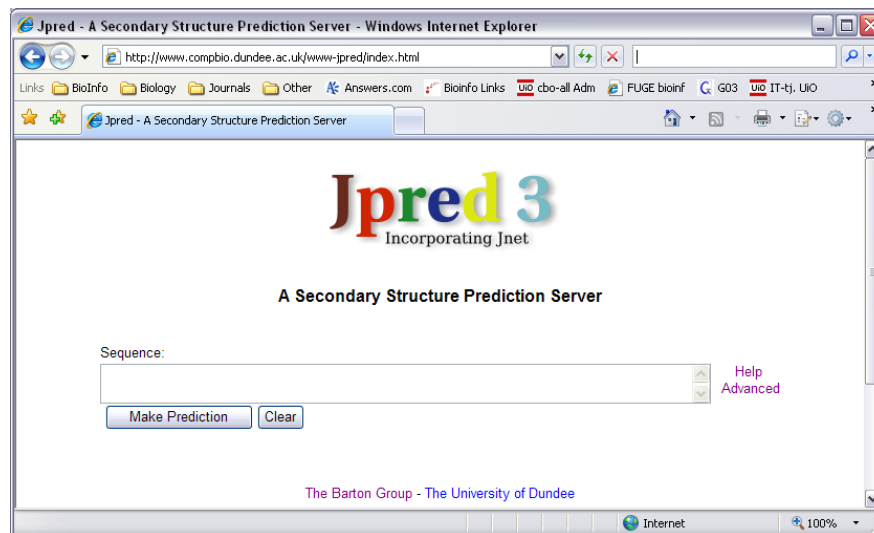
jpred        : -----HHHHH-----EEEE-----EE-HHHHHHHHHHH-----HHHHHHHHHHHH-----HHHHHHHHHHHHHH-----EEE : jpred

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-----B-----B-----B----- : Jnet_0
Jnet Rel    : 68888774110389831202254570799558841644325999998826841489999999997587998187899999998860525874 : Jnet Rel
```

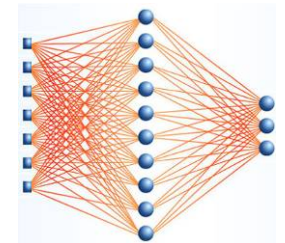
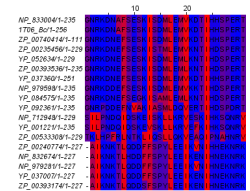
Puehringer *et al. BMC Biochemistry* 9:8 (2008)

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



Predictors – *common features*

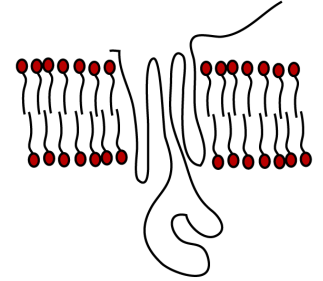
- Use propensities/intrinsic tendencies of single residues or short sequence segments to be in a certain state (e.g. secondary structure state, order/disorder state, signal sequence)
- Include local interactions, *i.e.* take into account states in up- and downstream sequence
- Use homologous sequences to get predictions from many sequences with same structure/function
- Use neural networks or similar methods in predictions
- Consensus from many tools is better than just a single result (e.g. metaservers)



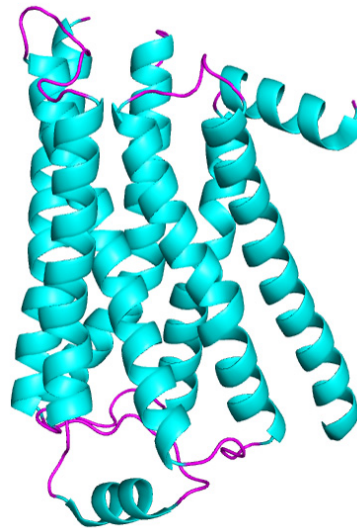
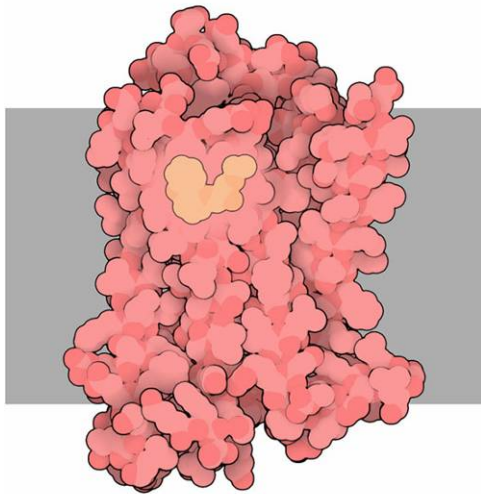
Transmembrane (TM) proteins

- ~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

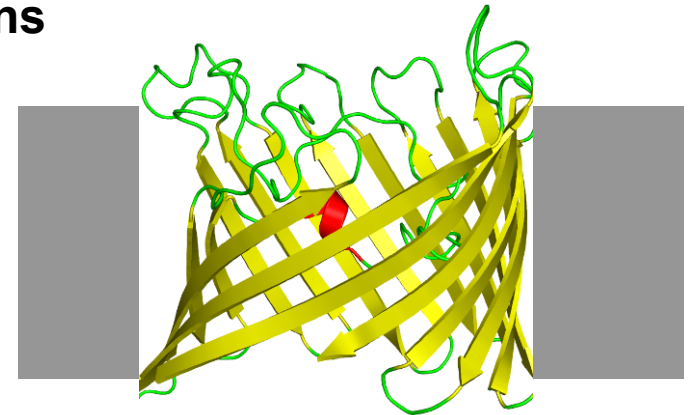


Can usually NOT use the same predictors for secondary structure and other properties as for globular proteins



PDB Apr. 08 "Molecule of the Month"

2RH1, Human adrenergic receptor



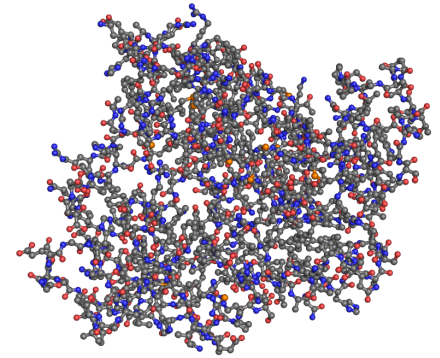
Porin

3D Structure Modeling

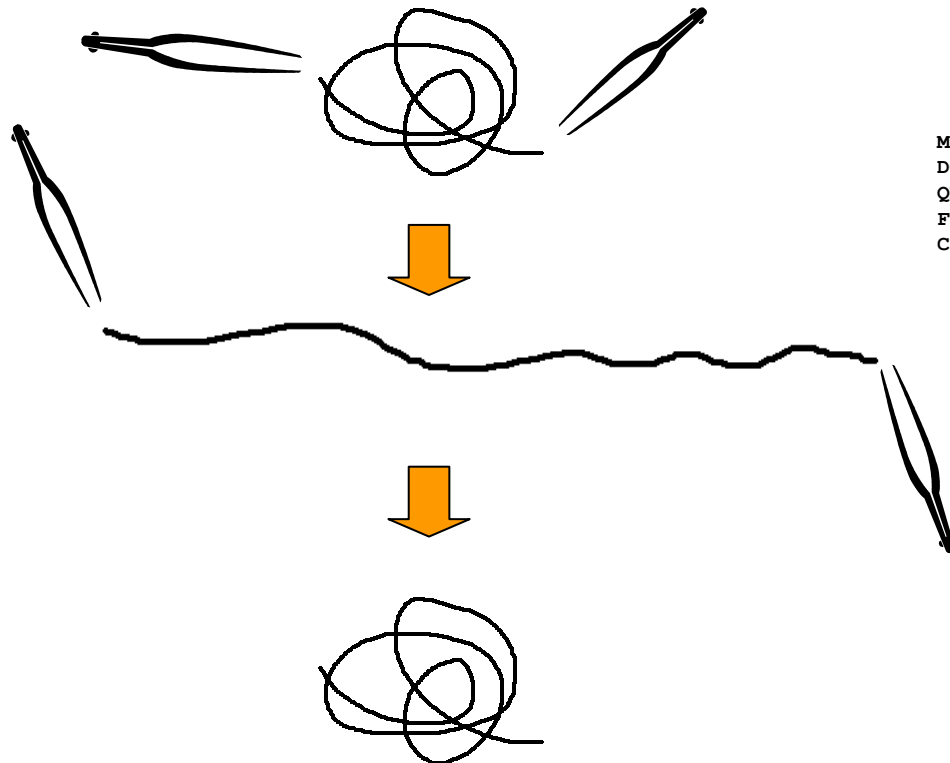
Modeling of 3D structure

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)



```
MPARALLPRRMGHRRTLASTPALWASIPCPRSELRLDLVLPQGQSFWRQSPAHWSGVLA  
DQVWTLTQTEEQQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF  
QEVAQKFGVRLLRQDPICLFSFICSSNNIARI TGMVERLCQAFGPRLIQLDDVTYHG  
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAI LEEQGGLAWLQQLRESSYEEAHKAL  
CILPGVGTKVADCI CLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPPQTNKELG
```

The sequence
determines the 3D
structure!

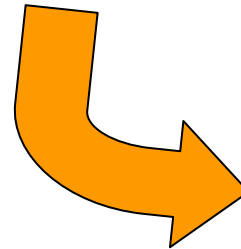
Nobel Prize in chemistry
1972 to Christian B.
Anfinsen

Protein folding

