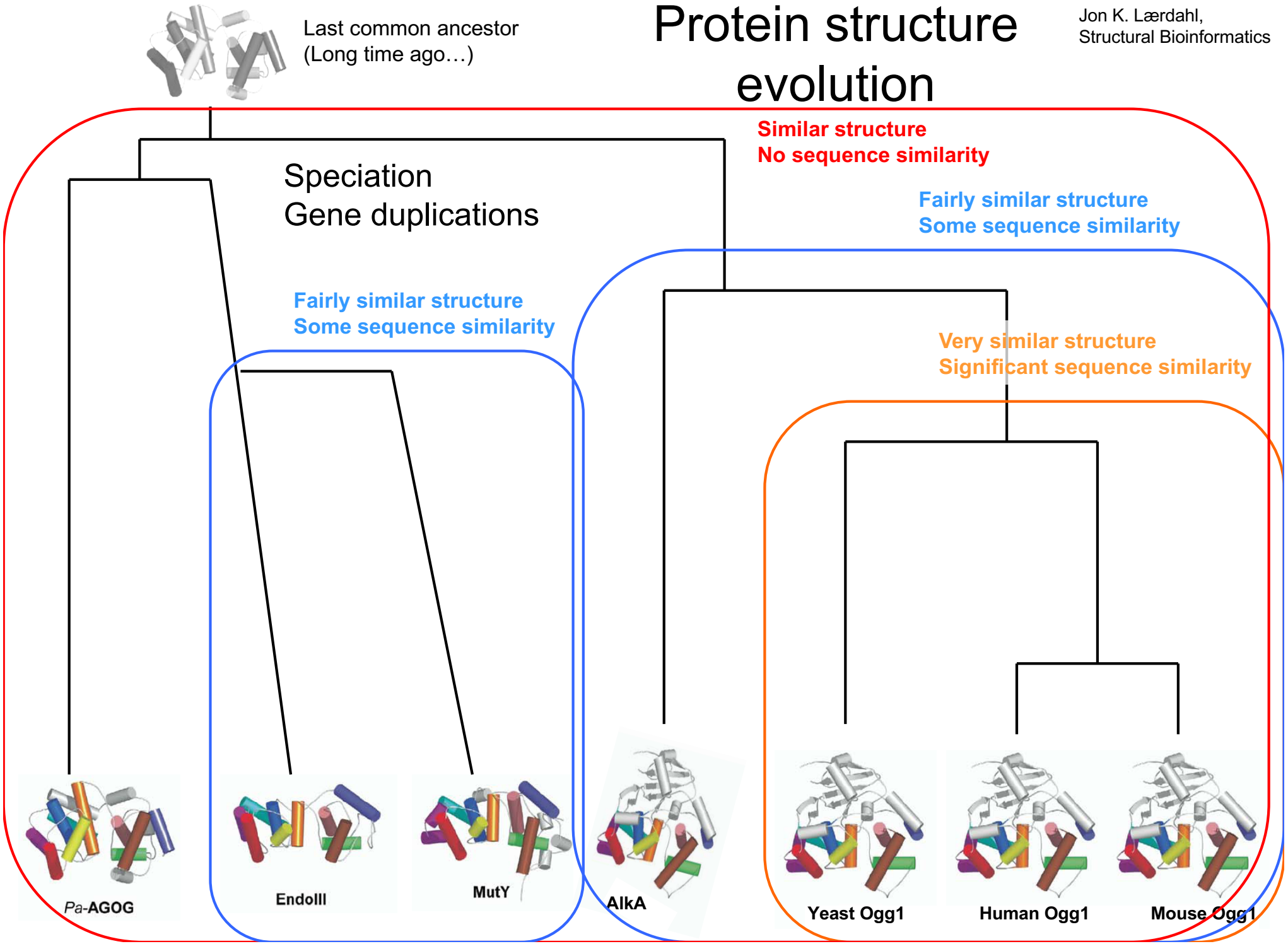
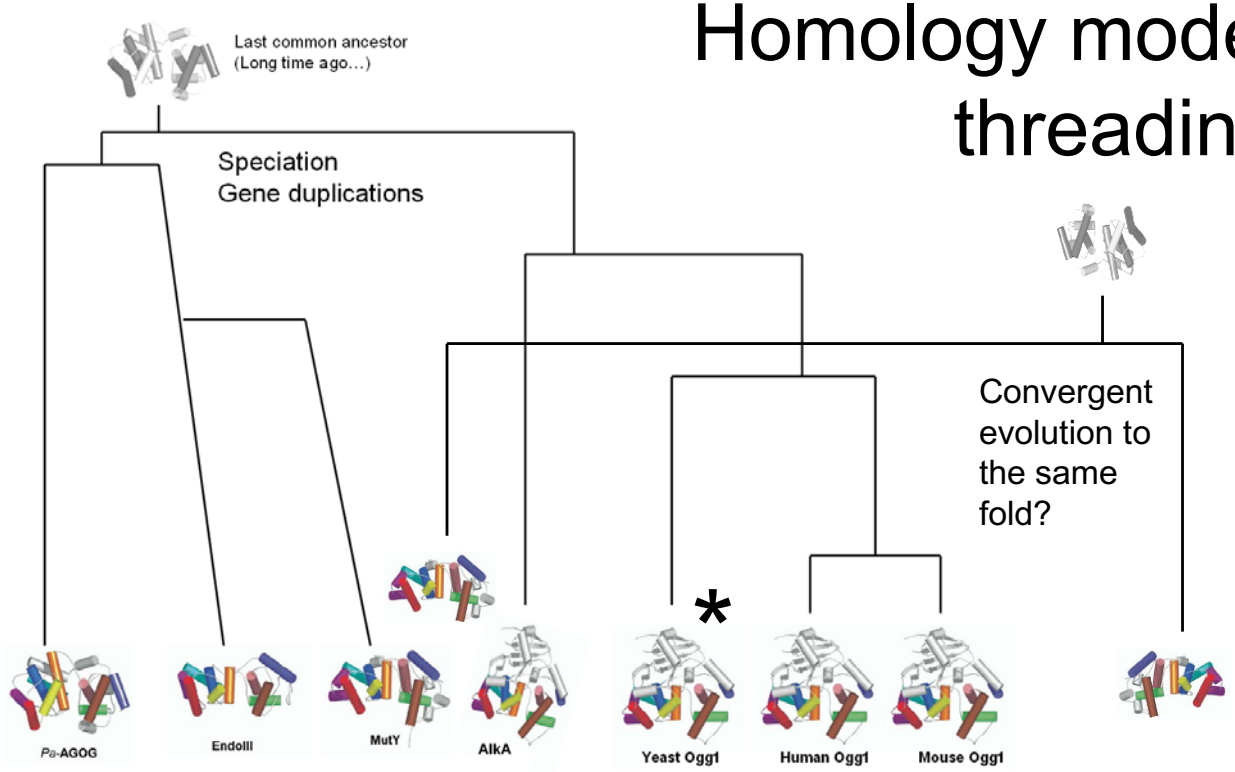
Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics



Homology modeling and threading

Jon K. Lærdahl,
Structural Bioinformatics



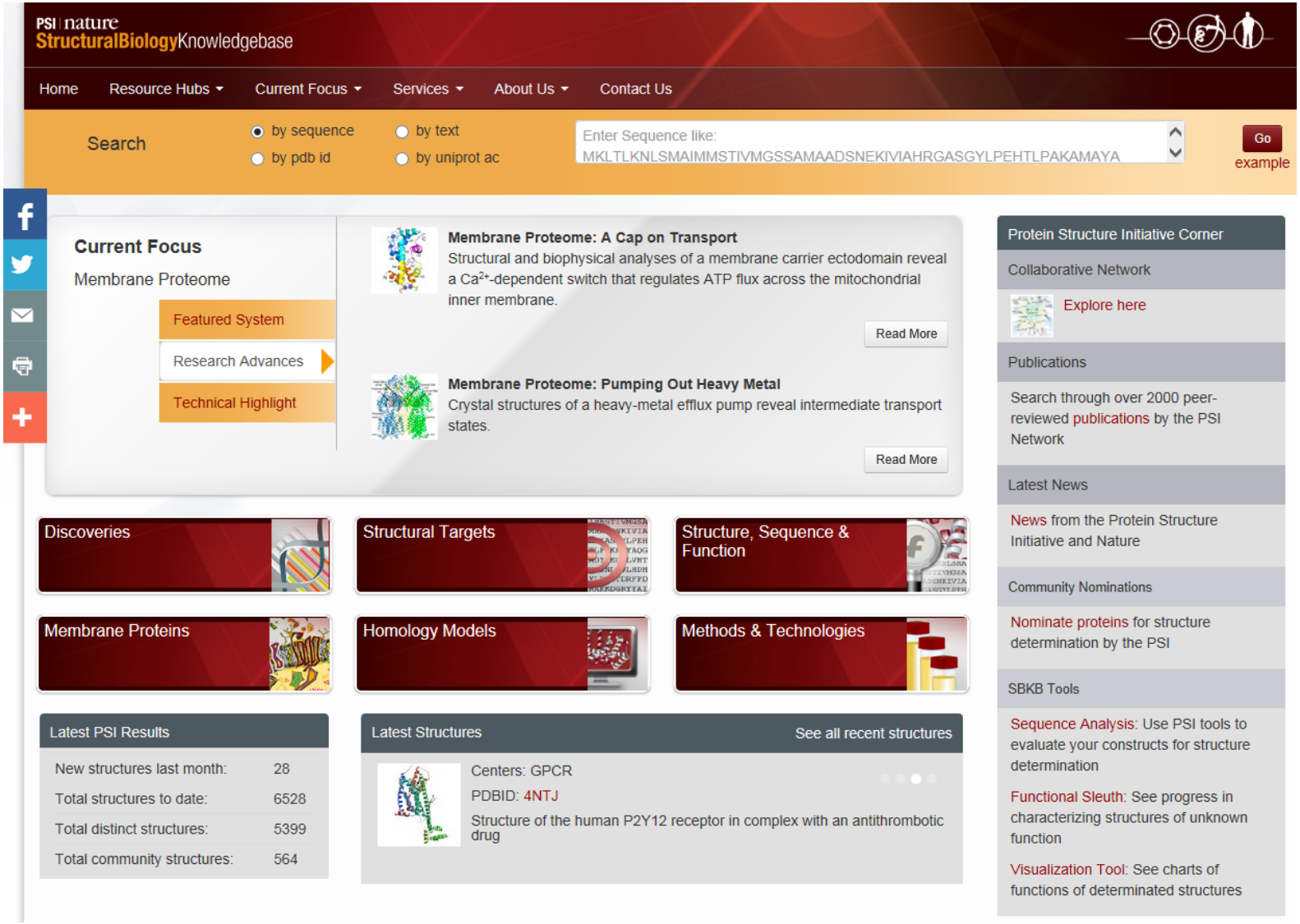
Important goal to have at least one structure in all structural superfamilies!

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



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

MCSG

Midwest Center for Structural Genomics

PSI

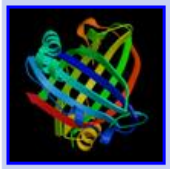
Structure Gallery

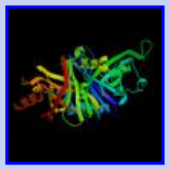
• XML Files • Target List • Progress • Statistics • Log in • Site Search: Go

- 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

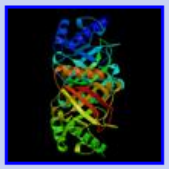
959 targets in PDB (28 new folds)


APC006 [ref]
1SQE ident: 23.9%
annotation

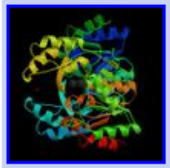

APC007
1XBW ident: 64.5%
annotation

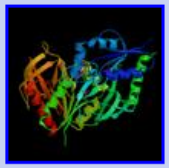

APC008
2AP3 ident: <20%
annotation

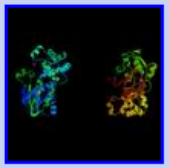

APC009 [ref]
1P99 ident: <20%
annotation

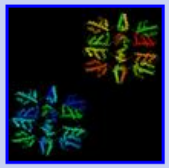

APC010 [ref]
1NG5 New Fold
annotation


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


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

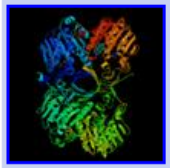

APC037 [ref]
1KXJ ident: 100%
annotation

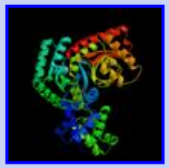

APC038 [ref]
1M6Y ident: <20%
annotation

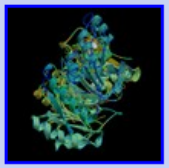

APC042
1WPB ident: <20%
annotation


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

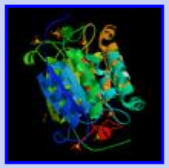

APC046
1JI0 ident: 33.5%
annotation


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

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!

PSI concluded in 2015 (7000 structures)

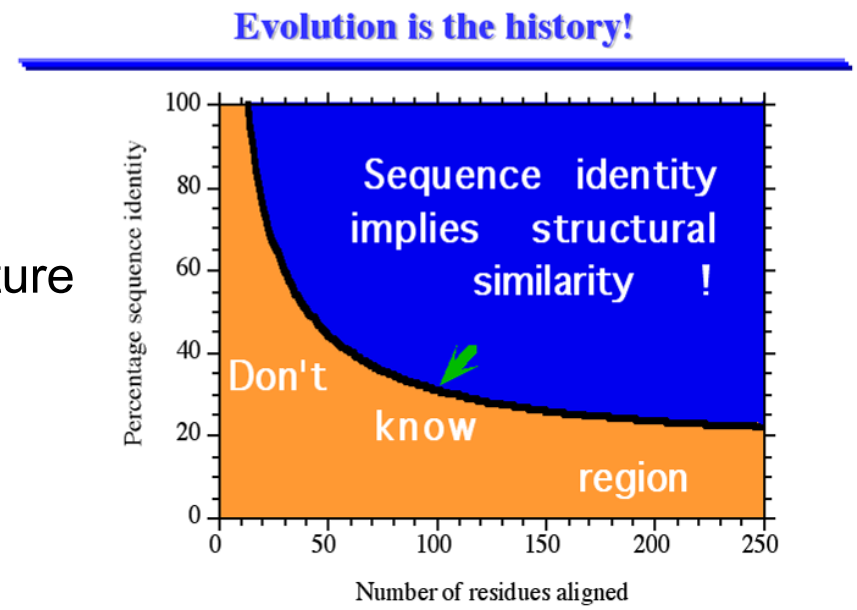
Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

- 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



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

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

- 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?

