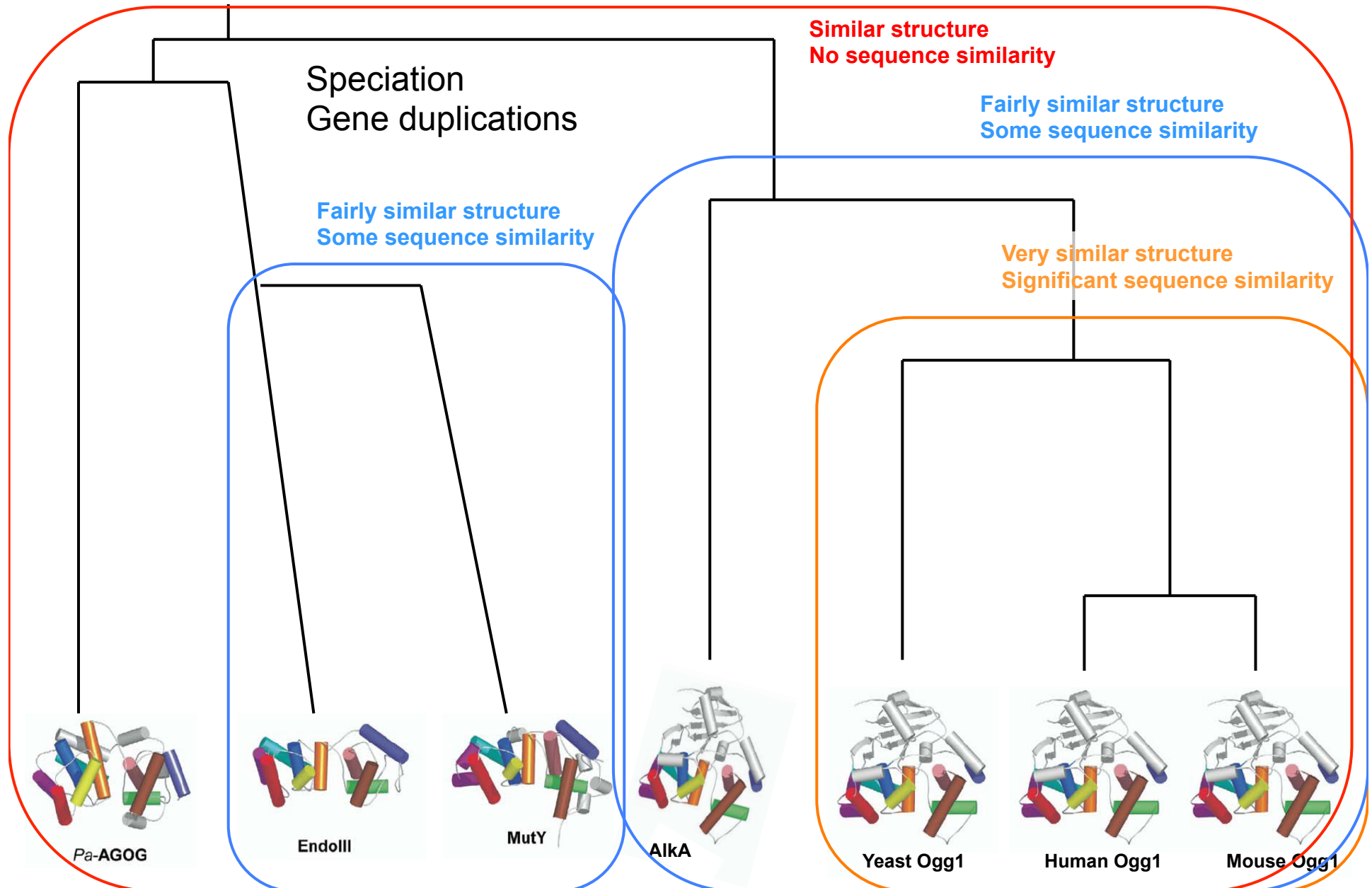




Last common ancestor
(Long time ago...)