```
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVLPSPGQSFWRREQSPAHWSGVLA  
DQVWTLTQTTEEQLHCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF  
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG  
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGLAWLQQLRESSYEEAHKAL  
CILPGVGTKVADCICLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPPQTNKELG  
NFFRSLWGPYAGWAQATPPSYRCCSVPTCANPAMLRSHQQAERVPKGRKARWGTLDKEI
```

**The sequence determines
the 3D structure!**

In the computer

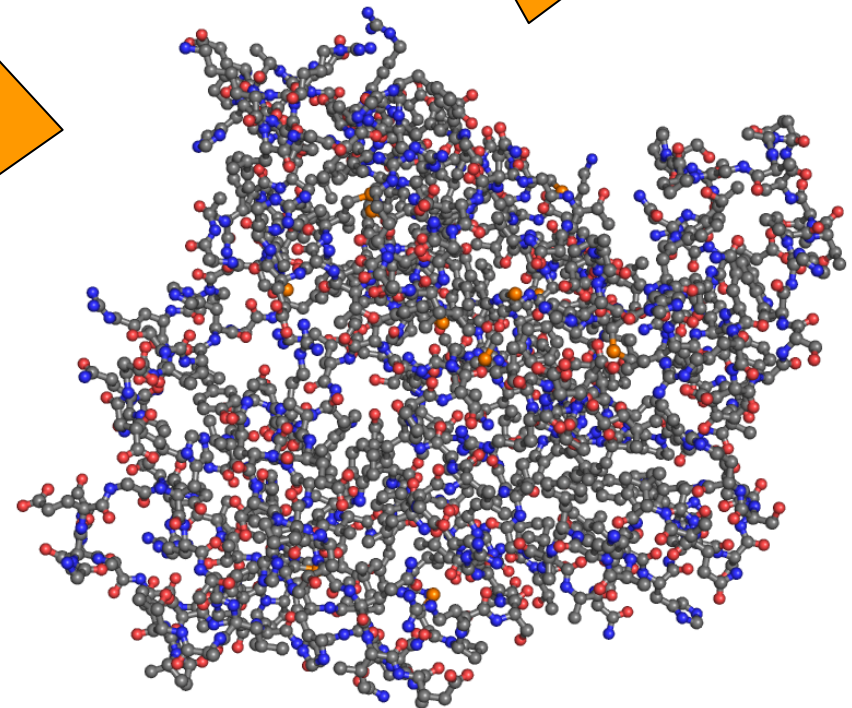
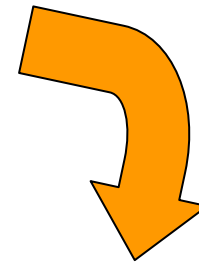


***Ab initio/de novo* structure prediction**

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

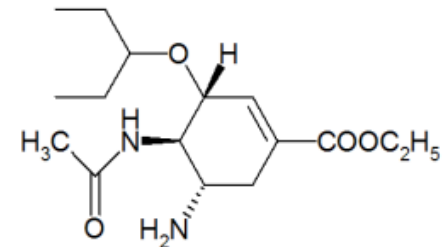
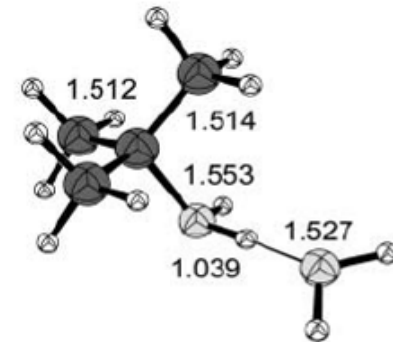
Folding is
spontaneous in the
cell

In the cell



Ab initio structural prediction

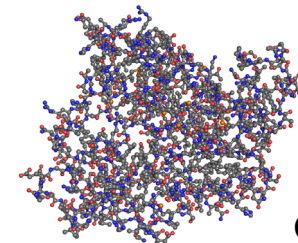
- Determine the tertiary structure for a protein based on amino acid sequence and chemical and physical laws only
- Does *not* use prior knowledge of structure from the PDB
- *Ab initio* quantum chemistry is pure “*ab initio*”
 - Based on solving the Schrödinger equation
 - Is routinely used for chemical 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

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

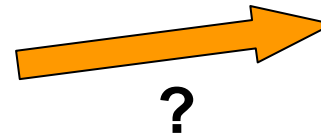
- All problems with *ab initio* modeling will never be completely solved?
- They have certainly not been solved yet!



or



```
MPARALLPRMGHRTLASTPALWASIPCPRSELRLDLVLPSSQSFWRREQSPAHWSGVLA  
DQVWTLTQTEEQHLCTVYRGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVDSHF  
QEVAQKFQGVRLLRQDPIECLFSFICSSNNNIARITGMVERLCQAFGPRLIQLDDVTYHG  
FPSLQALAGPEVEAHLRKLGLGYRARYVSASARAILEEQGLAWLQQLRESSYEEAHKAL  
CILPGVGTKVADCI CLMALDKPQAVPVDVHMWHIAQRDYSWHPTTSQAKGPSQTNKELG
```



Structural bioinformatics

Jon K. Lærdahl,
Structural Bioinformatics

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



[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/INNS/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/INNS/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

Charts

Top500 List:

06/2008

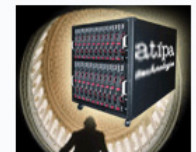
Chart Type:

Vendors

Development



Innovation that
matters.



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 structure¹ — 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-

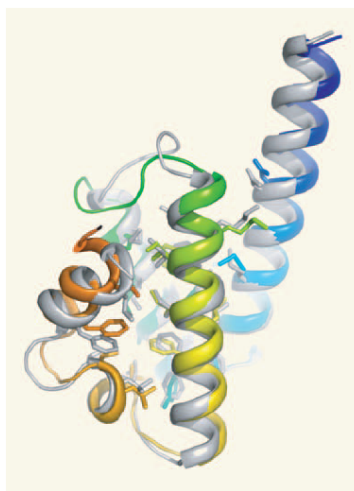


Figure 1 | Model test. Qian *et al.*² have developed a computational method for predicting the three-dimensional structure of a protein from its amino-acid sequence. Here, their predicted structure (grey) of a protein is overlaid with the experimentally determined crystal structure (shown in colour) 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’ searches⁵ 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' models² 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 structure¹ — 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-

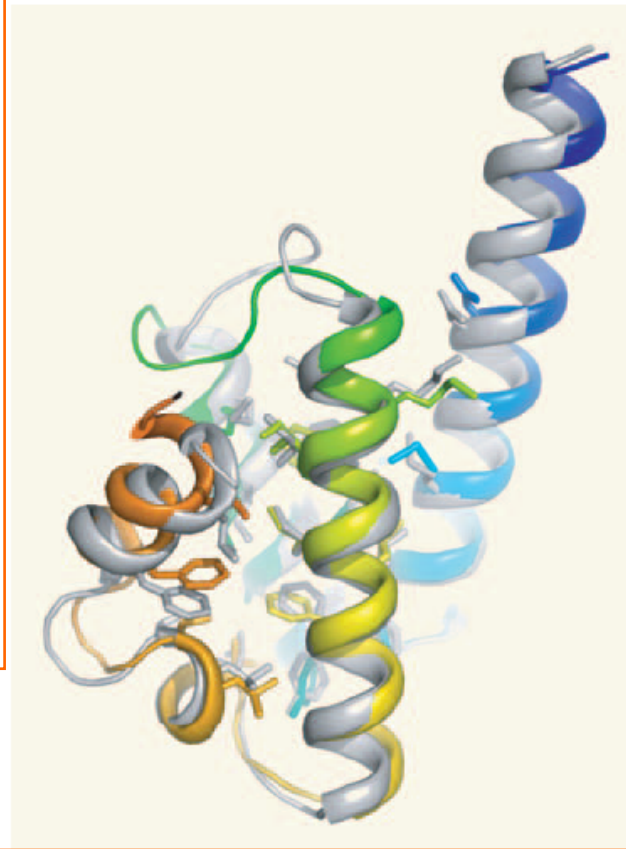


Figure 1 | Model test. Qian *et al.*² have developed a computational method for predicting the three-dimensional structure of a protein from its amino-acid sequence. Here, their predicted structure (grey) of a protein is overlaid with the experimentally determined crystal structure (shown in colour) 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’ searches⁵ 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' models² 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-

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@home⁶, 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.

The screenshot shows the Rosetta@home website in a Windows Internet Explorer browser window. The address bar shows the URL <http://boinc.bakerlab.org/rosetta/>. The page features the Rosetta@home logo and navigation links for Protein Folding, Design, and Docking. A search bar is present. The main content area includes a "What is Rosetta@home?" section, a "Join Rosetta@home" section with a list of links, an "About" section with a list of links, and a "Returning participants" section. A "User of the day" section features a profile for dori_doreau. A "Server Status" section displays real-time statistics as of 30 Oct 2007 10:57:39 UTC, including scheduler running status, queued tasks, in-progress tasks, successes, users, hosts, credits, and TeraFLOPS estimate. A "News" section highlights a "Predictor of the day" for Wade Cazabat and an article in Nature.

Rosetta@home Protein Folding, Design, and Docking

What is Rosetta@home?

Rosetta@home needs your help to determine the 3-dimensional shapes of proteins in research that may ultimately lead to finding cures for some major human diseases. By running the Rosetta program on your computer while you don't need it you will help us speed up and extend our research in ways we couldn't possibly attempt without your help. You will also be helping our efforts at designing new proteins to fight diseases such as HIV, Malaria, Cancer, and Alzheimer's (See our [Disease Related Research](#) for more information). Please [join us](#) in our efforts! **Rosetta@home is not for profit.** [login/out]

Join Rosetta@home

1. [Rules and policies](#)
2. [System requirements](#)
3. [Download, install, and run BOINC](#)
(enter the project URL: <http://boinc.bakerlab.org/rosetta/>)
4. [A welcome from David Baker](#)

About

- [10 reasons why users crunch Rosetta@home](#)
- [Quick Guide to Rosetta@home and its Graphics](#)
- [Rosetta@home FAQ](#)
- [Rosetta@home Science FAQ](#)
- [Disease Related Research](#)
- [Research Overview](#)
- [News & Articles about Rosetta](#)
- [David Baker's Rosetta@home Journal](#)
- [Rosetta@home promo video](#)
- [Technical news](#)

Returning participants

- [Your account - view stats, modify preferences](#)

User of the day

 [dori_doreau](#)

Server Status as of 30 Oct 2007 10:57:39 UTC

[Scheduler running] Queued: 19,018
In progress: 324,535
Successes last 24h: 226,915
Users [↓](#) (last day [↓](#)) : 166,545 (+205)
Hosts [↓](#) (last day [↓](#)) : 435,054 (+504)
Credits last 24h [↓](#) : 6,255,810
Total credits [↓](#) : 2,670,214,176
TeraFLOPS estimate: 62.538

Oct 29, 2007
Predictor of the day: Congratulations to [Wade Cazabat](#) (Team [Invaders](#)) for predicting the lowest energy structure for workunit [hpr_200_BOINC_MFR_ABRELAX_PICKED_1972_0](#)!
[...more](#) [XML](#) Available as an [RSS feed](#)..

News

Oct 17, 2007
[An article about Rosetta@home](#) is in Nature. Congrats and thank you to all the volunteers that made this possible!

Rosetta@home

Rosetta@home - Windows Internet Explorer

http://boinc.bakerlab.org/rosetta/

Links BioInfo Biology Journals Other Answers.com cbo-all Adm FUGE bioinf G03 Google UIO IT-tj. UIO PubMed SGP ABC Startsiden Wikipedia

Rosetta@home

Rosetta@home

Protein Folding, Design, and Docking

[What is Rosetta@home?](#)



Rosetta@home needs your help to determine the 3-dimensional shapes of proteins in research that may ultimately lead to finding cures for some major human diseases. By running the Rosetta program on your computer while you don't need it you will help us speed up and extend our research in ways we couldn't possibly attempt without your help. You will also be helping our efforts at designing new proteins to fight diseases such as HIV, Malaria, Cancer, and Alzheimer's (See our [Disease Related Research](#) for more information). Please [join us](#) in our efforts! **Rosetta@home is not for profit.**

[login/out]

Site search

Join Rosetta@home

- [Rules and policies](#)
- [System requirements](#)
- [Download, install, and run BOINC](#)
(enter the project URL: <http://boinc.bakerlab.org/rosetta>)
- [A welcome from David Baker](#)

About

- [10 reasons why users crunch Rosetta@home](#)
- [Quick Guide to Rosetta@home and Its Graphics](#)
- [Rosetta@home FAQ](#)
- [Rosetta@home Science FAQ](#)
- [Disease Related Research](#)
- [Research Overview](#)
- [News & Articles about Rosetta](#)
- [David Baker's Rosetta@home Journal](#)
- [Rosetta@home promo video](#)
- [Technical news](#)

Returning participants

- [Your account - view stats, modify preferences](#)

User of the day



[dori_doreau](#)

Server Status as of 30 Oct 2007 10:57:39 UTC

[Scheduler running] Queued: 19,018
In progress: 324,535
Successes last 24h: 226,915
Users [↓](#) (last day [↓](#)) : 166,545 (+205)
Hosts [↓](#) (last day [↓](#)) : 435,054 (+504)
Credits last 24h [↓](#) : 6,255,810
Total credits [↓](#) : 2,670,214,176
TeraFLOPS estimate: 62.558

Oct 29, 2007
Predictor of the day: Congratulations to [Wade Cazabat](#) (Team [Invaders](#)) for predicting the lowest energy structure for workunit [hpr_200_BOINC_MFR_ABRELAX_PICKED_1972_0](#) !

[...more](#)

XML Available as an [RSS feed](#)..

News

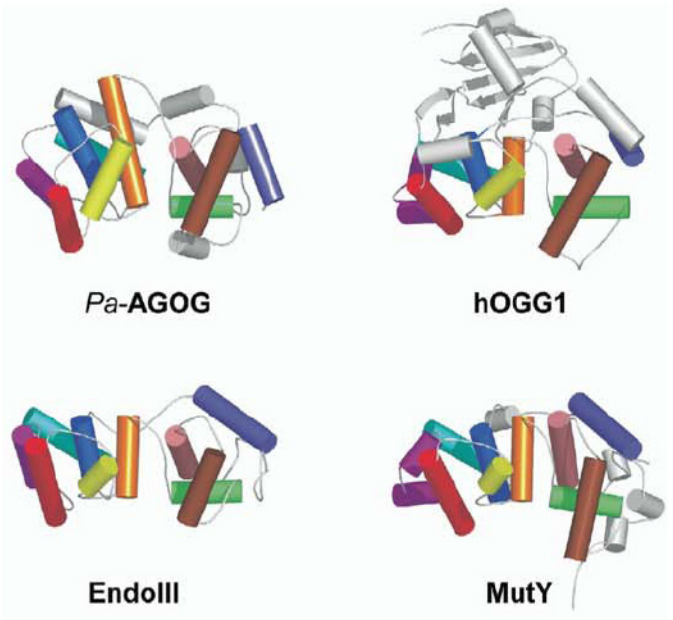
Oct 17, 2007
[An article about Rosetta@home](#) is in Nature. Congrats and thank you to all the volunteers that made this possible!

Protein structure evolution

OGG1_YEAST/1-376
OGG1_MOUSE/1-345
OGG1_RAT/1-345
OGG1_HUMAN/1-345
OGG1_FLY/1-343