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

→ Model of protein 2

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

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

Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (target)

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

```
I want to model this!

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

Homology modeling

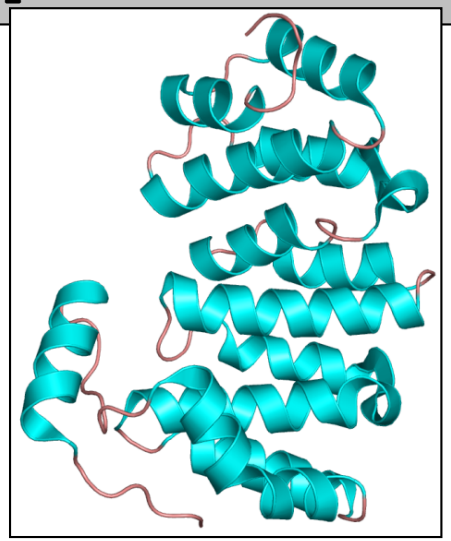
Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

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

Do sequence search in all "PDB sequences"

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

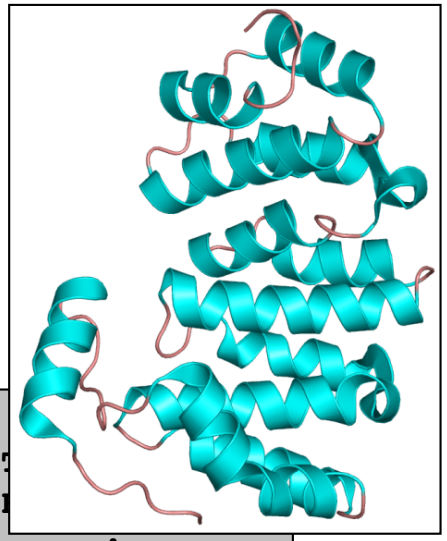


Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

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



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

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

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (tar

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

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   FRIIVRELWDLPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL  120
EF3068    LCQFIEAYYQKTEREYQYVAIDLALQNVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG  122
          :   :.  :: .***:* .*:*:  :  : :.  .:  ::  :  *:***:**:    *
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF  180
EF3068    SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF  171
          . : *:  **:.  :. : .::*:* :*.*: :**  *:* ::**  .*   :::***
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS    237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------     217
          ******* **:*:**:*: * * :::  *:**::**. *:: :
```

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

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Obtaining the correct alignment
ant step!! in

St

1.

t, template
(50-100?) of
ine, T-
a different

2.

e alignment

e template
all indels

3.

4.

heets

5.

6. Valid

```
XP_968170_Tribolium_castaneum   KTLGTGSFGRVMIVQHKPT-KEYYAMKILDKQK
Q4JIV3_Lymnaea_stagnalis        KTLGTGSFGRVMLVQHKGENKAYYAMKILDKQK
```

Where is the correct position of the gap?

```
XP_968170_Tribolium_castaneum   KTLGTGSFGRVMIVQHK-PTKEYYAMKILDKQK
Q4JIV3_Lymnaea_stagnalis        KTLGTGSFGRVMLVQHKGENKAYYAMKILDKQK
```

The MSA gives the answer!!

```
Homolog2_Petromyzon_marinus                 KTLGTGSFGRVMLVKHK-ATDRYFAMKILDKQK
ENSCJAP00000040924_Callithrix_jacchus       KTLGIGSFGRVVLVSHR-ESGSHYAMKILNKEK
P22612_Homo_sapiens                         RTLGMGSFGRVMLVRHQ-ETGGHYAMKILNKQK
XP_968170_Tribolium_castaneum               KTLGTGSFGRVMIVQHK-PTKEYYAMKILDKQK
Q4JIV3_Lymnaea_stagnalis                    KTLGTGSFGRVMLVQHKGENKAYYAMKILDKQK
ENSTRUP00000015108_Takifugu_rubripes        KTLGTGSFGRVMLVKHK-ETNQFYAMKILDKQK
```

```
                                                  DSIVPTFL   120
                                                  DAWRKFFG   122
                                                  *:     *
                                                  LHSSKEFF   180
                                                  DRTTEEFF   171
                                                   ::::***
Bc_AlkD   IQKAIGWVREVAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS     237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------     217
          ******* **:*:**:*: * * :::  *:**::**. *:: :
```

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

Homology modeling

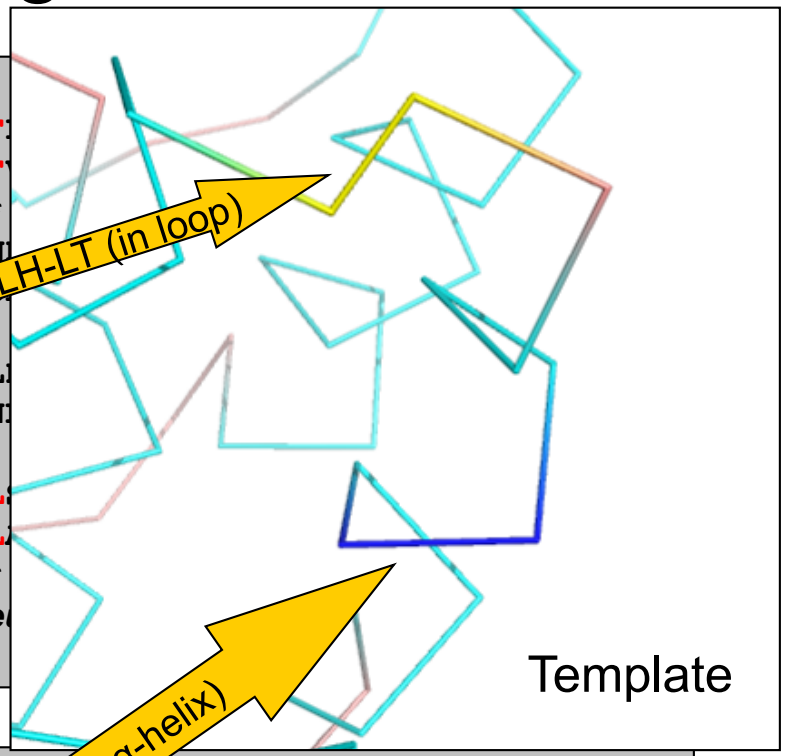
Jon K. Lærdahl,
Structural Bioinformatics

Start with

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

```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHF
EF3068    --------MDTLQFQKNPETAAKMSAYMKHQF
                   : :  :****.*  *: ***::*
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMMQKYKKYIN
EF3068    LCQEIEAYYQKTEREYQYVAIDLALQNV
          :   :.  :: .***:* .*.*:  :  : :.
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAIL
EF3068    SWVALH-LTELPTIFALFYGAENFWNRRVALN
          . : *:    :.: :. : .::*:* :*.*:
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPL
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPL
          ******* **:*:**:*: * * :::  *:**
```

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



indel at PT-IF (in α-helix)

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

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

Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

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!