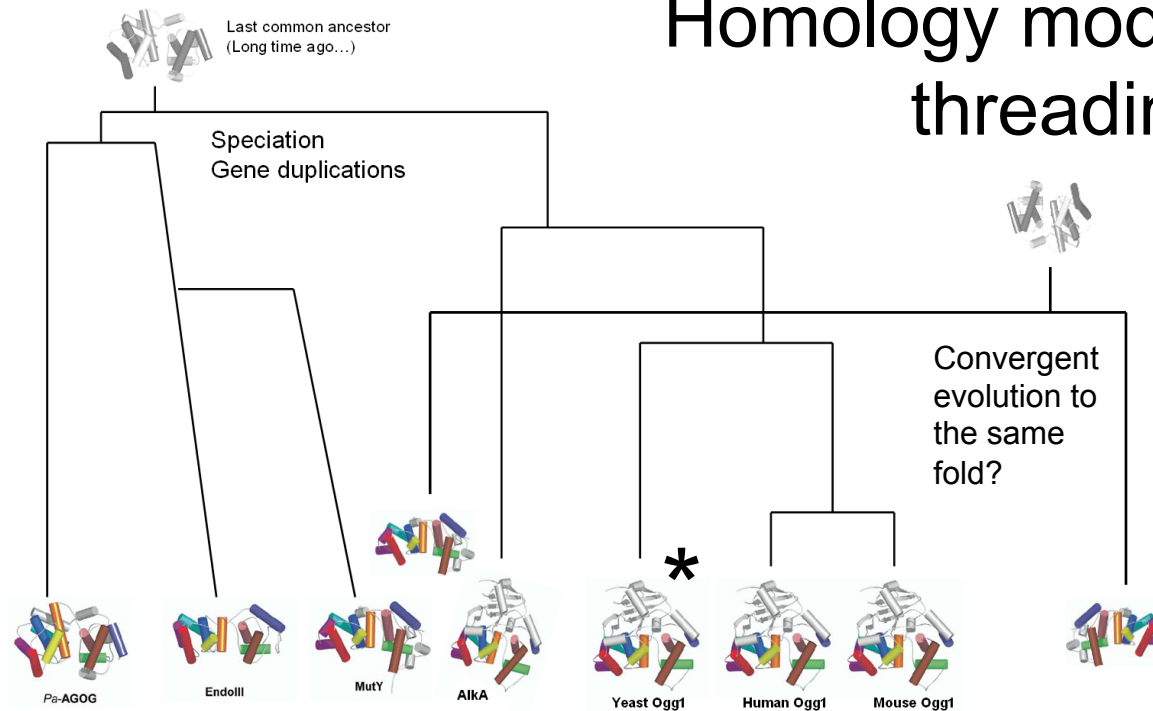
Protein structure evolution

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 (actually domains) in a superfamily have the same overall structure/fold
- If we know (from experiment) the structure of one 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 Biology Knowledgebase website. At the top, there is a navigation bar with links for Home, Resource Hubs, Current Focus, Services, About Us, and Contact Us. Below this is a search bar with options to search by sequence, text, pdb id, or uniprot ac. The main content area is divided into several sections:

- Current Focus:** Membrane Proteome. It features three sub-sections: Featured System, Research Advances, and Technical Highlight.
- Membrane Proteome: A Cap on Transport:** Structural and biophysical analyses of a membrane carrier ectodomain reveal a Ca²⁺-dependent switch that regulates ATP flux across the mitochondrial inner membrane.
- Membrane Proteome: Pumping Out Heavy Metal:** Crystal structures of a heavy-metal efflux pump reveal intermediate transport states.
- Protein Structure Initiative Corner:** Collaborative Network, Publications, Latest News, Community Nominations, and SBKB Tools.
- Discoveries, Structural Targets, Structure, Sequence & Function, Membrane Proteins, Homology Models, Methods & Technologies:** A grid of six red-themed boxes with icons representing different research areas.
- Latest PSI Results:** A table showing statistics on new structures, total structures, and community structures.
- Latest Structures:** A section highlighting recent structures, such as the human P2Y12 receptor in complex with an antithrombotic drug (PDBID: 4NTJ).

Traditionally:
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

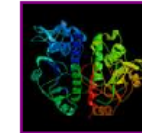
Structural genomics/The Protein Structure Initiative (PSI)

Jon K. Lærdahl,
Structural Bioinformatics

Security /
Privacy Notice



Midwest Center
for
Structural Genomics



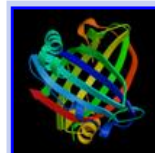
Structure Gallery

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

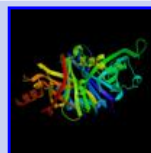
Consortium
Project
Investigators
Targets
3-D Structures
Related Publications
SG Sites
SG Progress
NIH
MCSG Resources
Job opportunities
Collaborators
Internals
Technologies

GALLERY OF MCSG STRUCTURES IN PDB

959 targets in PDB (28 new folds)



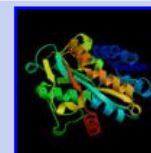
APC006 [\[ref\]](#)
[LSQE](#) ident: 23.9%
[annotation](#)



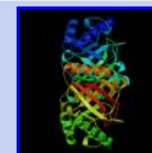
APC007
[LXBW](#) ident: 64.5%
[annotation](#)



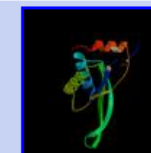
APC008
[2AP3](#) ident: <20%
[annotation](#)