```

174 SRATTEAKLRELGFGYRAKYIIETARKLVNDKAEANITSDTTYLQSICCKDAQYEDVREHLMSYNGVGPKVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISAN 276
189 GPEAETHLRKLGLGYRARYRASAKAILEEOGGP-----AWLQQLRV-APYEEAHKALCTLPGVGAKVADCICLMALDKPOAVPVDVHWQIAHRDYGWHPK 284
189 GPEVETHLRKLGLGYRARYVSASAKAILEEOGGP-----AWLQQLRV-ASYEEAHKALCTLPGVGTKVADCICLMALDKPOAVPVDIHWQIAHRDYGWHPK 284
189 GPEVETAHLRKLGLGYRARYVSASARAILEEOGGL-----AWLQQLRE-SSYEEAHKALCILPGVGTKVADCICLMALDKPOAVPVDVHMHIAQRDYSWHPPT 284
191 CEDLNAQLRAAKFGYRAKFIAQTLQEIQKKGGQ-----NWFISLKS-MPFEKAREELTLLPGIGYKVADCICLMSMGHLESVPVDIHIYRIAQNYLPHLT 285
    
```



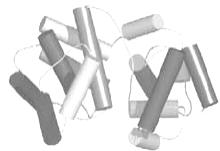
- Reason for similarities in sequence/structure is common ancestry, the sequences/structures are homologs
- Structures evolves slowly
- Sequence evolves faster
 - Many mutations does not change the structure
- Only some few 1000 superfamilies in the PDB
- Only a factor 2-10(???) as many superfamilies/folds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

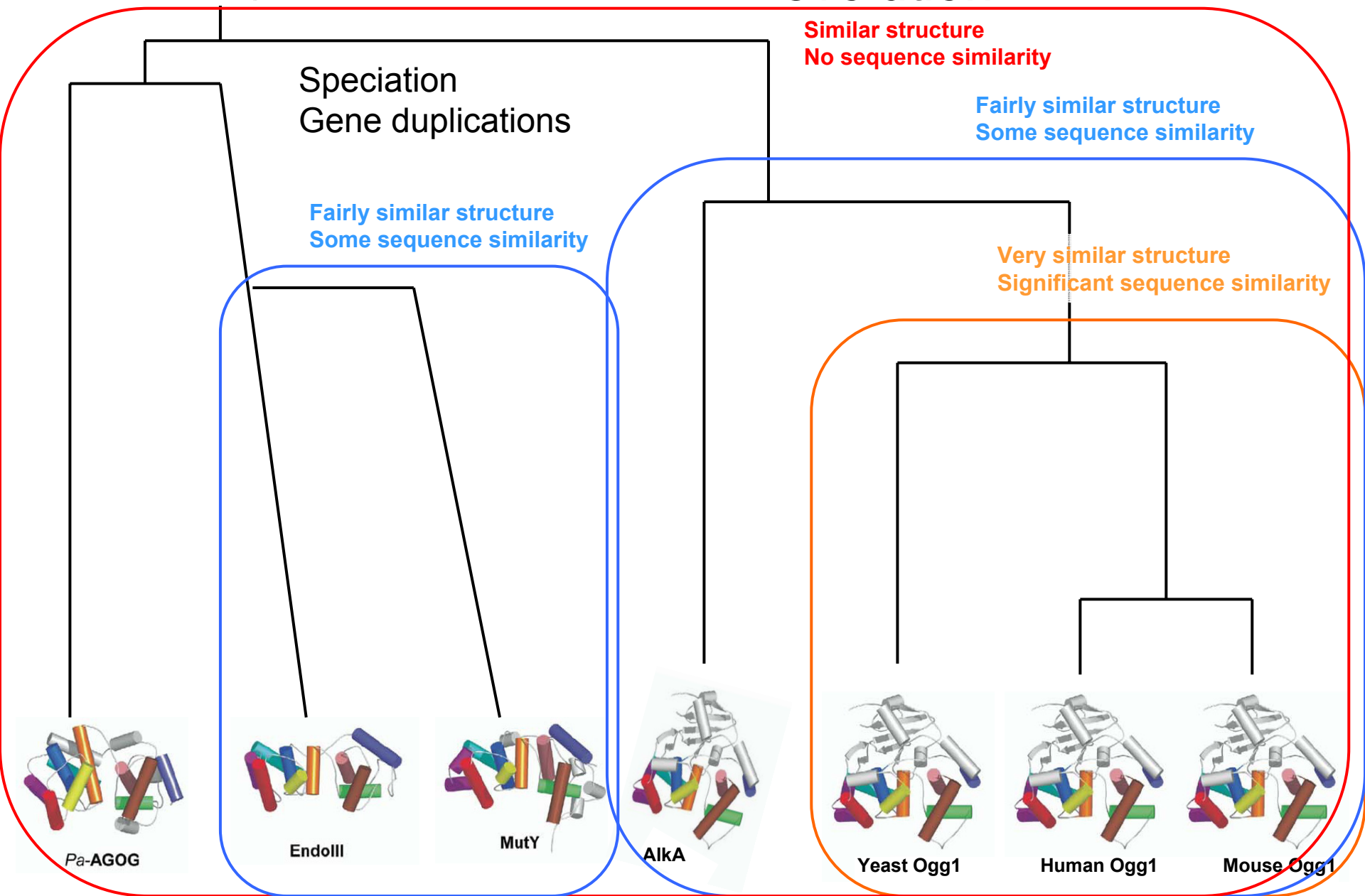
Class	Architecture	Topology	Homologous Superfamily
1	5	310	682
2	20	196	438
3	14	512	956
4	1	92	102
Total	40	1110	2178



Last common ancestor
(Long time ago...)

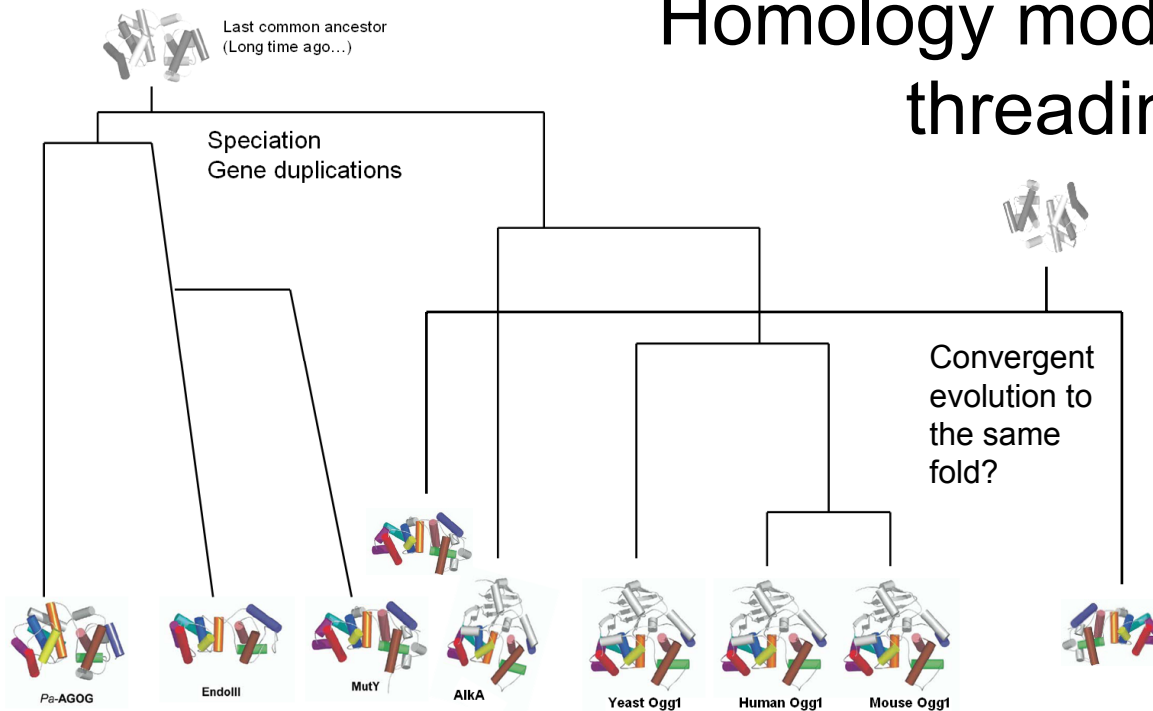
Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics



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

The screenshot shows the PSI | nature Structural Genomics Knowledgebase website. The header features the PSI | nature logo and the text 'Structural Genomics Knowledgebase'. Navigation icons include a molecular structure, a DNA helix, and a human figure. The main content area is divided into several sections: 'structural genomics update' (May 2009) with a monthly update description; 'research advances' featuring articles like 'Controlling p53' and 'Mitotic checkpoint control', and technical highlights like 'Improve your cloning efficiency' and 'Faster solid-state NMR'; 'featured molecule' highlighting the 'Bacterial leucine transporter, LeuT'; and 'news' with 'PSI in the Spotlight' and 'The beneficial side of prions'. A sidebar on the left contains navigation links like 'home', 'structural genomics update', 'research advances', etc. A right sidebar offers 'e-alerts', 'RSS (monthly updates)', and 'RSS (new molecules)'. A search section at the bottom right allows searching by text, sequence, or structure.

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 to have a new fold

Structural genomics/The Protein Structure Initiative (PSI)

Jon K. Lærdahl,
Structural Bioinformatics

Security /
Privacy Notice



Midwest Center
for
Structural Genomics



Structure Gallery

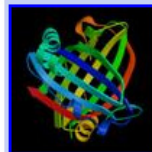
• XML Files • Target List • Progress • Statistics • Log in • Site Search:

Go

Consortium
Project
Investigators
Targets
3-D Structures
Related Publications
SG Sites
SG Progress
NIH
MCSG Resources
Job opportunities
Collaborators
Internals
Technologies

GALLERY OF MCSG STRUCTURES IN PDB

959 targets in PDB (28 new folds)



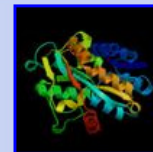
APC006 [ref]
[1SQE](#) ident: 23.9%
[annotation](#)



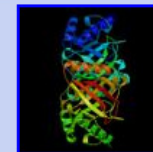
APC007
[1XBW](#) ident: 64.5%
[annotation](#)



APC008
[2AP3](#) ident: <20%
[annotation](#)



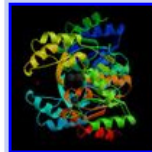
APC009 [ref]
[1P99](#) ident: <20%
[annotation](#)



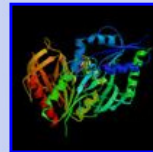
APC010 [ref]
[1NG5](#) New Fold
[annotation](#)



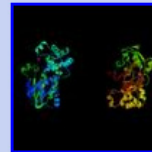
APC012 [ref]
[1KR4](#) ident: <20%
[annotation](#)



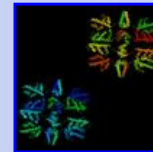
APC014 [ref]
[1KYT](#) ident: <20%
[annotation](#)



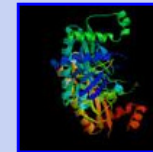
APC037 [ref]
[1KXJ](#) ident: 100%
[annotation](#)



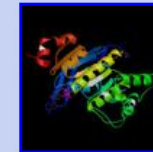
APC038 [ref]
[1M6Y](#) ident: <20%
[annotation](#)



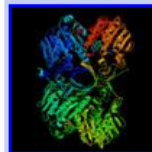
APC042
[1WPB](#) ident: <20%
[annotation](#)



APC043 [ref]
[1KUT](#) ident: <20%
[annotation](#)



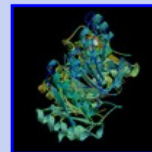
APC046
[1J10](#) ident: 33.5%
[annotation](#)



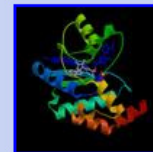
APC047 [ref]
[1JQ3](#) New Fold
[annotation](#)



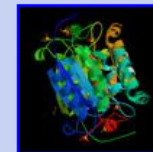
APC048 [ref]
[1MKM](#) ident: <20%
[annotation](#)



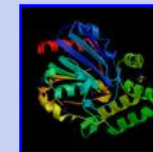
APC049
[1TS7](#) ident: <20%
[annotation](#)



APC050 [ref]
[1EJ2](#) ident: <20%
[annotation](#)



APC063 [ref]
[1MKZ](#) ident: 30%
[annotation](#)



APC064 [ref]
[1M33](#) ident: 26.2%
[annotation](#)

Structural genomics/The Protein Structure Initiative (PSI)

Jon K. Lærdahl,
Structural Bioinformatics

Security / Privacy Notice

MCSG

Midwest Center for Structural Genomics

PSI

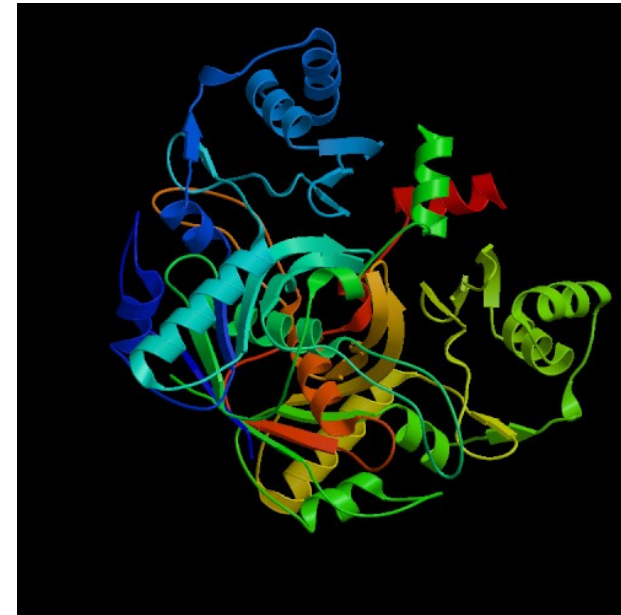
Structure Gallery

XML Files • Target List • Progress • Statistics • Log in • Site Search: Go

GALLERY OF MCSG STRUCTURES IN PDB

959 targets in PDB (28 new folds)

 APC006 [ref] LSQE ident: 23.9% annotation	 APC007 LXRY ident: 64.5% annotation	 APC008 2AP3 ident: <20% annotation	 APC009 [ref] IP99 ident: <20% annotation	 APC010 [ref] LNG5 New Fold annotation	 APC012 [ref] IKR4 ident: <20% annotation
 APC014 [ref] LKYI ident: <20% annotation	 APC037 [ref] LKNJ ident: 100% annotation	 APC038 [ref] LM6Y ident: <20% annotation	 APC042 LWPE ident: <20% annotation	 APC043 [ref] LKUT ident: <20% annotation	 APC046 LJI0 ident: 33.5% annotation
 APC047 [ref] LJQ3 New Fold annotation	 APC048 [ref] LMKM ident: <20% annotation	 APC049 LTS7 ident: <20% annotation	 APC050 [ref] LEJ2 ident: <20% annotation	 APC063 [ref] LMKZ ident: 30% annotation	 APC064 [ref] LM33 ident: 26.2% annotation



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!

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

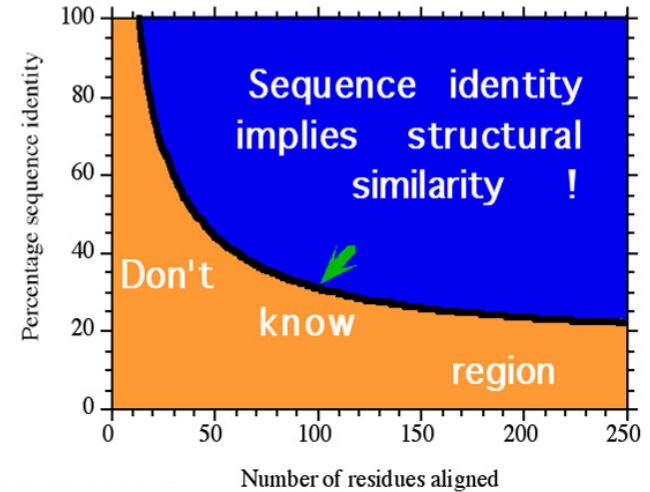
Known (experimental) structure of protein 1 (*template*)

&
Sequence alignment with protein 2 (*target*)



Model of protein 2

Evolution is the history!



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

- 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?

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!

Homology modeling

Start with a protein sequence (target)

1. Template selection:

- Find template sequences