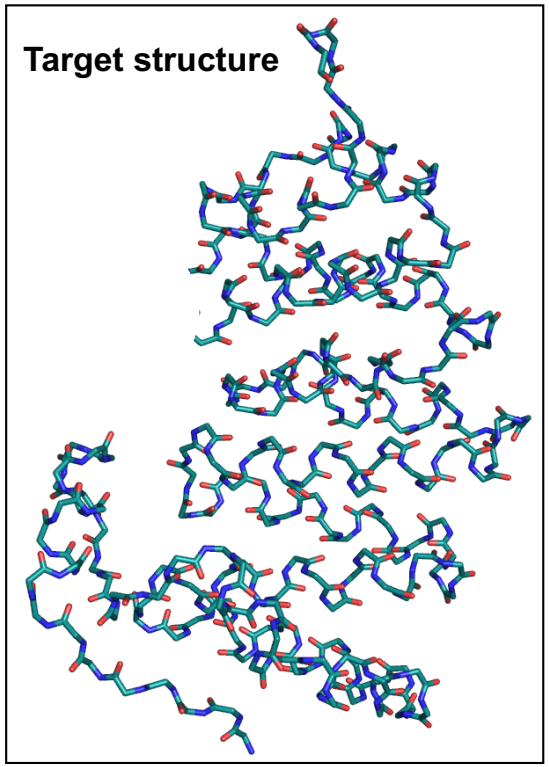
Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

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

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



Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Sta (tar

1. nd

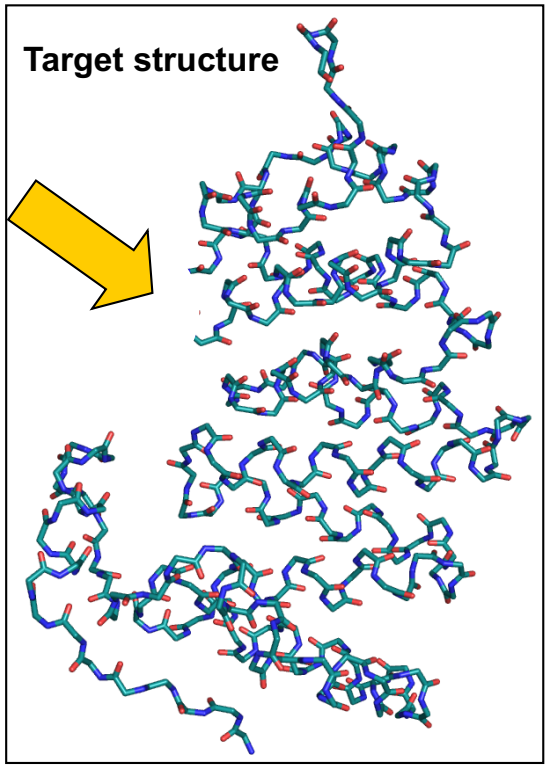
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

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

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

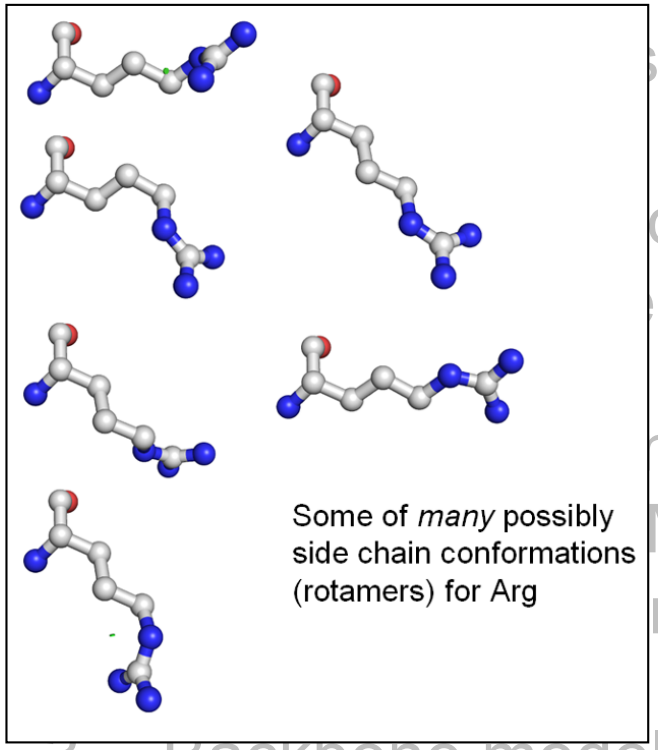
***Ab initio*: Generates random loops and chooses the one with**
- **Lowest energy scores**
- **Ok Ramachandran plot**
- **No clashes**

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



Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics



sequence (tar

on:

in PDB and

nts

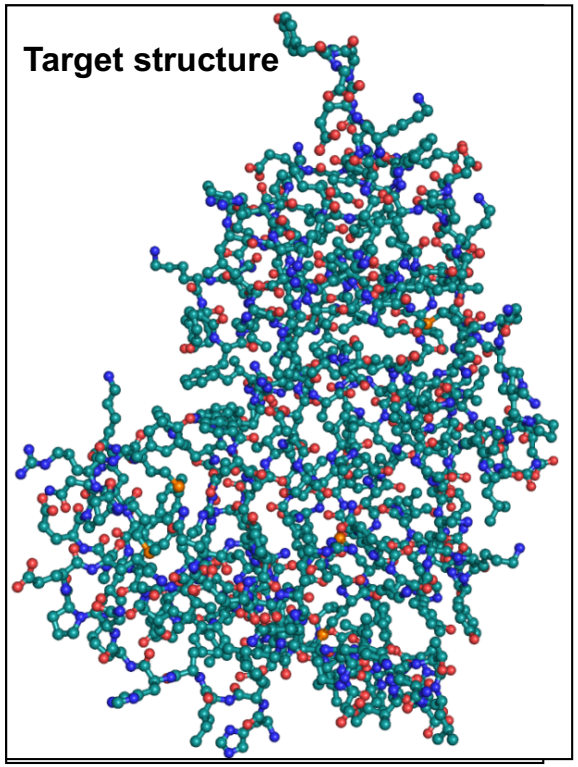
MSA programs

ment of insertions

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!

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**



Homology modeling

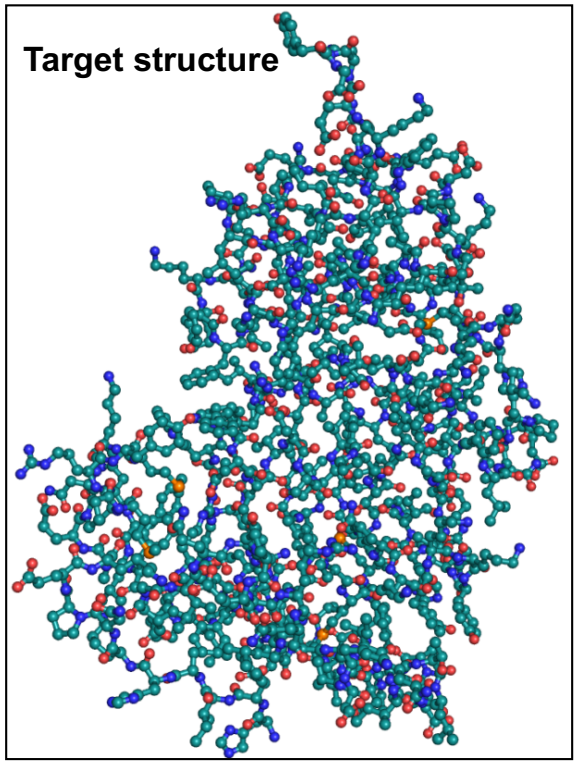
Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

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

Do a few hundred iterations of energy minimization?

- **Will hopefully remove clashes and very unfavorable conformations**
- **Too many iterations will most likely destroy structure**
- **Not always necessary (depends on the program)**



Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Check if model makes sense?