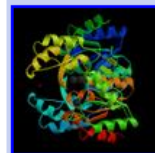
APC009 [\[ref\]](#)
[1P99](#) ident: <20%
[annotation](#)



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



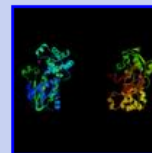
APC012 [\[ref\]](#)
[1KR4](#) ident: <20%
[annotation](#)



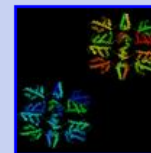
APC014 [\[ref\]](#)
[1KYT](#) ident: <20%
[annotation](#)



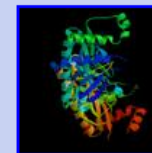
APC037 [\[ref\]](#)
[1KXJ](#) ident: 100%
[annotation](#)



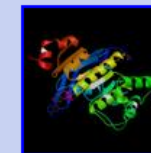
APC038 [\[ref\]](#)
[1M6Y](#) ident: <20%
[annotation](#)



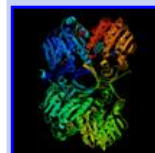
APC042
[1WPB](#) ident: <20%
[annotation](#)



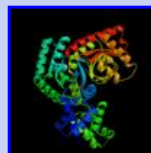
APC043 [\[ref\]](#)
[1KUT](#) ident: <20%
[annotation](#)



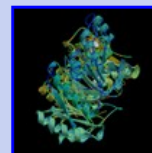
APC046
[1J10](#) ident: 33.5%
[annotation](#)



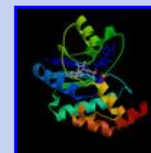
APC047 [\[ref\]](#)
[1JQ3](#) **New Fold**
[annotation](#)



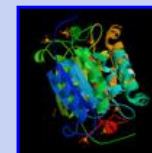
APC048 [\[ref\]](#)
[1MKM](#) ident: <20%
[annotation](#)



APC049
[1T57](#) 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

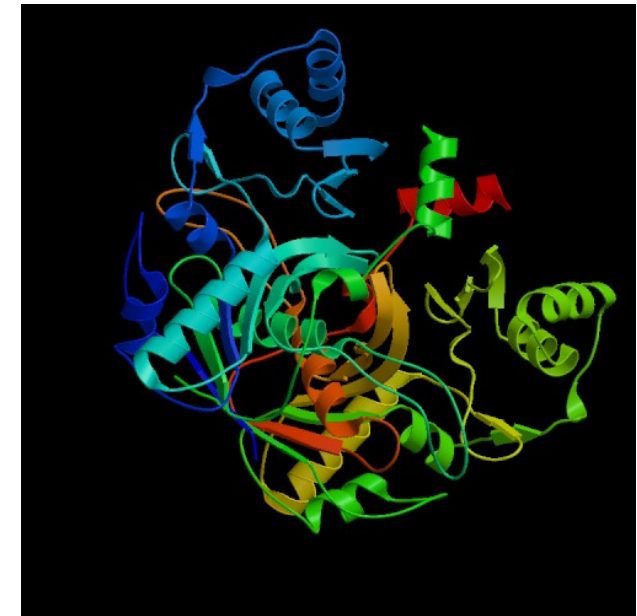
Structure Gallery

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

GALLERY OF M-CSG 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] 1KYI ident: <20% annotation	 APC037 [ref] 1KXJ ident: 100% annotation	 APC038 [ref] 1MGV ident: <20% annotation	 APC042 1WPE 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



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

10 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*)

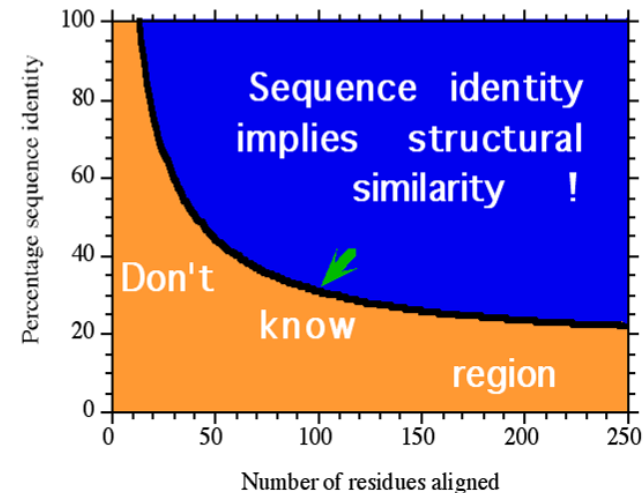
&

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 (in textbooks)
- 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!
```

2. Correct alignment

- Use the best alignment
- Correct placement and deletion

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

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!