```
I want to model this!
```

```
>gi|84618885|emb|CAJ31885.1| methylpurine-DNA  
glycosylase [Bacillus cereus]
```

2. Correct alignment

- Use the best alignment
- Correct placement of insertions and deletions

```
MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQT  
PERRQLLKDVIIQIHTLPDPKDFRIIVRELWDLPEREFQA  
AALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVP  
TFLGNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLK  
YKQKMDEELLFWVIGQLHSSKEFFIQKAIGWVLREYAKT  
KPDVVWEYVQNNELAPLSRREAIAIKHIKENYGINNEKIGE  
TLS
```

3. Backbone modeling

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!

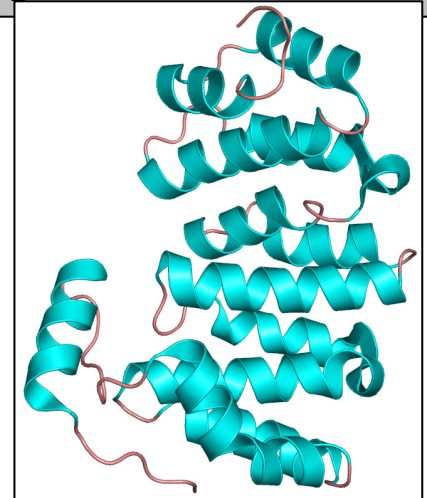
Homology modeling

Start with a protein sequence (target)

1. Template selection:
 - Find template in PDB and compare sequences
2. Correct alignments
 - Use the best MSA program
 - 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 sequence search in all "PDB sequences"

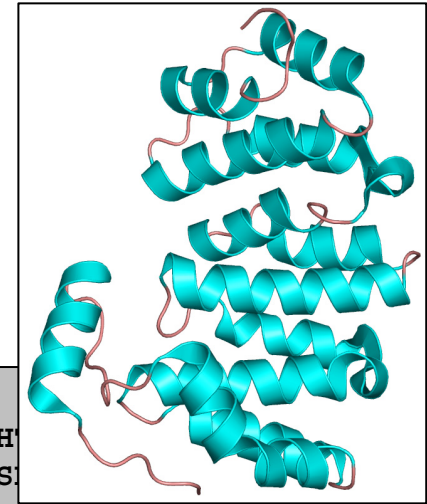
- Useful templates have 30% or higher sequence identity to target (but sometimes even lower)
- Several templates?
 - Resolution?
 - Highest sequence identity?
 - Cofactors?
 - Use the structure that best fits your task



Homology modeling

Start with a protein sequence (target)

1. Template selection:
 - Find template in PDB and align sequences
2. Correct alignments



Sequence alignment

```

Bc_AlkD  MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQTPERRQLLKDVIQIH
EF3068  -----MDTLQFQKNPETAAKMSAYMKHQFVFAGIPAPERQALSKQLLKES!
          :  :  :****.*  * :  ***::*: *  *  :***:  *  *:::  :
Bc_AlkD  FRIIVRELWDLPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL 120
EF3068  LCQEIEAYYQKTEREYQYVAIDLALQNVQRFSLEEYVAFKAYVPQKAWWDSVDAWRKFFG 122
          :  :.  ::  .***:*  .**:*  :  :  :.  .:  ::  :  *::***:**:  *
Bc_AlkD  GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQMDEELLFWVIGQLHSSKEFF 180
EF3068  SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIYDRTTEEFF 171
          .  :  * :  **:.  :.  :  .::**:*  :*. * :  **  *:*  :::**  .*  ::::***
Bc_AlkD  IQKAIGWVLREYAKTKPDVWVEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068  IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----- 217
          *****  **::***:*  *  *  :::  *::***:**.  *::  :
    
```

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

5. Refine and optimize model
6. Validate and check model quality!

Homology modeling

Start with a protein sequence (target)

1. Template selection:
 - Find template in PDB and homologous sequences
2. Correct alignments
 - Use the best MSA program
 - Correct placement of insertions and deletions

Check indels!

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

3. Backbone
4. Model building
 - Rigid body
 - Loop closure
5. Refinement
6. Validation