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

e (target)

and align

grams

nsertions

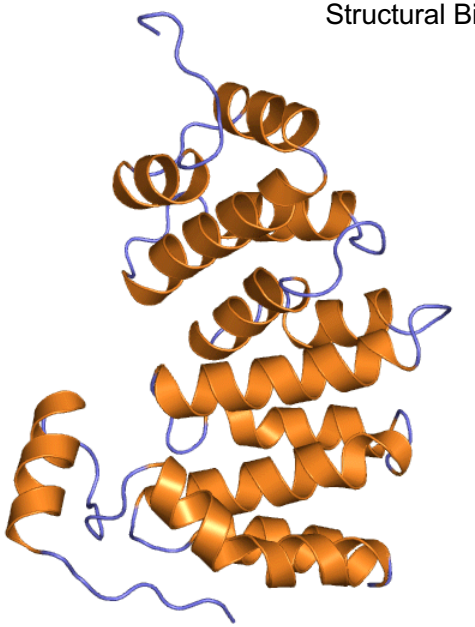
g

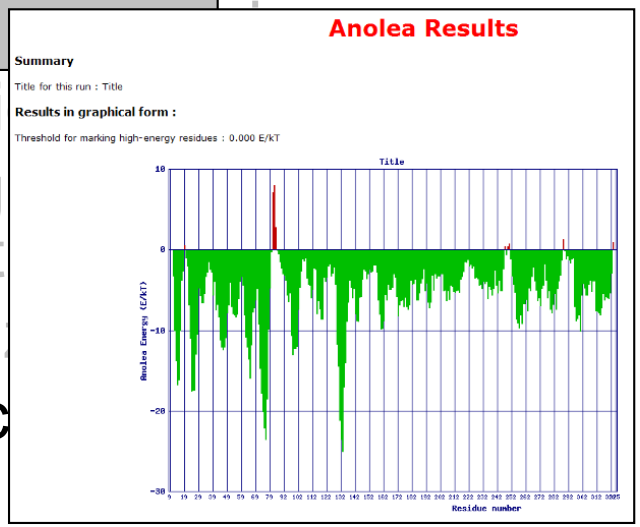
– Rotamer librari

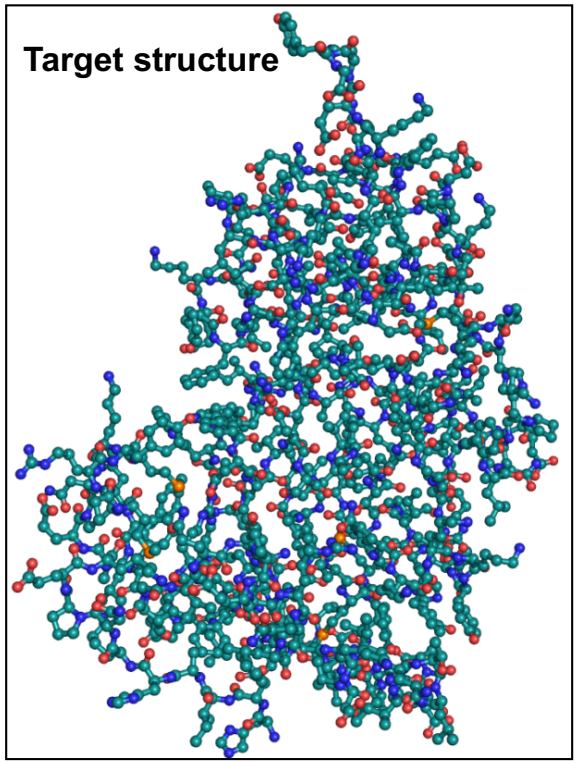
– Loop modeling
or *ab initio* met

5. Refine and optimi

6. Validate and chec

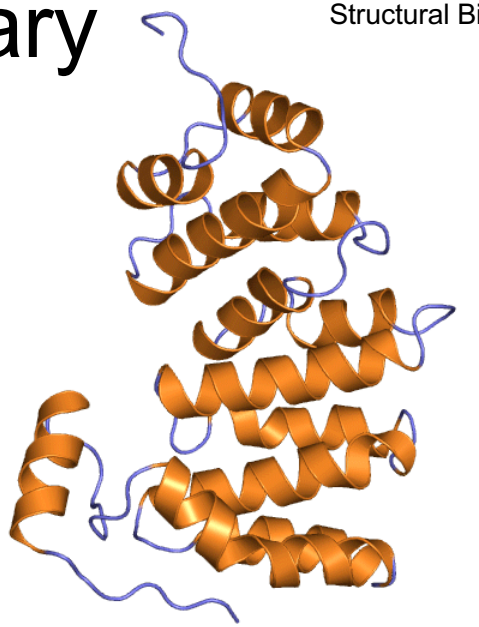






Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling summary



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

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

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