```
MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQT  
PERRQLLKDVIQIHTLPDPKDFRIIVRELWDLPEREFQA  
AALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVP  
TFLGNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLK  
YKQKMDEELLFWVIGQLHSSKEFFIQKAIGWVLREYAKT  
KPDVVWEYVQNNELAPLSRREAIAIKHIKENYGINNEKIGE  
TLS
```

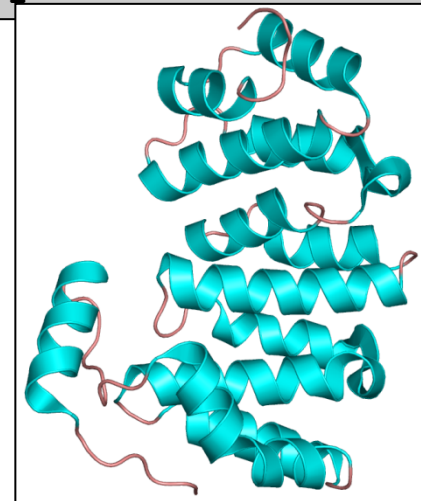

Homology modeling

Start with a protein sequence (target)

1. Template selection:
 - Find template in PDB and other 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

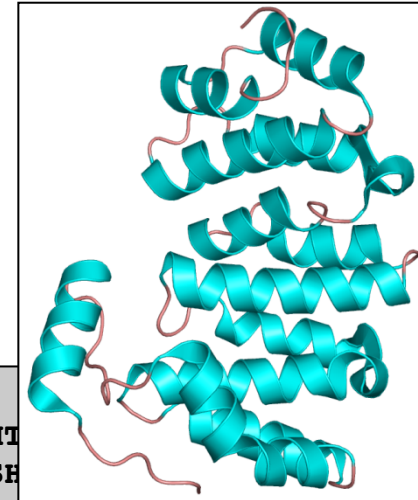


Homology modeling

Start with a protein sequence (target)

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

2. Correct alignments



Sequence alignment

```

Bc_Alkd  MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQTPERRQLLKDVIQIHT
EF3068   -----MDTLQFQKNPETAAKMSAYMKHQFVFAGIPAPERQALSKQLKKESH
           :  :  :****.*  * :  ***:*** * ** :***: * *:::  :.
Bc_Alkd  FRIIVRELWDLPERFQAAALDMMQKYKYINETHIPFLEELIVTKSWWDTVDSIVPTFL 120
EF3068   LCQEIEAYYQKTEREYQYVAIDLALQNVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG 122
           :  :.  :. :****:*  .*: : : :.  .:  : :  : * :***:**:  *
Bc_Alkd  GNIFLQHPELISAYIPKWIASDNIWLQRAILFQLKYKQKMDEELLFWVIGQLHSSKEFF 180
EF3068   SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF 171
           . : * :  **:.  :. :  :*:** :*. : **  *:*  :*:**  .*  :*:**
Bc_Alkd  IQKAIGWVLREYAKTKPDVWVEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068   IQKAIGWSLRQYSKTNPQVELMKELVLSPLAQREGSKYLAKASE----- 217
           ***** **:*:**: * * :: : ***:**. * : :
    
```

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

- 3.
- 4.
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 other 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

```

Sequence alignment
Bc_Alkd  MHPFVKALQEHFIAHKNPEKAEPMARY
EF3068  -----MDTLQFQKNPETAAKMSAY
          : : :****.* *: *
Bc_Alkd  FRIIVRELWDLPEREFQAAALDMMQYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL 120
EF3068  LCQEIFAIYQKTEREYQVVAIDLALQNVQRFSLSEVVAFKAYVPQKAWWDSVDAWRKFFG 122
          :  :.  :. :***:* .***:  :  :.  :.  :  : *:*:*:*: *
Bc_Alkd  GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF 180
EF3068  SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEFF 171
          . : * : **:.  :.  :  :***: *:*: **  ** :***  .*  :***
Bc_Alkd  IQKAIGWVLRVAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068  IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----- 217
          ***** **:***:*: * * :. : *:*:*:*: *:: :
    