```

Sequence alignment
Bc_Alkd  MHPFVKALQEHFIAHKNPEKAEPMARY
EF3068  -----MDTLQFQKNPETAAKMSAY
          : : :****.* *: *
Bc_Alkd  FRIIVRELWDLPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL 120
EF3068  LCQRIEAIYQKTEREFQYVAIDLALQNVQRFSLSEEVVAFKAYVPQKAWWDSVDARWRF 122
          : . :. :. :****.* .*: : : :. :. :. : * :****:*** : *
Bc_Alkd  GNIFLQHPHELISAYIPKWIASDNIWLRRAAILFQLKYKQKMDDELLFWVIGQLHSSKEFF 180
EF3068  SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEFF 171
          . : * : ** :. : : :****.* :*.* : ** *:* :**** * : : :****
Bc_Alkd  IQKAIGWVLRVAKTKPDVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068  IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQRREGSKYLAKASE----- 217
          ***** **:*:**:* : * * :. : * :****:***. * : : :
    
```

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

Homology modeling

Start with

1. Temp

– Fir

se

2. Corre

– Us

– Co

and deletions

3. Back

4. Mod

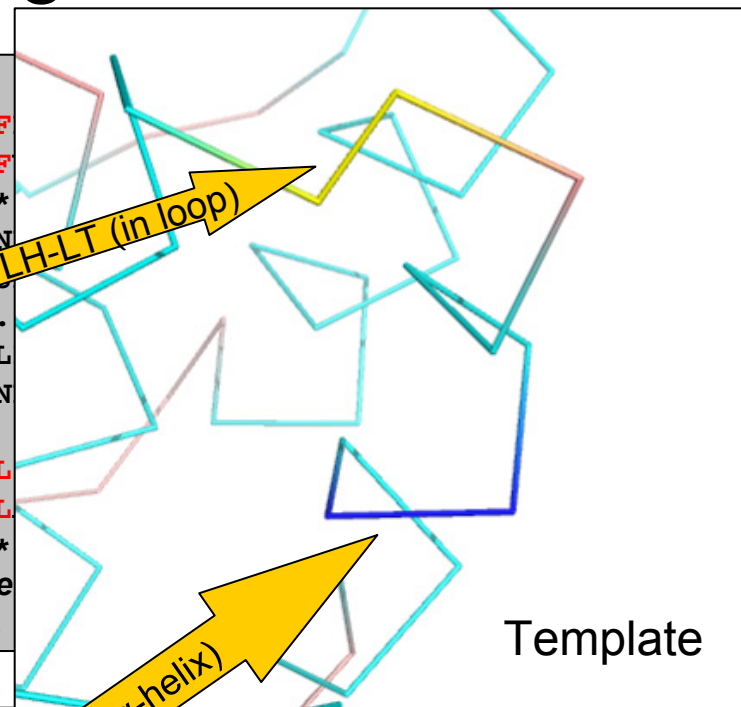
– R

– L

5. Refi

6. Valid

```
Sequence alignment
Bc_AldD MHPFVKALQEHFIAHKNPEKAEPMARYMKNHF
EF3068 -----MDTLQFQKNPETAAKMSAYMKHQF
      : : :****.* *: ***:**
Bc_AldD FRIIVRELWDLPEREFQAAALDMMQKYKYYIN
EF3068 LCQEIEAYIQKTERTYQVVAIDLALOM
      : : : : .****.* .*. : : : :
Bc_AldD GNIFLQHPHELISAYIPKWIASLNIWLQRAAIL
EF3068 SWVALH-LTELPTIFALFYGAENFWNRRVALN
      . : * : : : : : : : : : : : : : : : : : : : : : : : : : :
Bc_AldD IQKAIGWVLRVYAKTKPDVWEYVQNNELAPL
EF3068 IQKAIGWSLRQYSKTNPQWVEELMKELVLSPL
      ***** **:*:**:*: * * : : : : * : **
CORRECTED Alignment of the sequences of B. cere
hypothetical protein EF3068 (template from MCSG).
```