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

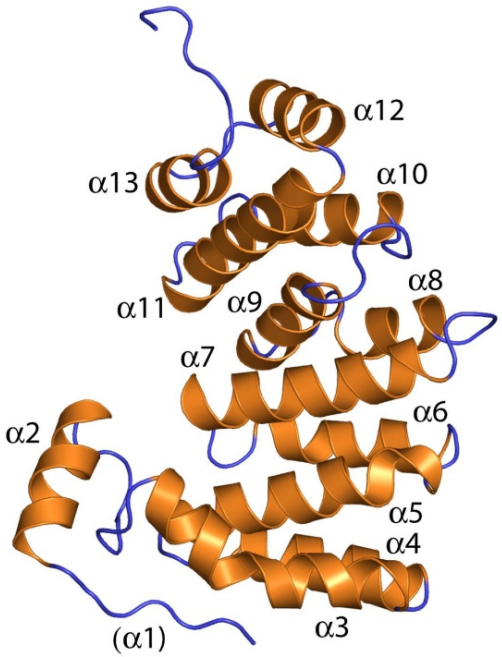
Structural bioinformatics

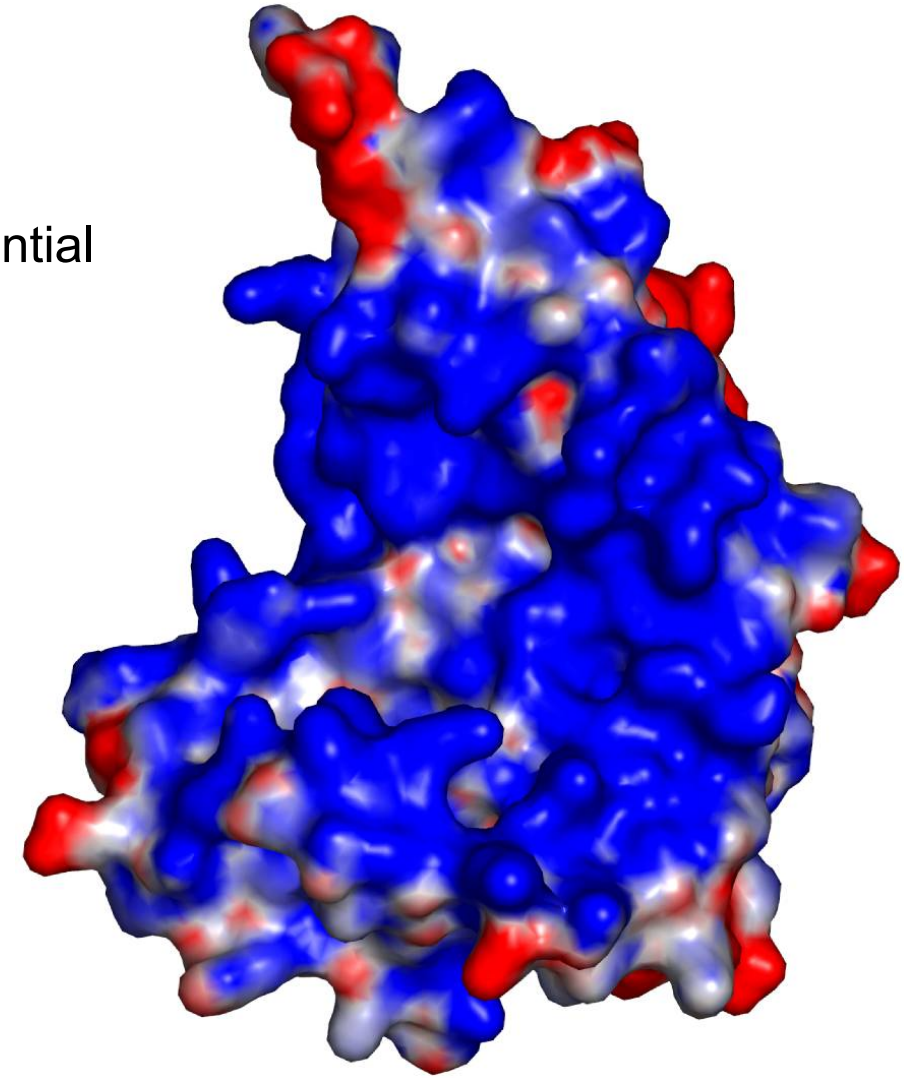
Jon K. Lærdahl,
Structural Bioinformatics

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

Some examples:

B. cereus AlkD electrostatic potential

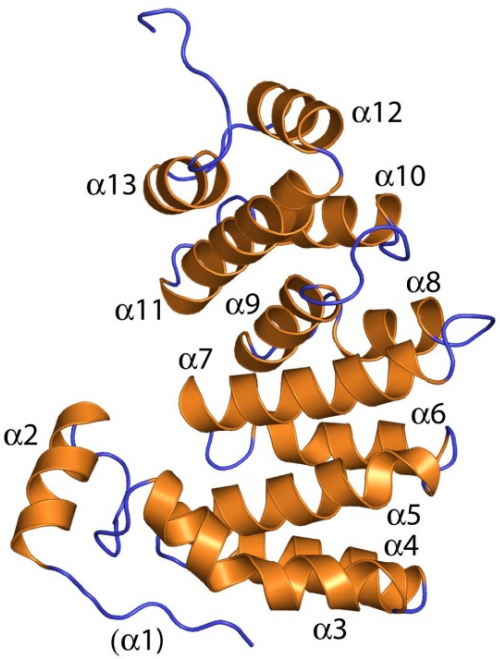


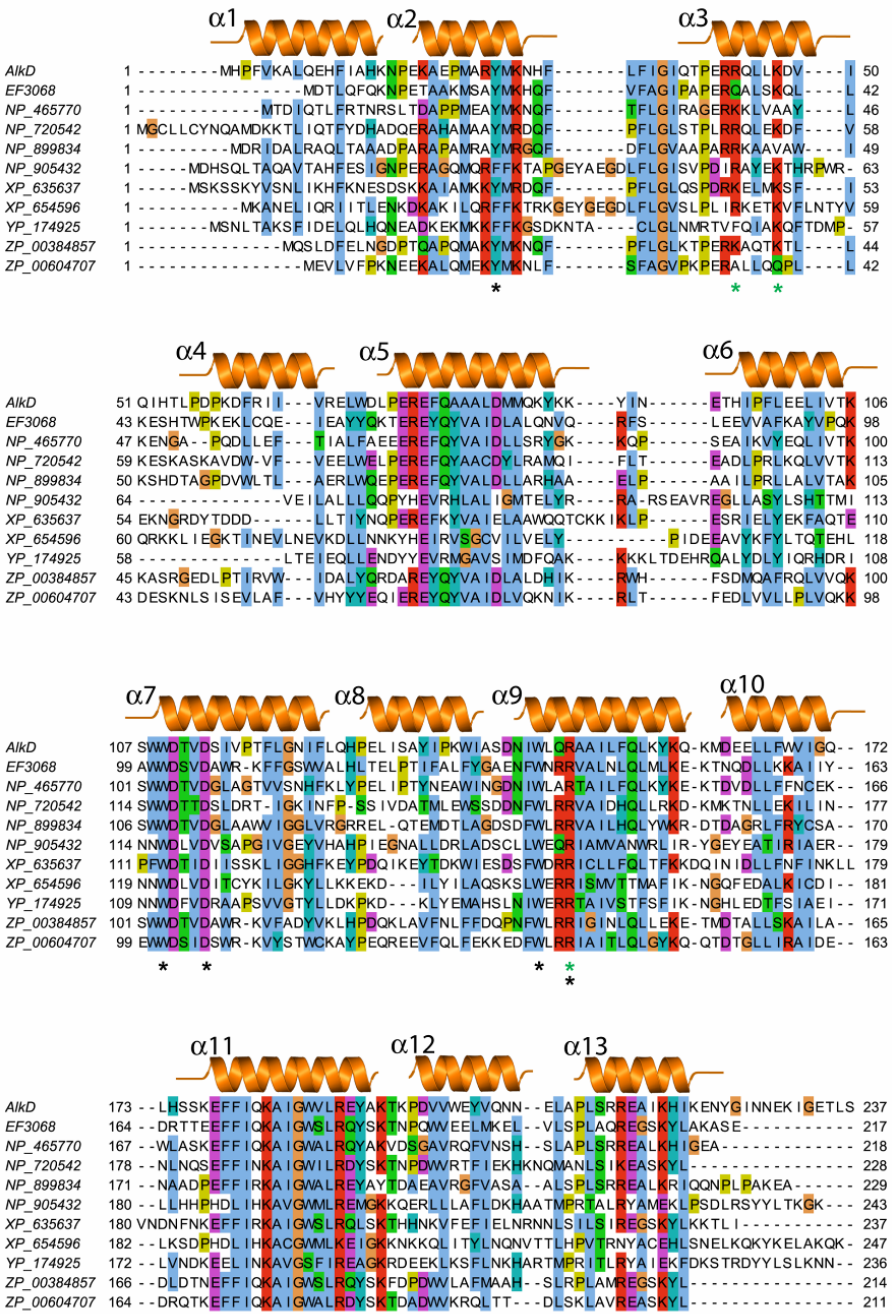


Structural bioinformatics

Jon K. Lærdahl,
Structural Bioinformatics

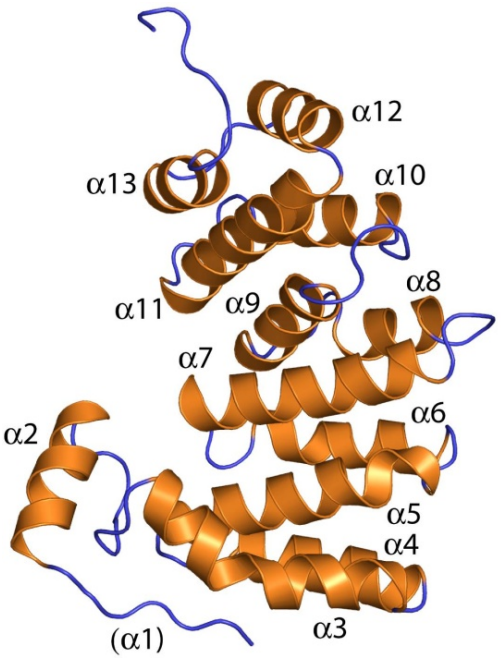
B. cereus AlkD sequence conservation from ConSurf:

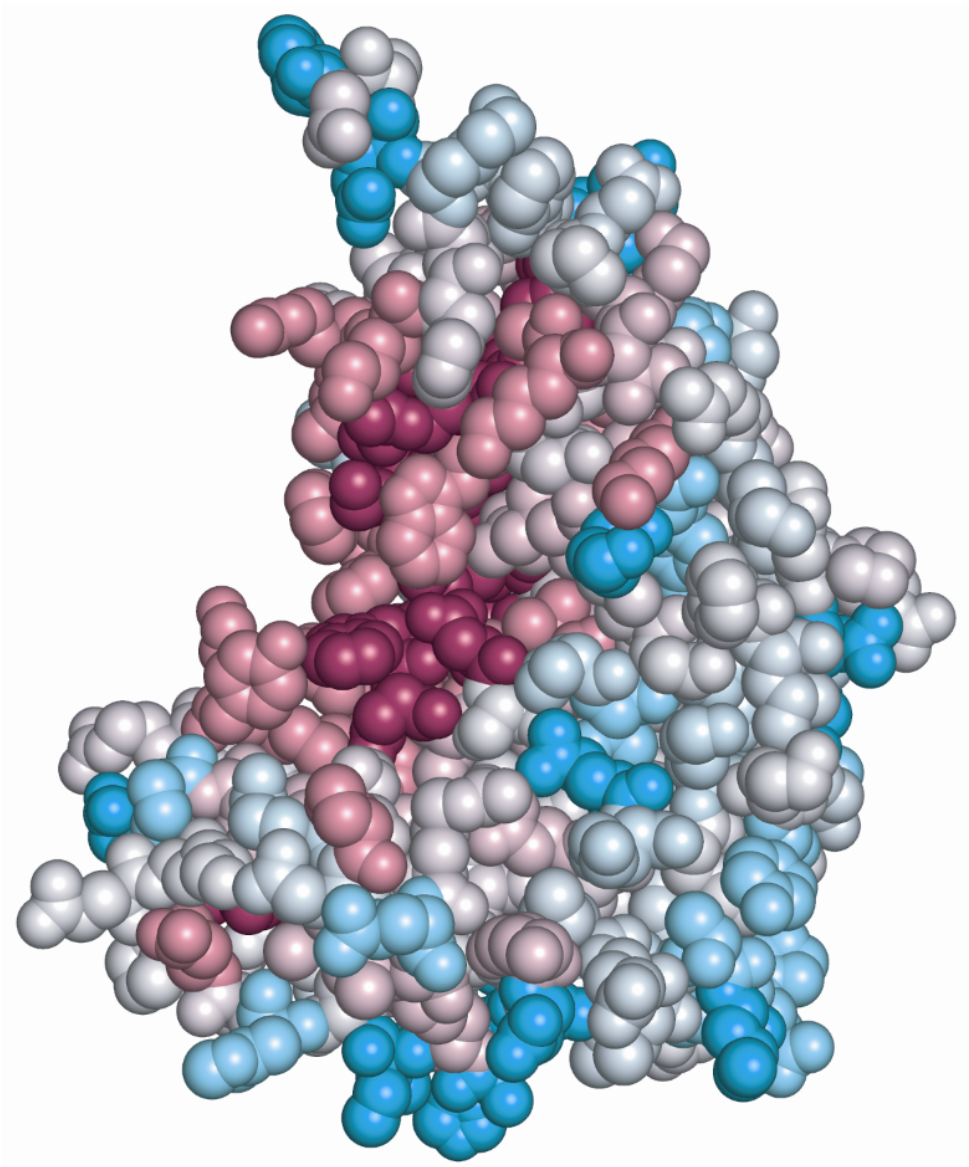




Structural bioinformatics

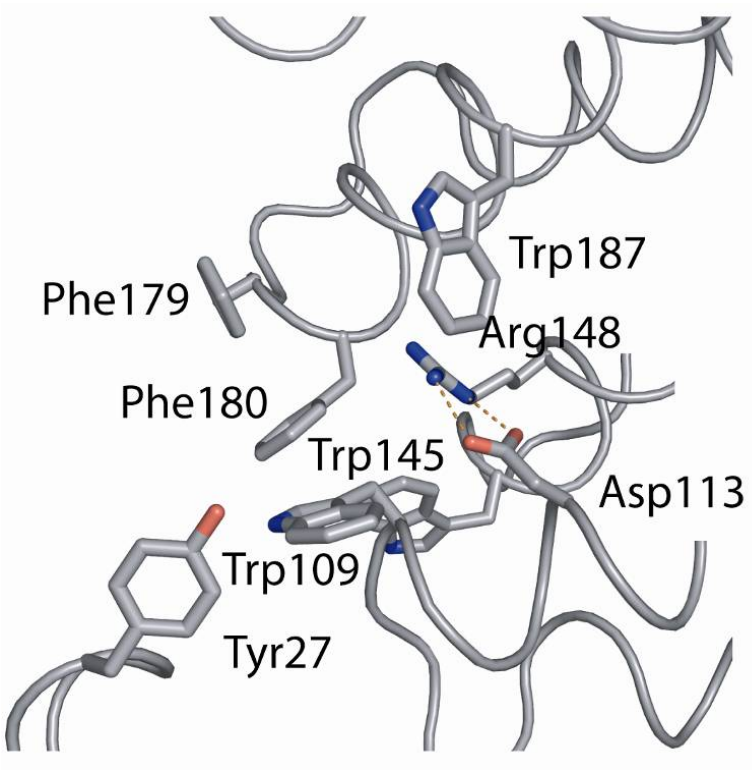
B. cereus sequence conservation from ConSurf:

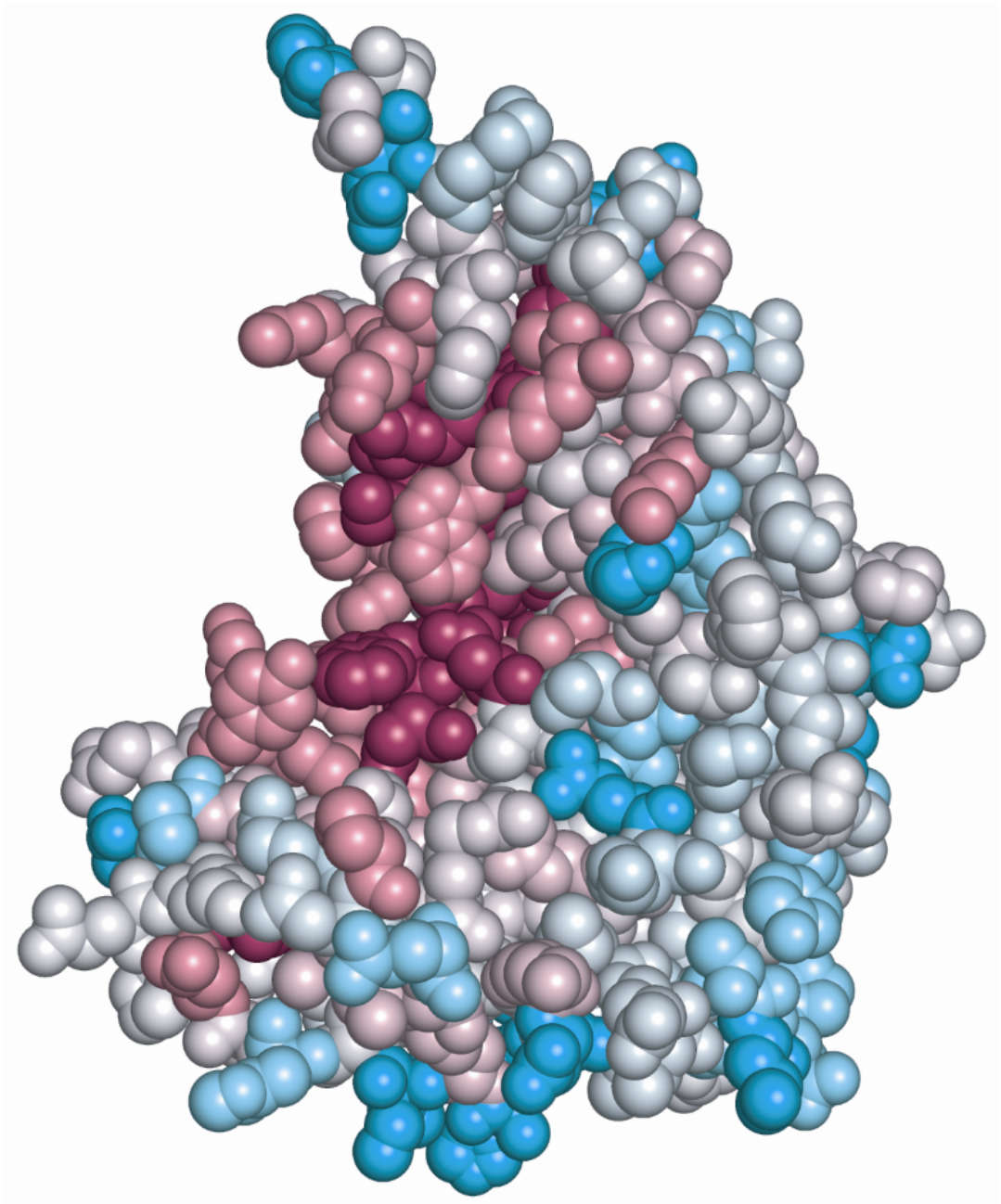




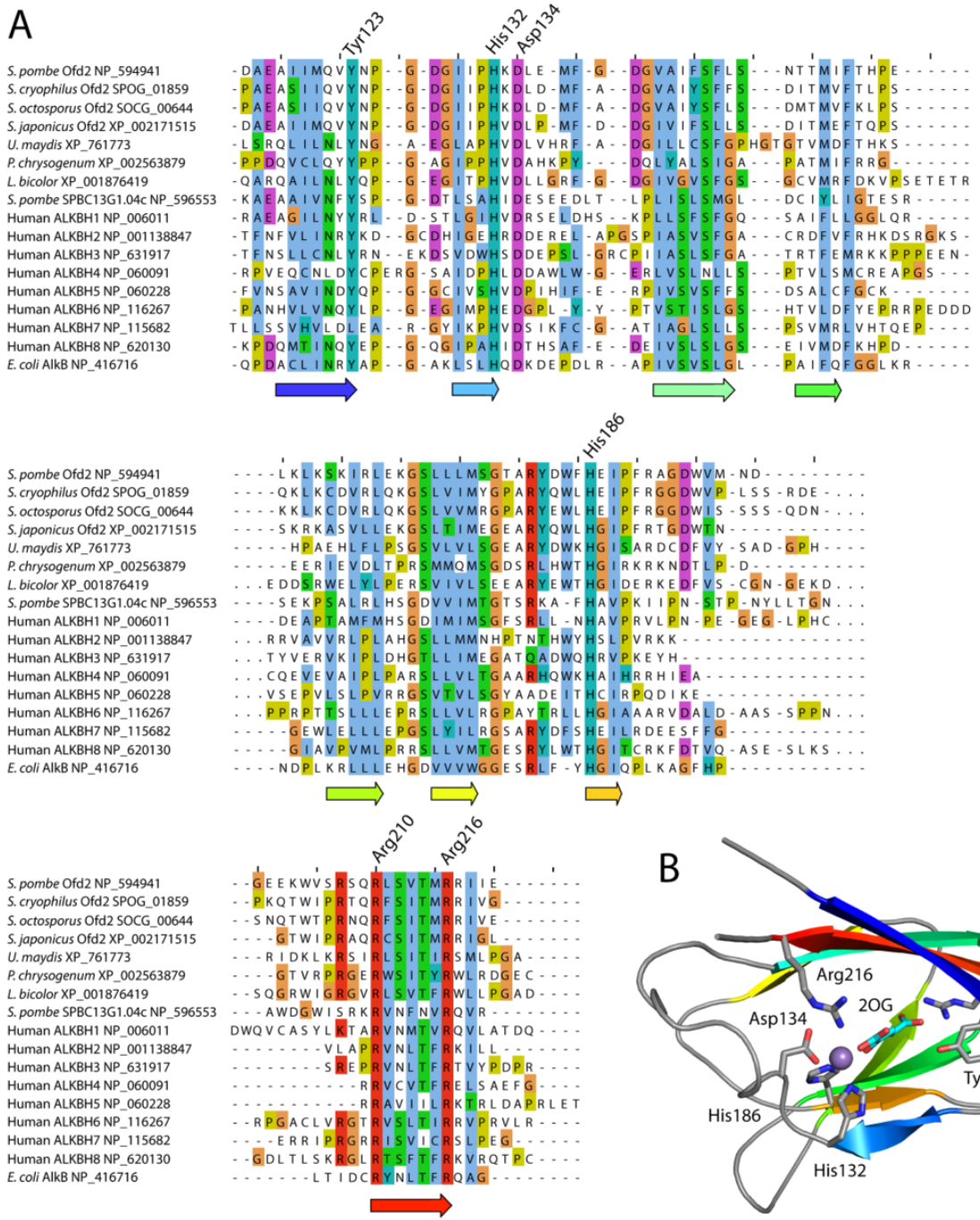
Structural bioinformatics

B. cereus sequence conservation from ConSurf:





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



Use MANY homologs to align two (or a few) homologs!

Korvald *et al.* PLOS One **6**, e25188 (2011)

Use MANY homologs to align some homologs!

Bioinformatics analysis

The sequences of *S. pombe* Ofd2 and 63 homologous sequences from fungal and metazoan species were obtained from the NCBI protein sequence databases [43]. The sequences for two additional *Schizosaccharomyces* species, *S. cryophilus* and *S. octosporus*, were retrieved from the Broad Institute Schizosaccharomyces group database [44] (http://www.broadinstitute.org/annotation/genome/schizosaccharomyces_group). The sequences were aligned with

Korvald *et al.* PLOS One **6**, e25188 (2011)

Expresso [45] and manipulated in Jalview [46]. The main bulk of sequences were subsequently removed in order to give a reliable alignment of Ofd2, AlkB and human and fungal homologs. Structural disorder predictions were performed with the VSL1 algorithm [47] and DISOPRED2 [48]. The structural model of the Ofd2 core domain was derived from an *E. coli* AlkB template from Yu and Hunt [32] (Protein Databank identifier 3I3Q) and the illustration was generated with PyMOL [49].