```

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

3. Back
4. Mod
 - R
 - L
5. Refi
6. Valid

Homology modeling

Obtaining the correct alignment

- 1.
- 2.
- 3.
- 4.
- 5.
- 6.

Where is the correct position of the gap?

The MSA gives the answer!!

XP_968170_Tribolium_castaneum	K	T	L	G	T	G	S	F	G	R	V	M	I	V	Q	H	K	P	T	-	K	E	Y	Y	A	M	K	I	L	D	K	K	
Q4JIV3_Lymnaea_stagnalis	K	T	L	G	T	G	S	F	G	R	V	M	L	V	Q	H	K	G	E	N	K	A	Y	Y	A	M	K	I	L	D	K	K	

XP_968170_Tribolium_castaneum	K	T	L	G	T	G	S	F	G	R	V	M	I	V	Q	H	K	-	P	T	K	E	Y	Y	A	M	K	I	L	D	K	K	
Q4JIV3_Lymnaea_stagnalis	K	T	L	G	T	G	S	F	G	R	V	M	L	V	Q	H	K	G	E	N	K	A	Y	Y	A	M	K	I	L	D	K	K	

Homolog2_Petromyzon_marinus	K	T	L	G	T	G	S	F	G	R	V	M	L	V	K	H	K	-	A	T	D	R	Y	F	A	M	K	I	L	D	K	K		
ENSCJAP00000040924_Callithrix_jacchus	K	T	L	G	I	G	S	F	G	R	V	V	L	V	S	H	R	-	E	S	G	S	H	Y	A	M	K	I	L	N	K	E	K	
P22612_Homo_sapiens	R	T	L	G	M	G	S	F	G	R	V	M	L	V	R	H	Q	-	E	T	G	G	H	Y	A	M	K	I	L	N	K	K		
XP_968170_Tribolium_castaneum	K	T	L	G	T	G	S	F	G	R	V	M	I	V	Q	H	K	-	P	T	K	E	Y	Y	A	M	K	I	L	D	K	K		
Q4JIV3_Lymnaea_stagnalis	K	T	L	G	T	G	S	F	G	R	V	M	L	V	Q	H	K	G	E	N	K	A	Y	Y	A	M	K	I	L	D	K	K		
ENSTRUP00000015108_Takifugu_rubripes	K	T	L	G	T	G	S	F	G	R	V	M	L	V	K	H	K	-	E	T	N	Q	F	Y	A	M	K	I	L	D	K	K		

ant step!! in
t, template
(50-100?) of
ine, T-
a different
e alignment
e template
all indels
heets

DSIVPTFL 120
DAWRKFFG 122
*: *
LHSSKEFF 180
DRTTEEFF 171
:::***

Bc_AlkD IQKAIGWVLR~~EVAKT~~KPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068 IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----- 217
***** **:***: * * ::: *:::***. *:::

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

Homology modeling

Start with

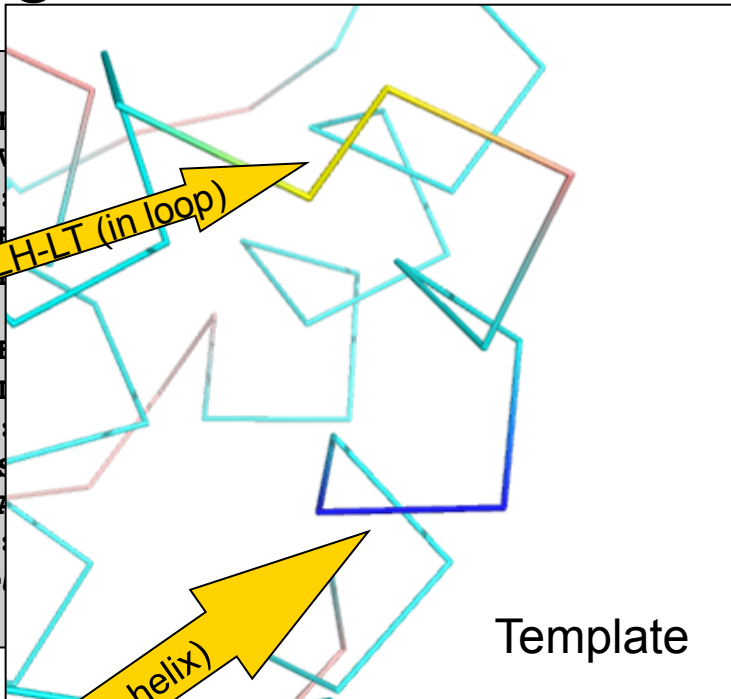
1. Template selection
2. Correction of insertions and deletions

Sequence alignment

```

Bc_Alkd  MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFI
EF3068  -----MDTLQFQKNPETAAKMSAYMKHQFY
          : : :****.* *: ***:.*
Bc_Alkd  FRIIVRELWDLPEREFQAAALDMMQYKKYIN
EF3068  LCQEIEAYIQKTEREYQYVAIDLALON
          : . : : .***:* .*: : : .
Bc_Alkd  GNIFLQHPELISAYIPKWIASDNIWLQRAAIL
EF3068  SWVALH-LTELPTIFALFYGAENFWNRRVALN
          . : * : . : : . : : : : * : * :
Bc_Alkd  IQKAIGWVLRVAKTKPDVVWEYVQNNELAPLS
EF3068  IQKAIGWSLRQYSKTNPQWVEELMKELVLSPL
          ***** **:*:**:*: * * : : : * : * :
    
```

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



3. Backbone
4. Model
5. Refinement
6. Validation

Sequence alignment