```
Sequence alignment
Bc_AldD MHPFVKALQEHFIAHKNPEKAEPMARYMKNHF...QTPEERRQLLKDVIIQIHTLPDPKD 60
EF3068 -----MDTLQFQKNPETAAKMSAYMK...AGIPAPERQALSQQLLKESHTWPKEK 52
      : : :****.* *: ** :***: * * : : : : : :
Bc_AldD FRIIVRELWDLPEREFQAAALDMM...KYYINETHIPFLEELIVTKSWWDTVDSIVPTFL 120
EF3068 LCQEIEAYIQKTERTYQVVAID...QNVQRFSLIEEVVAFKAYVPQKAWWDSVDAWRKFFG 122
      : : : : .****.* .*. : : : : : : : : : : : : : : : : : : : : *
Bc_AldD GNIFLQHPHELISAYIPKWIASLNIWLQRAAILFQLKYKQKMDDELLFWVIGQLHSSKEFF 180
EF3068 SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEFF 171
      . : * : ** : : : : : : : : : : : : : : : : : : : : * : : : : : : : : : :
Bc_AldD IQKAIGWVLRVYAKTKPDVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068 IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQRREGSKYLAKASE----- 217
      ***** **:*:**:*: * * : : : : * : ** : : : :
Alignment of the sequences of B. cereus AldD (target) and E. faecalis hypothetical protein
EF3068 (template from MCSG).
```

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

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!

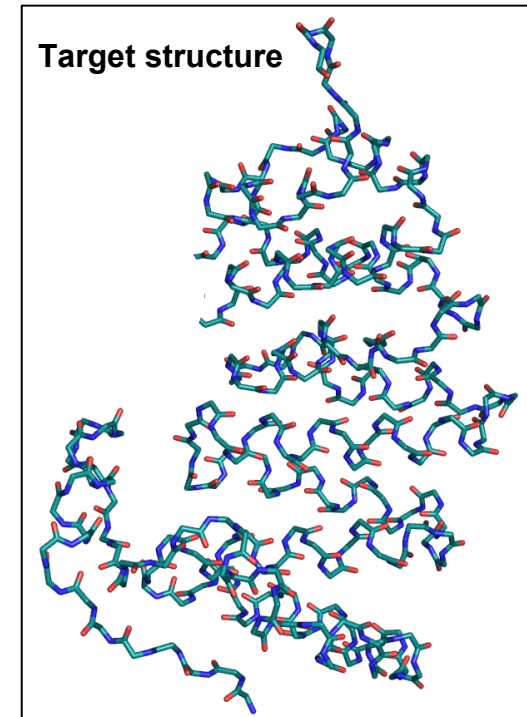
Homology modeling

Start with a protein sequence (target)

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



Homology modeling

Sta

1.

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

2.

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

3.

4.

- Use the best MSA programs
 - Correct placement of insertions and deletions
- Backbone model building
- Model loops and side-chains
- Rotamer libraries
 - Loop modeling using database or *ab initio* method

5.

6.

Refine and optimize model

Validate and check model quality!

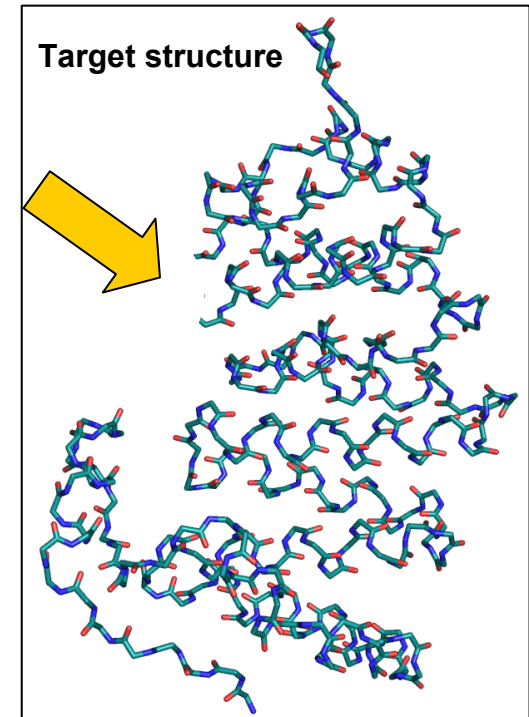
(tar

nd

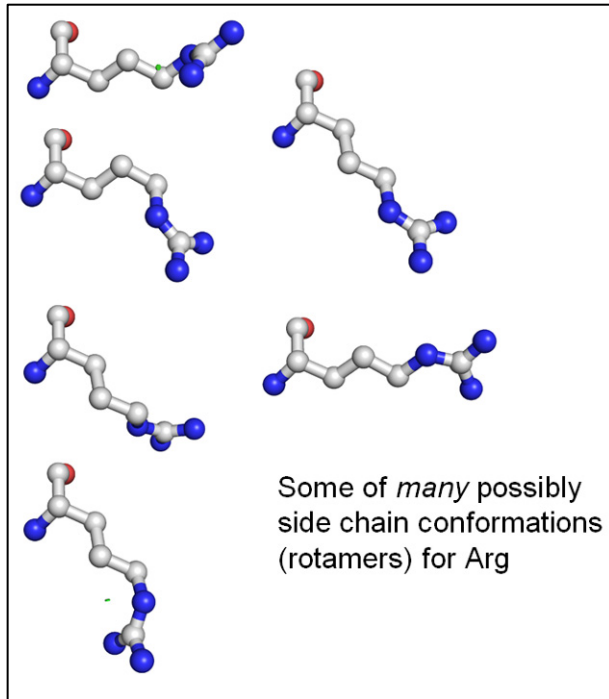
Ab initio: Generates random
loops and chooses the one with

- Lowest energy scores
- Ok Ramachandran plot
- No clashes

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



Homology modeling

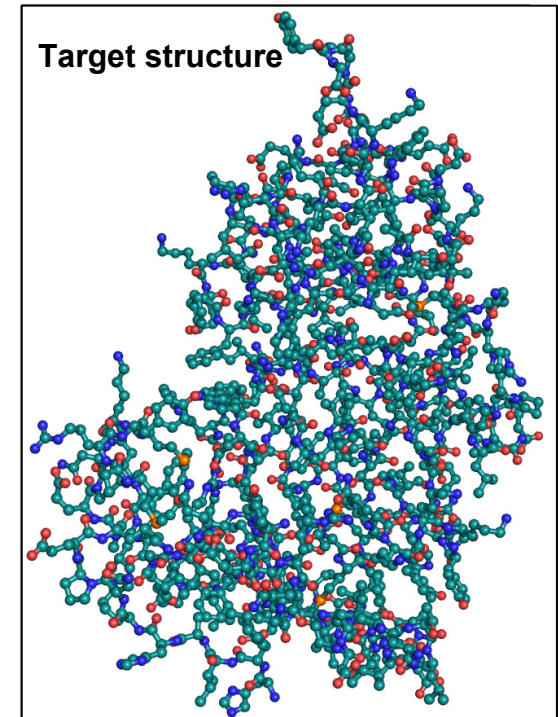


Get side chain conformations from rotamer libraries generated from known structures

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!



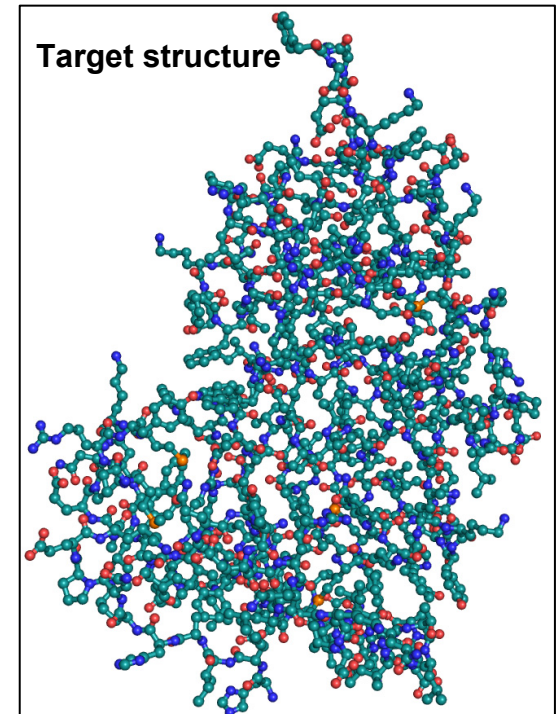
Homology modeling

Start with a protein sequence (target)

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)



Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Check if model makes sense?

- Ramachandran plot ok?
- No clashes?
- No funny bond

lengths/angles/conformations?

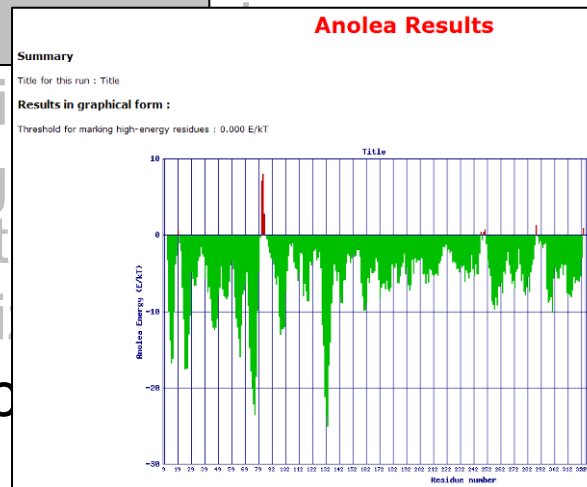
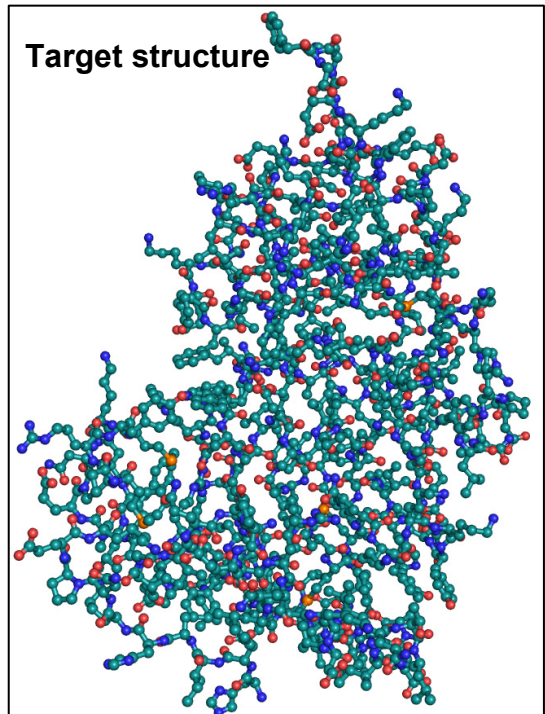
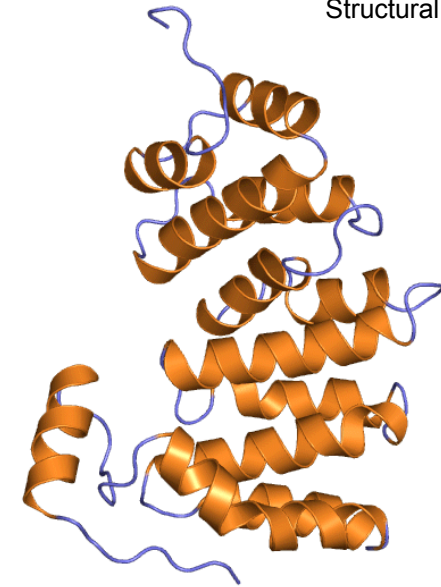
- Use programs such as:
 - Procheck
 - WHAT IF
 - ANOLEA
 - Verify3D
- These can only check if the chemical/physical properties are ok
- The model might still be 100% meaningless biologically and completely wrong!

e (target)

and align

grams

insertions



- Rotamer libraries
- Loop modeling or *ab initio* methods

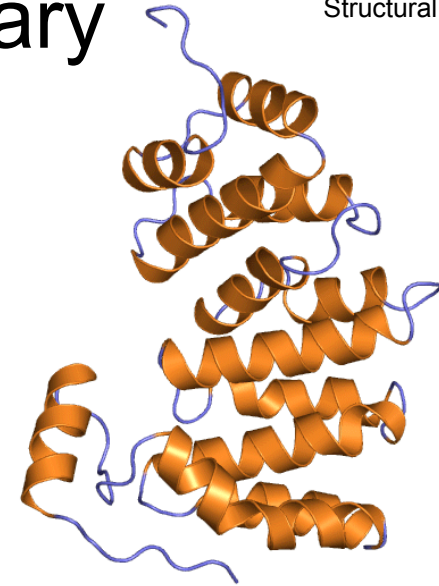
5. Refine and optimize

6. Validate and check

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 than manually generated models (if the modeler knows what she is doing...)



Tools:

- Modeller
- Swiss-Model
- 3D-JIGSAW

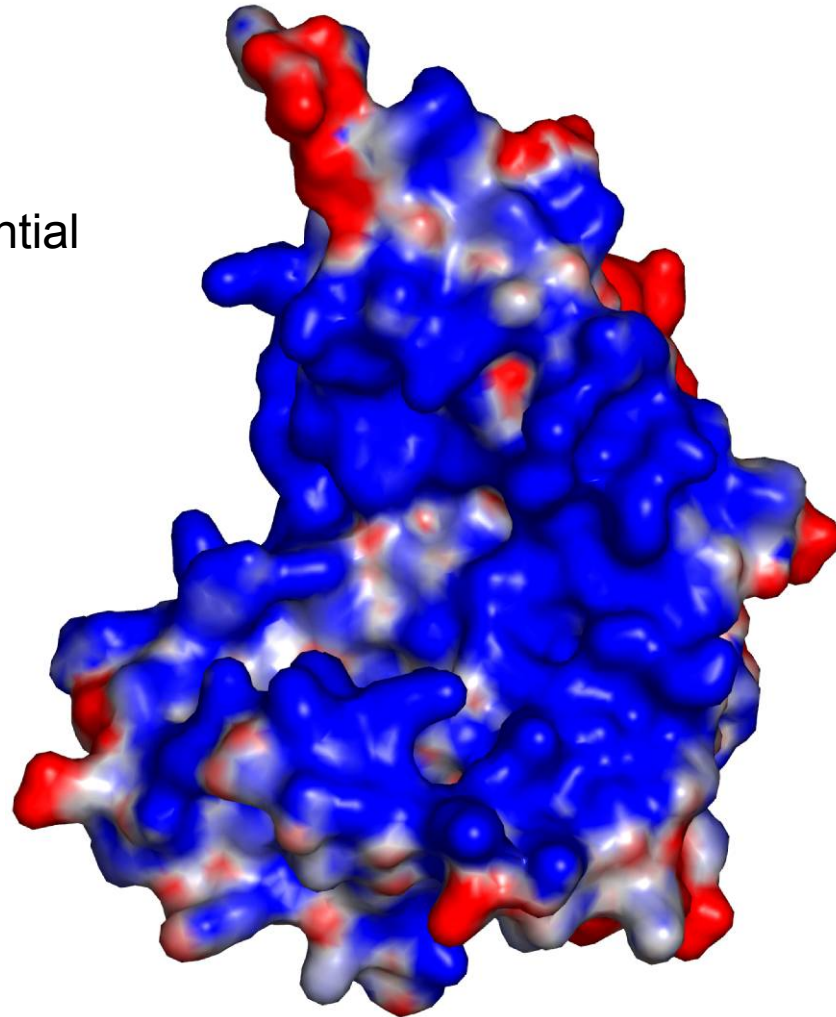
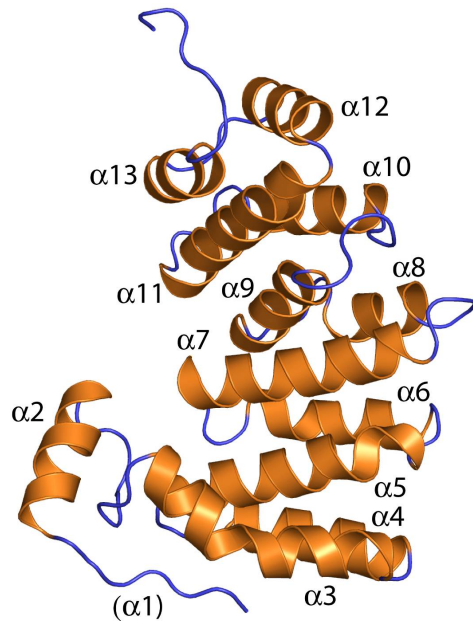
Homology model databases:

- Modbase (automatic modeling with Modeller)
- SWISS-MODEL Repository (automatic modeling with Swiss-Model)

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

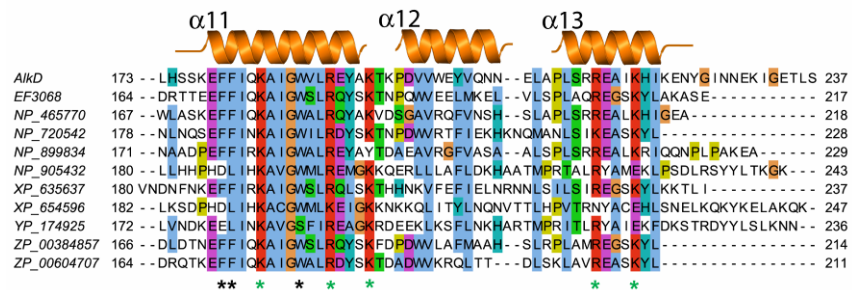
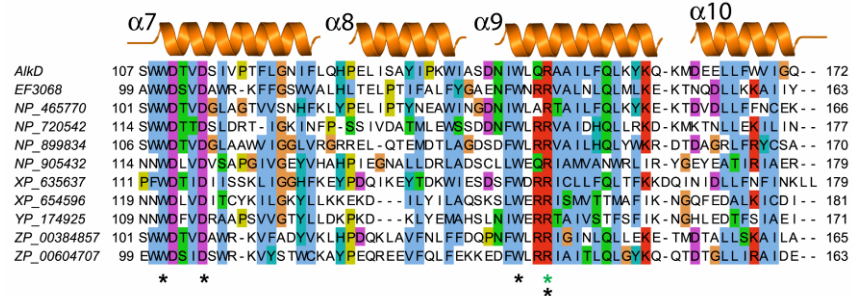
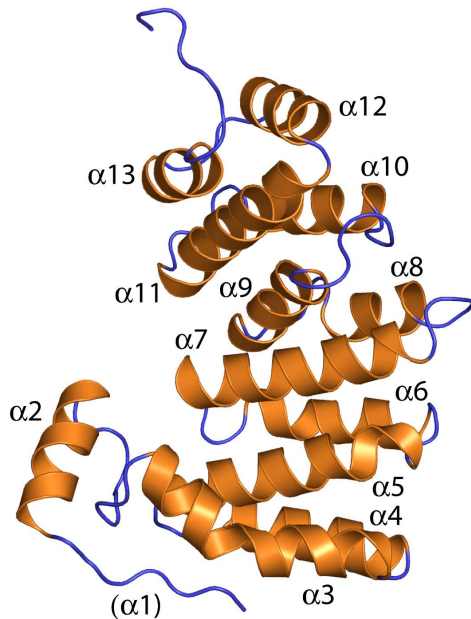
Some examples:

B. cereus AlkD electrostatic potential



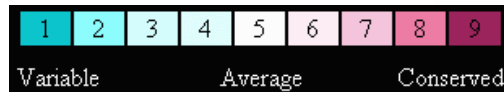
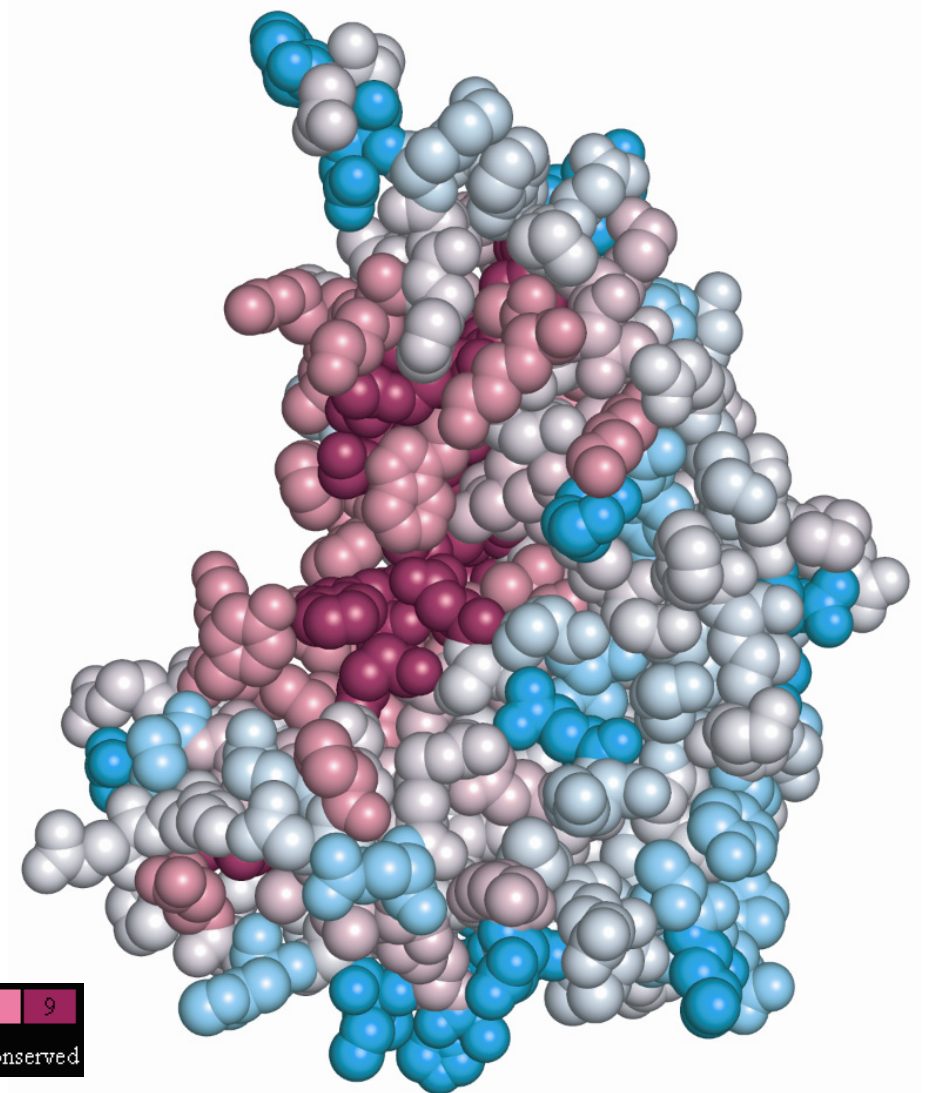
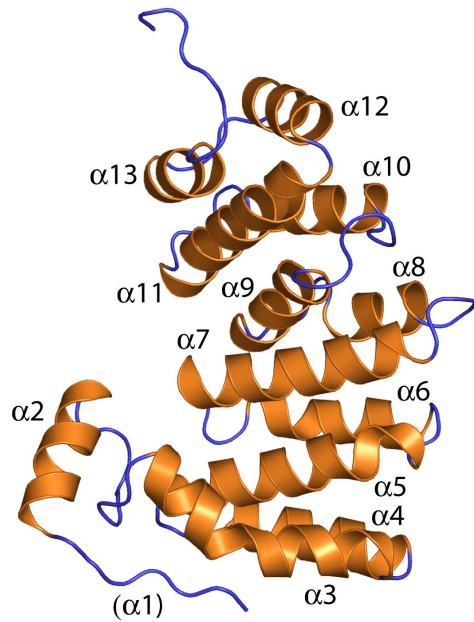
Structural bioinformatics

B. cereus AlkD sequence conservation from ConSurf:

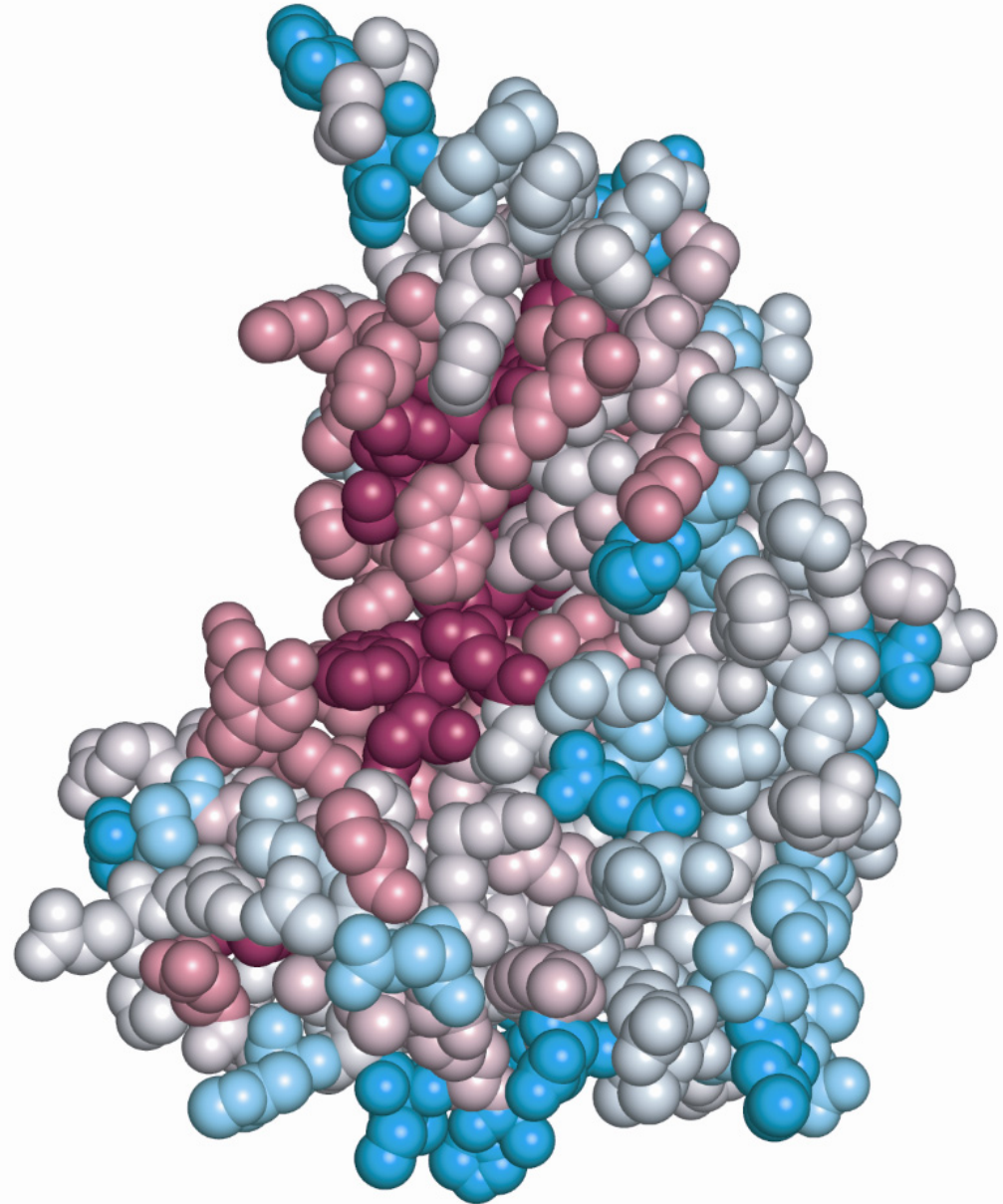
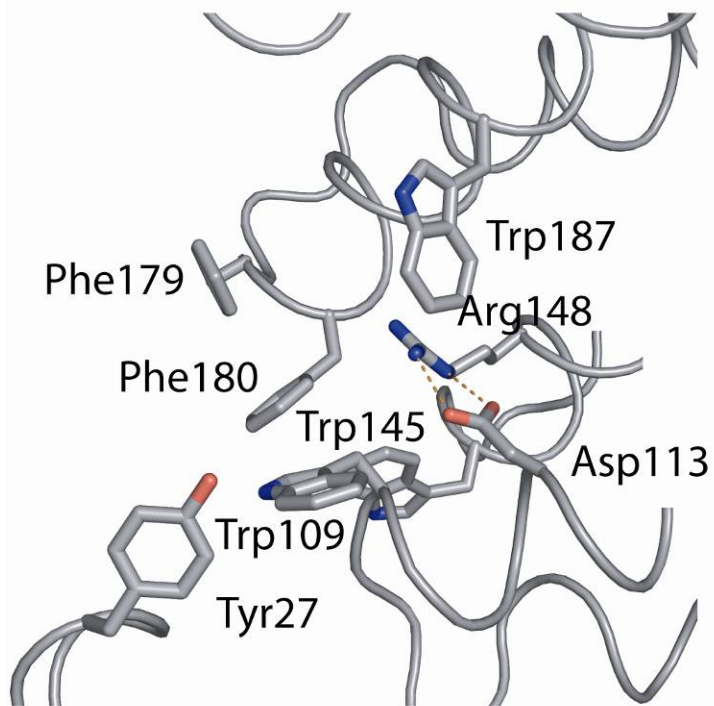


Structural bioinformatics

B. cereus sequence conservation from ConSurf:



B. cereus sequence conservation
from ConSurf:



When homology modeling does not work?

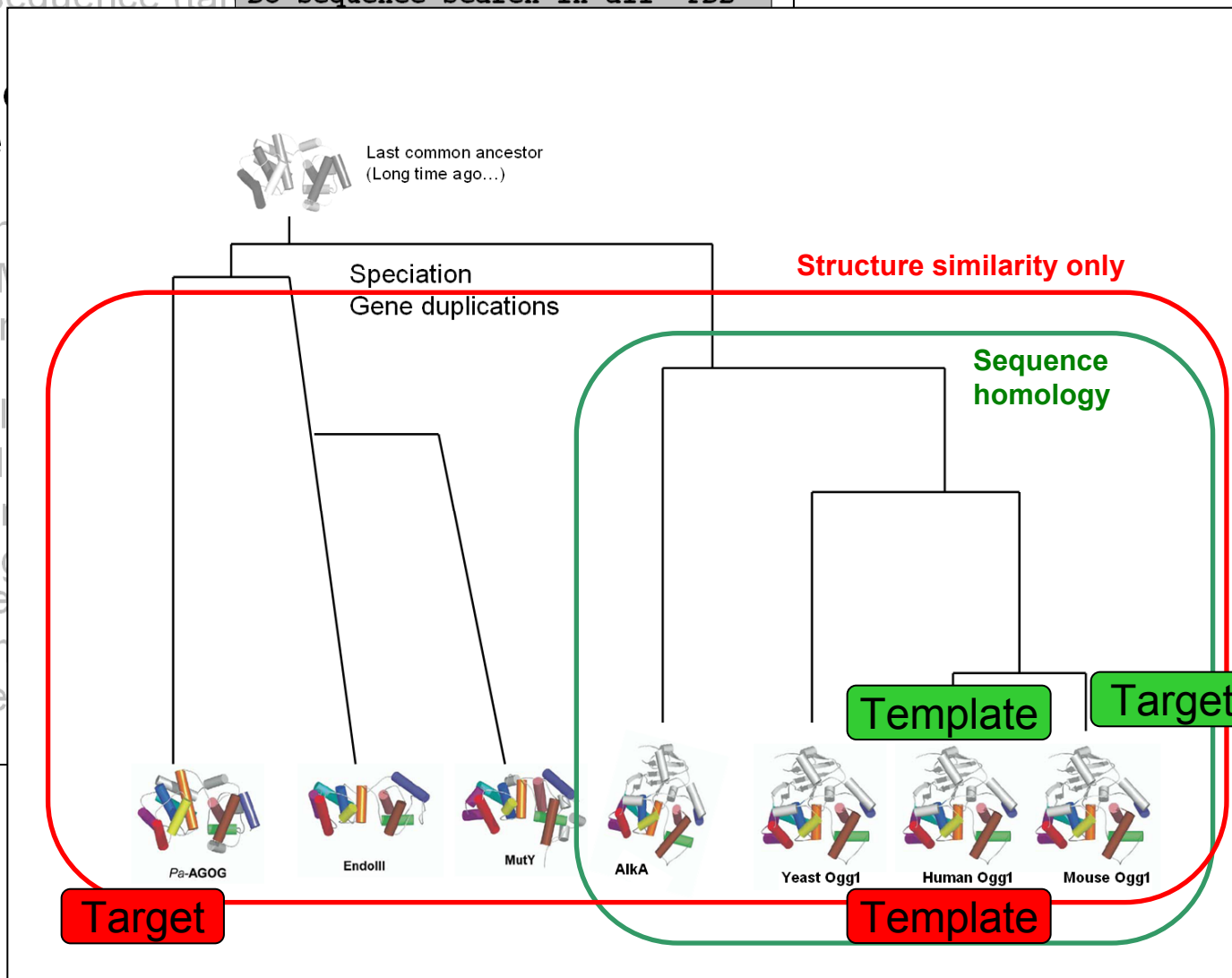
Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (far

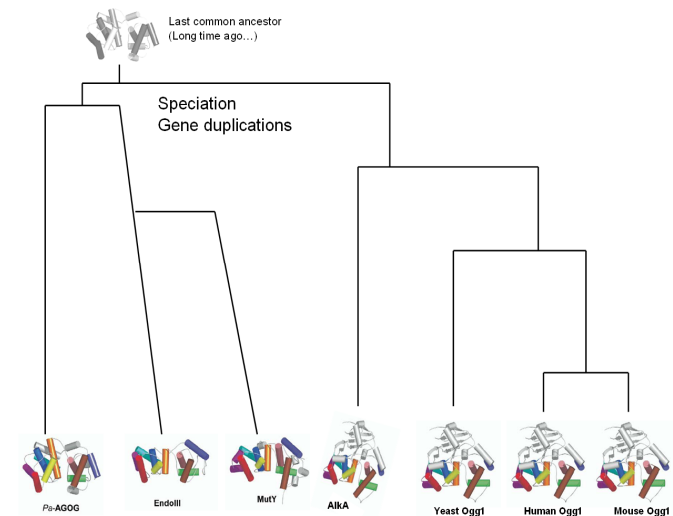
Do sequence search in all "PDB

1. Template selection
 - Find template sequences
2. Correct alignment
 - Use the best I
 - Correct placement and deletions
3. Backbone model
4. Model loops and
 - Rotamer library
 - Loop modeling or *ab initio* me
5. Refine and optim
6. Validate and che



Threading

- 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?



Threading/Fold recognition

```
>Unknown_protein  
MPARALLPRRMGHRTLA  
PSGQSFRWREQSPAHS  
RGDKSQASRPTPDELEA  
SHFQEVAQKFQGVRLLE  
GMVERLCQAFGPRLIQI  
LRKLGLGYRARYVSASA  
EAHKALCILPGVGTKVA  
IAQRDYSWHPTTSQAKG  
WAQATPPSYRCCSVPTC  
RWGTLDKIEIPQAPSPPE  
KARHPQIKQSVCTTRWC
```

What is the
structure of this?

PHYRE Protein Fold Recognition Server - Windows Internet Explorer

http://www.sbg.bio.ic.ac.uk/phyre/

File Edit View Favorites Tools Help

Links BioInfo Biology Journals Other Answers.com cbo-all Adm FUGE bioinf G03 Google UIO IT-tj. UIO PubMed SGP ABC Startsiden Wikipedia

PHYRE Protein Fold Recognition Server

phyre

Version 0.2

Protein Homology/analogY Recognition Engine

The Phyre webserver is for **Academic use only**
For in-house and/or commercial use please click [here](#)

Note: [Other tools available from our lab \(function prediction, docking, etc.\)](#)

E-mail Address

Optional Job description

Amino Acid Sequence

Quick Phyre Search

[News](#) - [Phyre Search](#) - [Help](#) - [Contact](#) - [Disclaimer](#) - [Example](#)

Done Internet 100%

Threading/Fold recognition

Quickphyre results for job AlkD - Windows Internet Explorer

D:\users\jonk\Desktop\Teaching\Quickphyre results for job AlkD_.mht

File Edit View Favorites Tools Help

Links BioInfo Biology Journals Other Answers.com cbo-all Adm FUGE bioinf G03 Google uio IT-tj, UIO PubMed SGF ABC Startsiden Wikipedia

Quickphyre results for job AlkD_

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily	Family
	c2jhnB_ (length:295) 18% i.d.	 Jmol MDL	2.9e-20	100 %	n/a	PDB header: hydrolase	Chain: B: PDB Molecule: 3-methyladenine dna-glycosylase;	PDBTitle: 3-methyladenine dna-glycosylase from archaeoglobus fulgidus
	d1orma_ (length:214) 18% i.d.	 Jmol MDL	1.7e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	Endonuclease III
	d1m3qa1 (length:190) 95% i.d.	 Jmol MDL	3.1e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	DNA repair glycosylase, 2 C-terminal domains
	d2abk_ (length:211) 17% i.d.	 Jmol MDL	3.1e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	Endonuclease III

Internet 100%

Threading/Fold recognition

The image shows two overlapping browser windows. The left window displays 'Quickphyre results for job AllK' with a 'Fold Recognition' table. The right window shows a detailed table of results for a specific protein query.

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	
	c2jhnB (length:295) 18% i.d.		2.9e-20	100 %	n/a	PDB header:hydrolase	Chai Mole methy glyco
	d1orna (length:214) 18% i.d.		1.7e-18	100 %	n/a	DNA-glycosylase	DNA
	d1m3qa1 (length:190) 95% i.d.		3.1e-18	100 %	n/a	DNA-glycosylase	DNA
	d2abk (length:211) 17% i.d.		3.1e-18	100 %	n/a	DNA-glycosylase	DNA


Category	Score	E-value	Delta G	Delta S	Delta H	Delta C	Delta T	Delta Q	Delta R	Delta A	Delta G	Delta H	Delta C	Delta T	Delta Q	Delta R	Delta A
HIGH	50.426	0.0004	-162.3	-4.7	205.0	129	144	345	1ngnA0	a.96.1.2							
HIGH	50.375	0.0004	-230.5	-1.5	189.0	141	182	345	2ofkA0	-							
MEDIUM	45.432	0.001	-198.1	-4.1	160.0	142	187	345	1lmzA0	a.96.1.4							
MEDIUM	45.427	0.001	-222.5	-4.7	155.0	133	186	345	2jq6A0	-							
LOW	35.934	0.012	-335.6	-4.2	42.0	259	356	345	2pgeA0	-							
LOW	34.659	0.016	-226.5	-6.1	76.0	147	179	345	2i10A0	a.4.1.9 a.121.1.1							

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

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

**I
M
P
O
R
T
A
N
T**

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 \(2008\)](#)

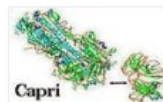
Proceedings

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](#)



[Automatic evaluation of prediction servers](#)

Message Board

CASP8 job fair

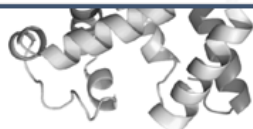
[Dear CASPers, we would like to take advantage of the CASP8 meeting to help young scientists looking for a position in Computational Biology meeting potential group leaders and vice versa. We pla ...](#)

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 ...](#)

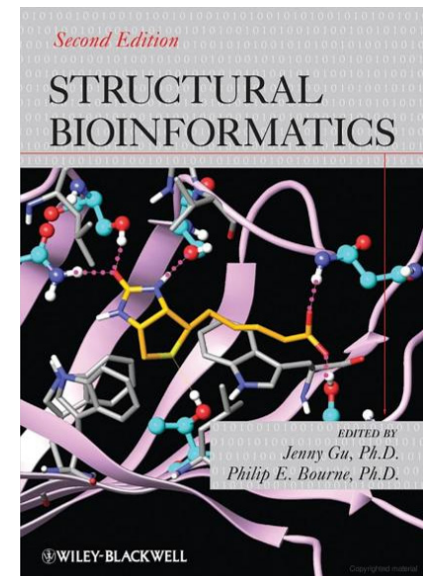
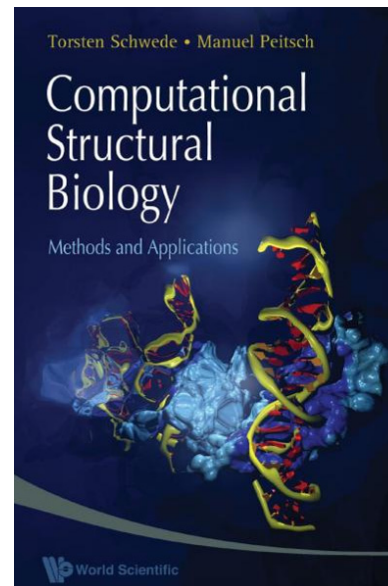


Literature & Useful links

- 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