```

Bc_Alkd  MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFIQTERRQLLKDVIIQIHTLPDPKD 60
EF3068  -----MDTLQFQKNPETAAKMSAYMKHGI PAPERQALSKQLLKESHTWPKEK 52
          : : :****.* *: * * :***: * * : : : :
Bc_Alkd  FRIIVRELWDLPEREFQAAALDMMQYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL 120
EF3068  LCQEIEAYIQKTEREYQYVAIDENVQRFSLSEEVVAFKAYVPQKAWWDSVDAWRKFFG 122
          : . : : .***:* .*: : : . : : : : * : * : * :
Bc_Alkd  GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF 180
EF3068  SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF 171
          . : * : * : . : : : : * : * : * : * : * : * : * : * : * :
Bc_Alkd  IQKAIGWVLRVAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS 237
EF3068  IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAOREGSKYLAKASE----- 217
          ***** **:*:**:*: * * : : : * : * : * : * : * :
    
```

Alignment of the sequences of *B. cereus* AlkD (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!

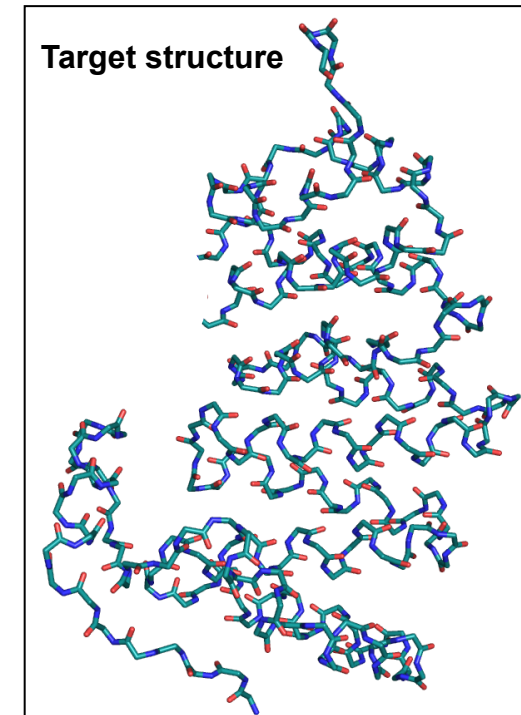
Homology modeling

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

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

1.

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

2.

- 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!

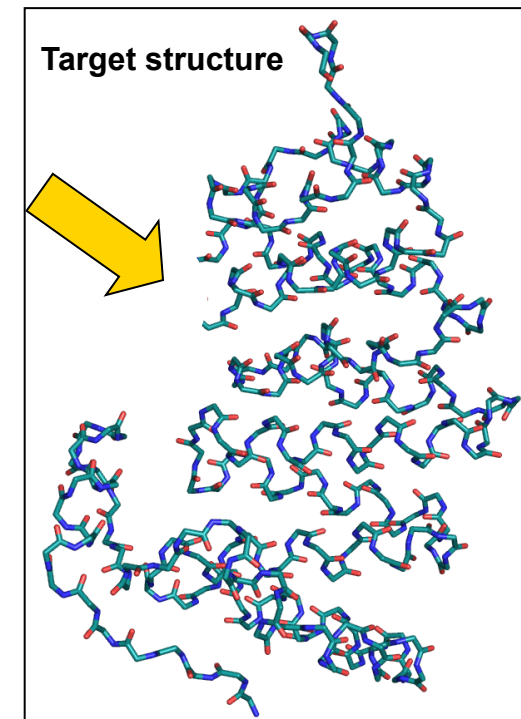
(tar

Ab initio: Generates random
loops and chooses the one with

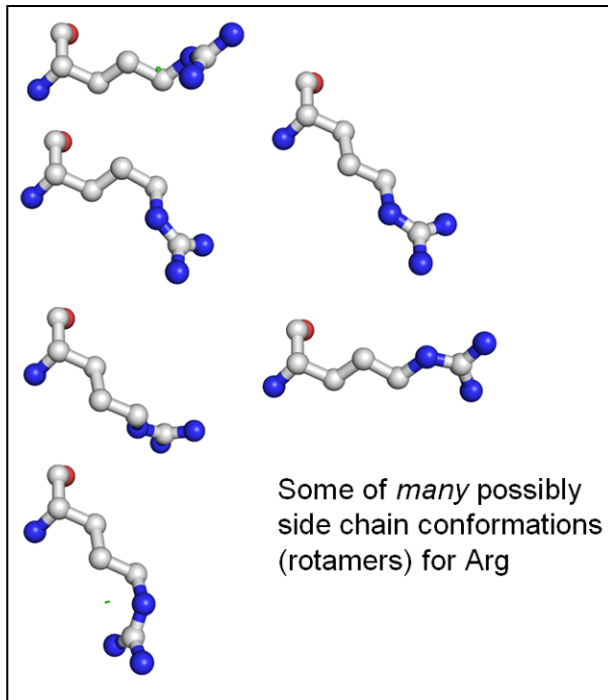
- Lowest energy scores
- Ok Ramachandran plot
- No clashes

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

nd



Homology modeling

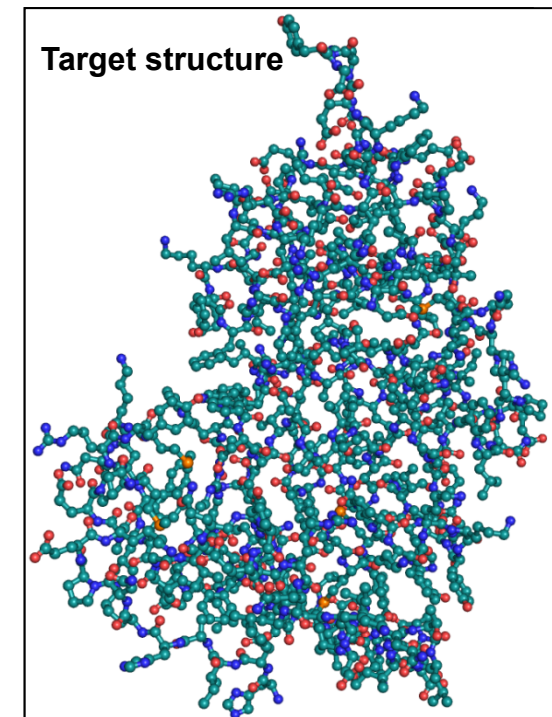


Get side chain conformations from rotamer libraries generated from known structures

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!



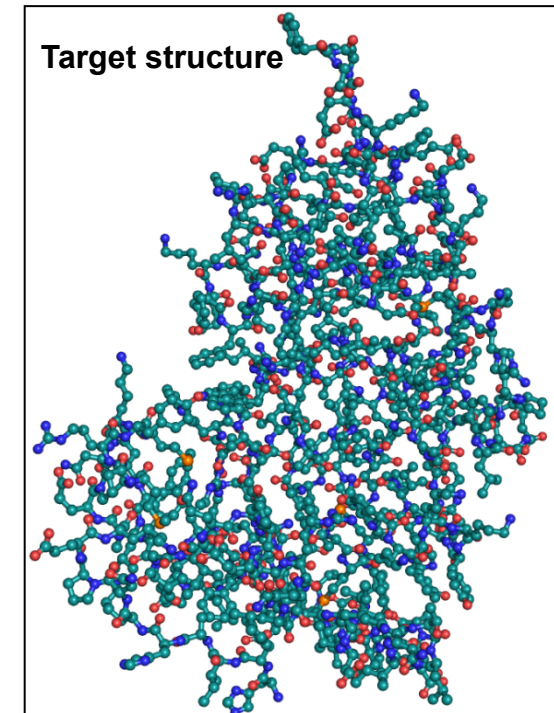
Homology modeling

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?

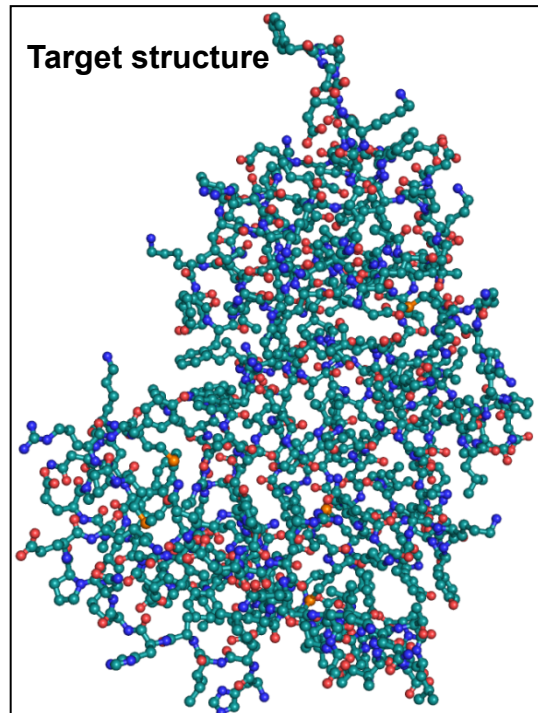
- Ramachandran plot ok?
- No clashes?
- No funny bond lengths/angles/conformations?
- Use programs such as:
 - Procheck
 - WHAT IF
 - ANOLEA
 - Verify3D
- These can only check if the chemical/physical properties are ok
- The model might still be 100% meaningless biologically and completely wrong!

e (target)

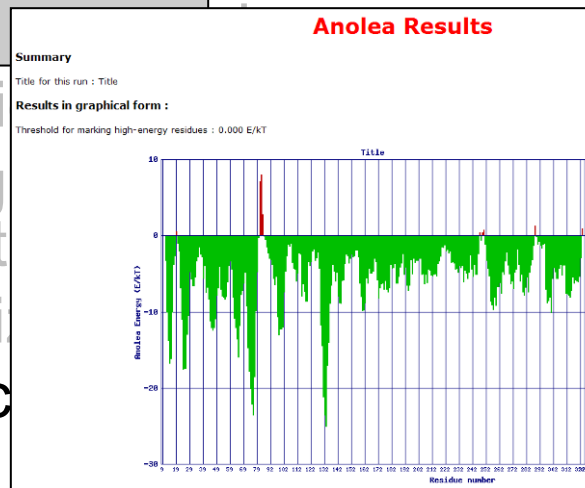
and align

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)

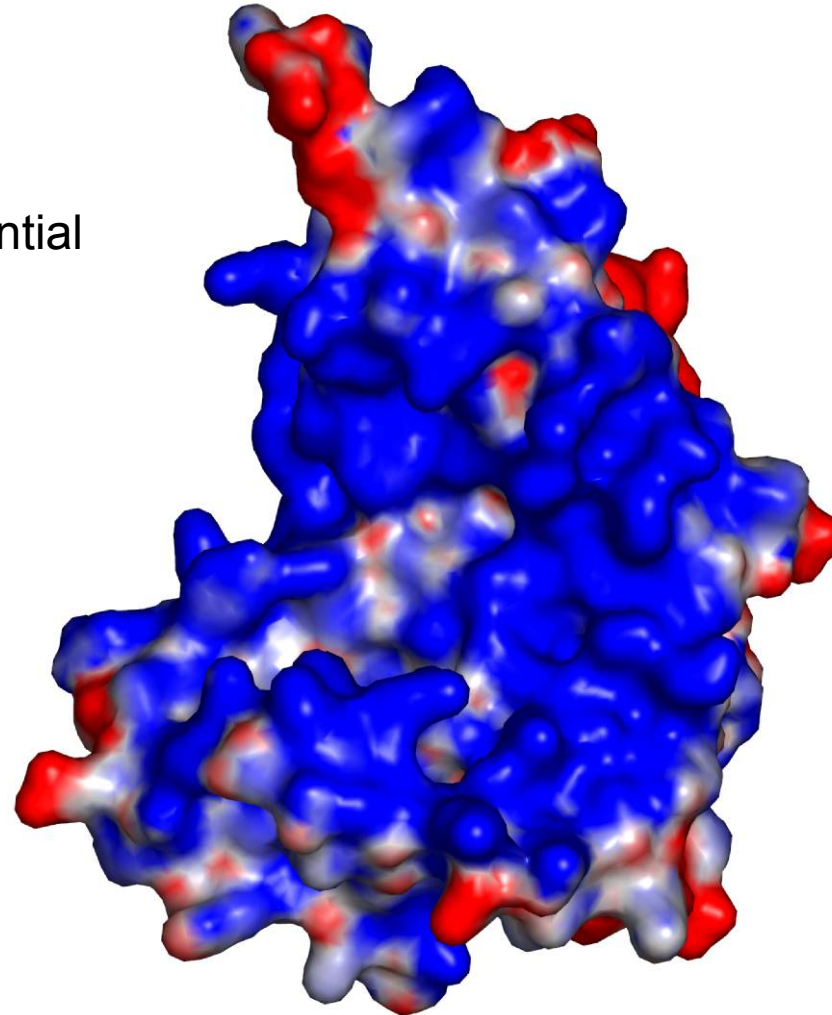
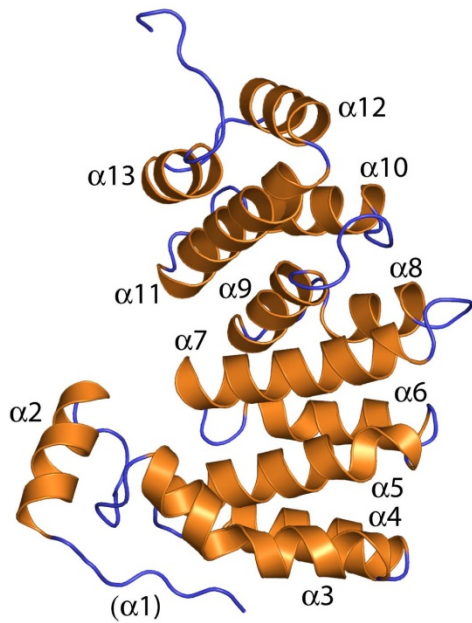
Structural bioinformatics

Jon K. Lærdahl,
Structural Bioinformatics

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

Some examples:

B. cereus AlkD electrostatic potential



Structural bioinformatics

Jon K. Lærdahl,
Structural Bioinformatics

B. cereus AlkD sequence
conservation from ConSurf:

