Protein structure evolution

Jon K. Lærdahl,
Structural Bioinformatics

Last common ancestor
(Long time ago…)

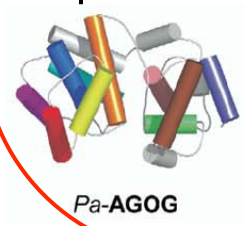
Speciation
Gene duplications

Similar structure
No sequence similarity

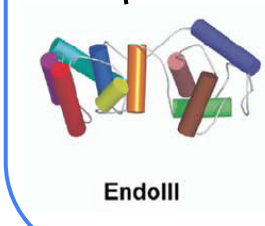
Fairly similar structure
Some sequence similarity

Fairly similar structure
Some sequence similarity

Very similar structure
Significant sequence similarity



Pa-AGOG



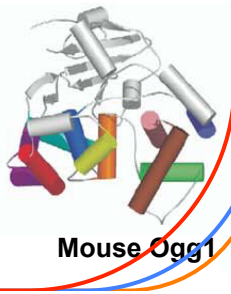
EndoIII

MutY

AlkA

Yeast Ogg1

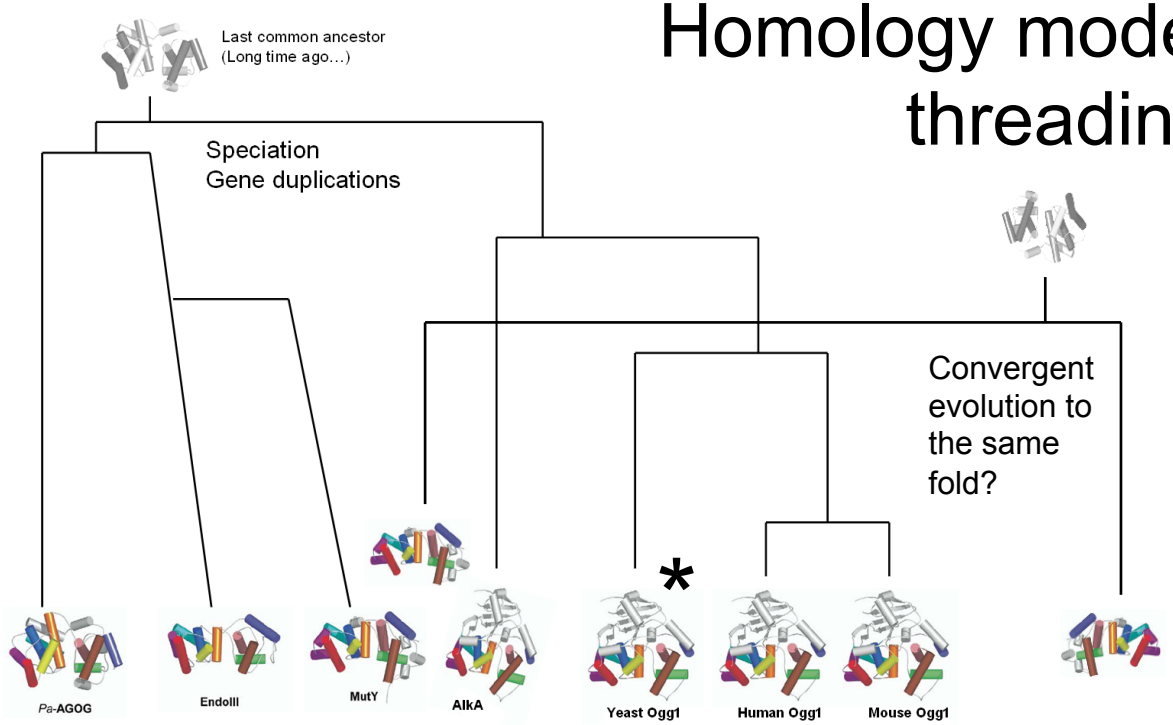
Human Ogg1



Mouse Ogg1

Homology modeling and threading

Jon K. Lærdahl,
Structural Bioinformatics



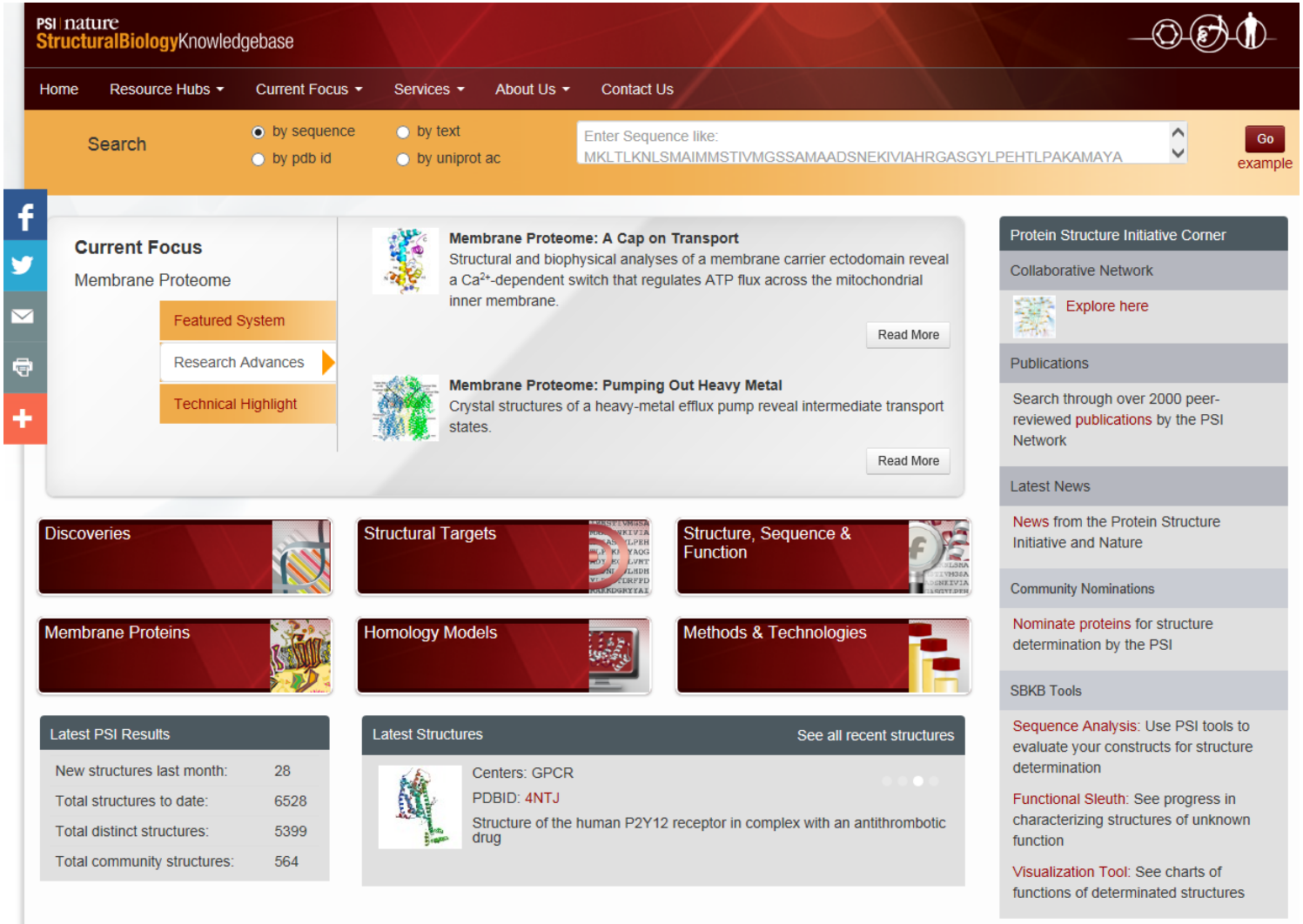
Important goal to have at least one structure in all structural superfamilies!

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

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!

Homology modeling

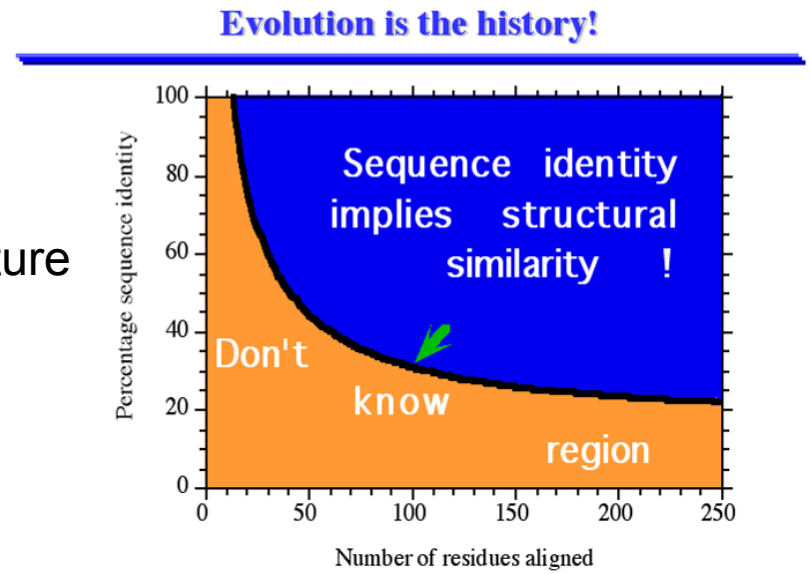
Jon K. Lærdahl,
Structural Bioinformatics

• 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

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



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

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

→ Model of protein 2

• 30% sequence identity necessary (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?

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

```
I want to model this!

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

Homology modeling

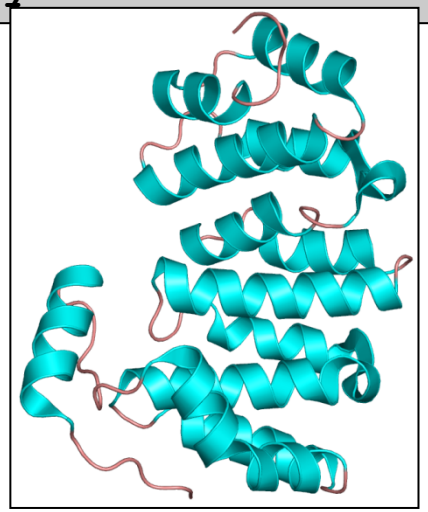
Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

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

Do sequence search in all "PDB sequences"

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

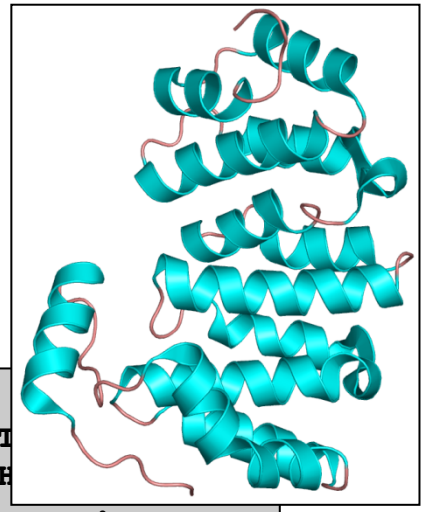


Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

1. Template selection:
   - Find template in PDB and align sequences
2. Correct alignments
3.
4.



```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFLFIGIQTPERRQLLKDVIQIHI
EF3068    --------MDTLQFQKNPETAAKMSAYMKHQFVFAGIPAPERQALSKQLLKESH
                  : :  :****.*  *: ***::*:* ** :***: * *::::       :.
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).

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

Homology modeling

Start with a protein sequence (tar

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

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

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

Use target/template alignment from this MSA

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

Check indels!

```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARY
EF3068    --------MDTLQFQKNPETAAKMSAY
                  : :  :****.*  *: *
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMMQKYKKYINETHIPFLEELIVTKSWWDTVDSIVPTFL  120
EF3068    LCQNIEAYYQKTEREYQYVAIDLALQNVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG  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

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

e alignment

e template
all indels

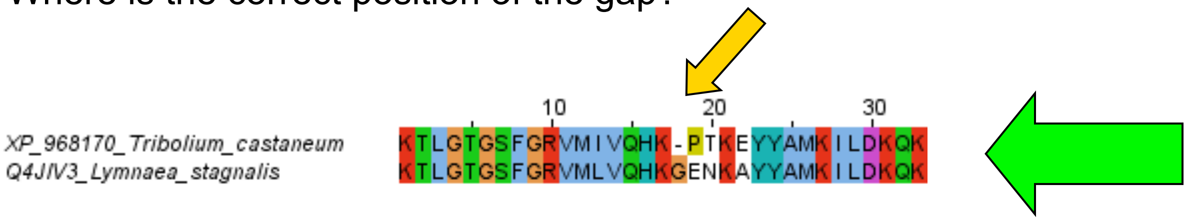
heets

St

1.

Where is the correct position of the gap?

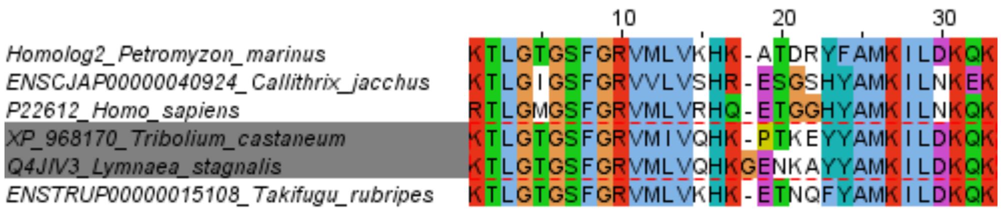
2.



The MSA gives the answer!!

3.

4.




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

5.

6. Valic

```
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS      237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------      217
          ******* **:*:**:*:  *  * :::  *:**::**. *:: :
```

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

Homology modeling

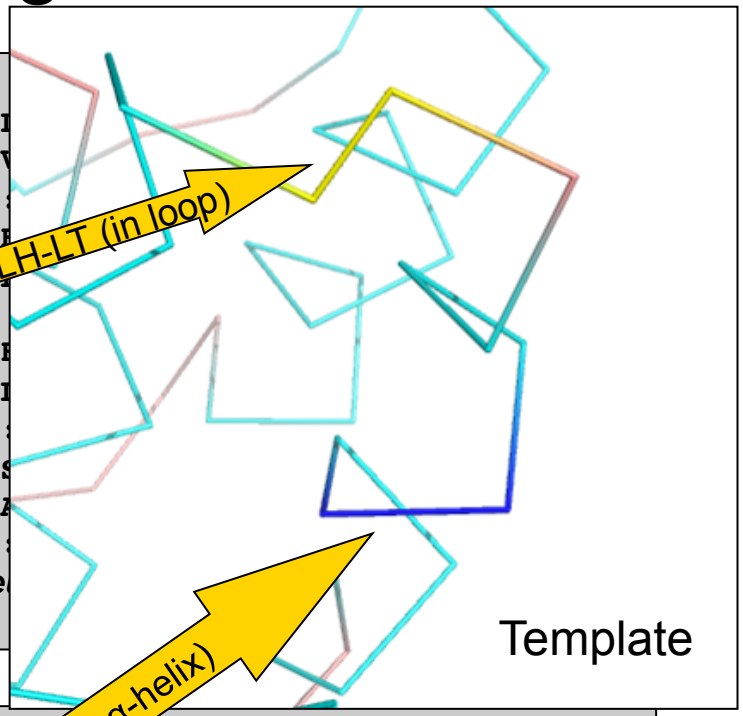
Jon K. Lærdahl,
Structural Bioinformatics

Start with

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

```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFI
EF3068    --------MDTLQFQKNPETAAKMSAYMKHQFV
                    : :  :****.*  *: ***::*:
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMMQKYKKYINE
EF3068    LCQEIEAYYQKTEREYQYVAIDLALQNVQRFSL
          :   :.  :: .***:* .*:*:  :  : :.
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAILF
EF3068    SWVALH-LTELPTIFALFYGAENFWNRRVALNL
          . : *:    :.: :. : .::*:* :*.*: :
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLS
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLA
          ******* **:*:**:*: * * :::  *:**:
```

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




```
Sequence alignment
Bc_AlkD   MHPFVKALQEHFIAHKNPEKAEPMARYMKNHFI  QTPERRQLLKDVIQIHTLPDPKD   60
EF3068    --------MDTLQFQKNPETAAKMSAYMKH  GIPAPERQALSKQLLKESHTWPKEK   52
                    : :  :****.*  *: *     :* ** :***: * *::::       :.
Bc_AlkD   FRIIVRELWDLPEREFQAAALDMM  YINETHIPFLEELIVTKSWWDTVDSIVPTFL  120
EF3068    LCQEIEAYYQKTEREYQYVAID  NVQRFSLEEVVAFKAYVPQKAWWDSVDAWRKFFG  122
          :   :.  :: .***:* .*:*:  :  : :.  .:  ::  :  *:***:**:    *
Bc_AlkD   GNIFLQHPELISAYIPKWIASDNIWLQRAAILFQLKYKQKMDEELLFWVIGQLHSSKEFF  180
EF3068    SWVALHLTELPT-IFALFYGAENFWNRRVALNLQLMLKEKTNQDLLKKAIIYDRTTEEFF  171
          . : *:  **:.  :. : .::*:* :*.*: :**  *:* ::::*  .*   :::****
Bc_AlkD   IQKAIGWVLREYAKTKPDVVWEYVQNNELAPLSRREAIKHIKENYGINNEKIGETLS     237
EF3068    IQKAIGWSLRQYSKTNPQWVEELMKELVLSPLAQREGSKYLAKASE----------     217
          ******* **:*:**:*: * * :::  *:**::**. *:: :
```

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

Jon K. Lærdahl,
Structural Bioinformatics

Homology modeling

Start with a protein sequence (target)

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

The most important step in homology modeling!

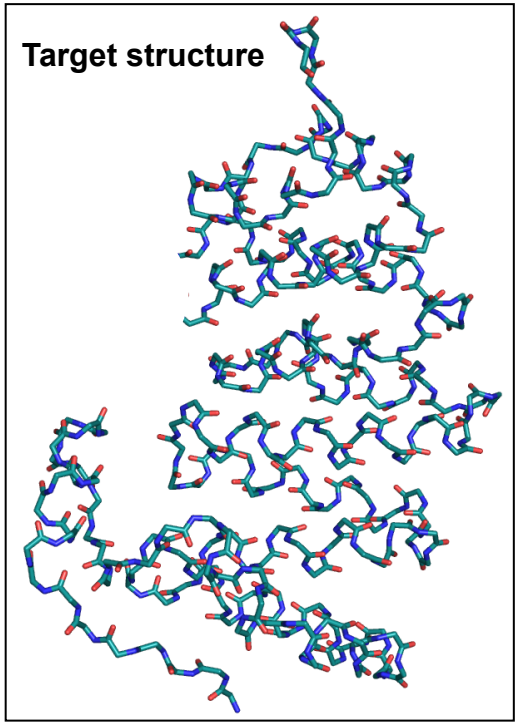
Homology modeling

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

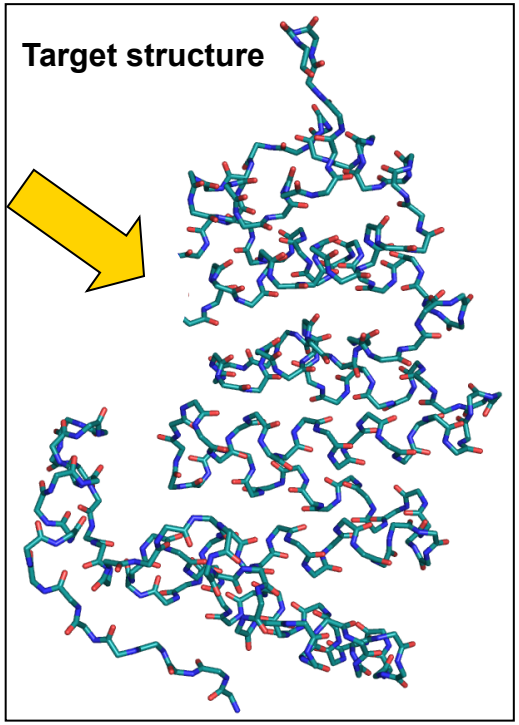
2. Correct alignments
   - Use the best MSA programs
   - Correct placement of insertions and deletions
3. Backbone model building
4. Model loops and side-chains
   - Rotamer libraries
   - Loop modeling using database or *ab initio* method
5. Refine and optimize model
6. Validate and check model quality!

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!!

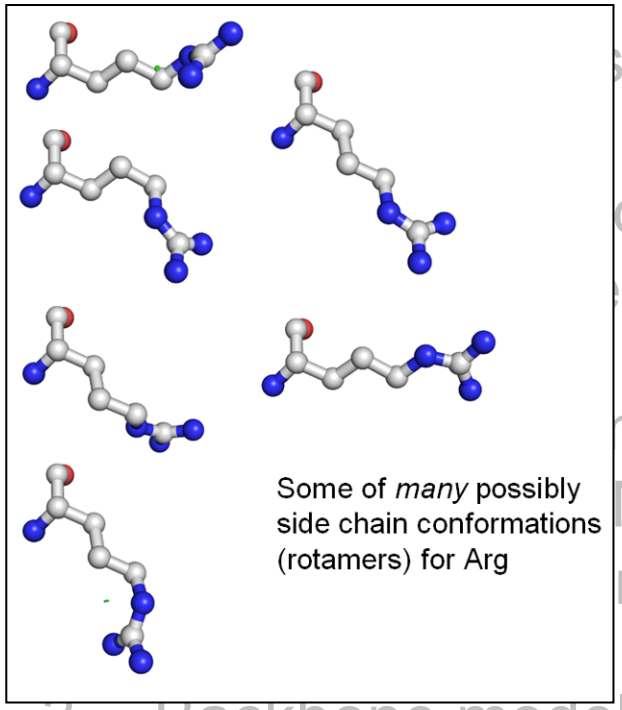
***Ab initio*: Generates random loops and chooses the one with**
- **Lowest energy scores**
- **Ok Ramachandran plot**
- **No clashes**

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



Homology modeling

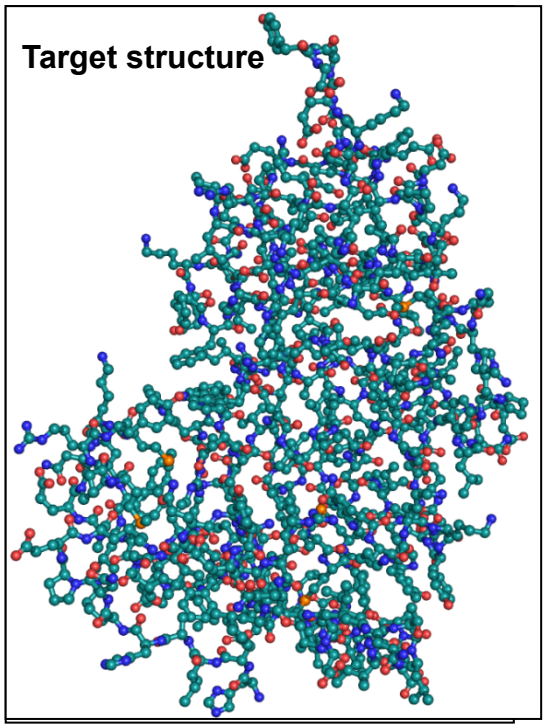
Jon K. Lærdahl,
Structural Bioinformatics



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!



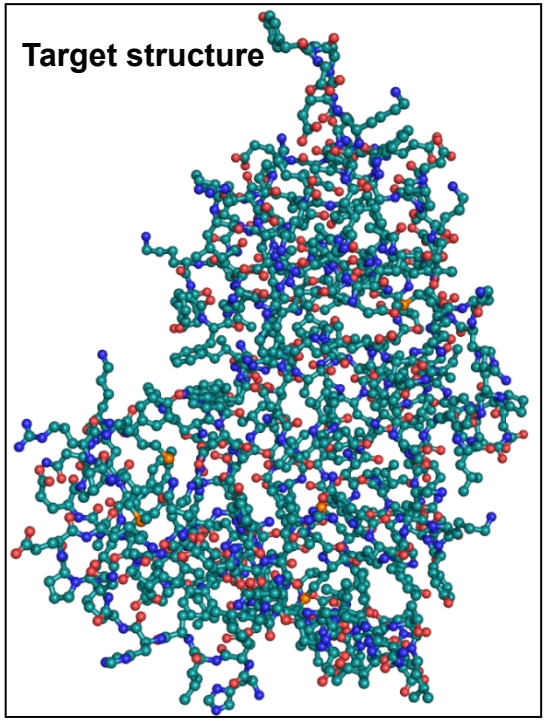
Homology modeling

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

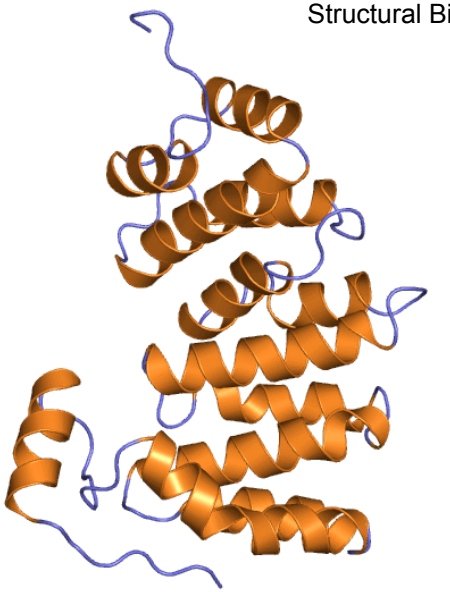
g

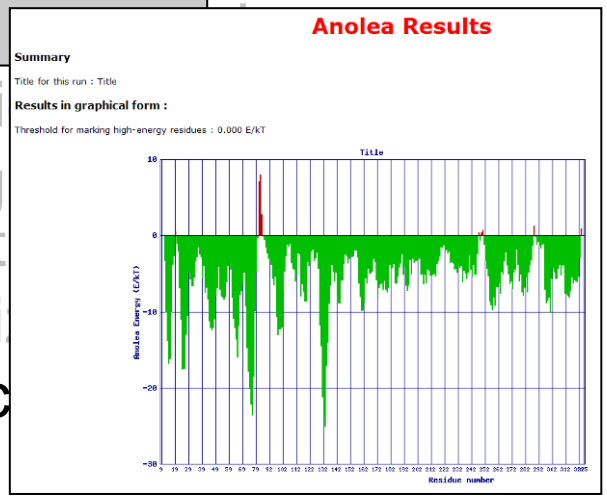
– Rotamer librari

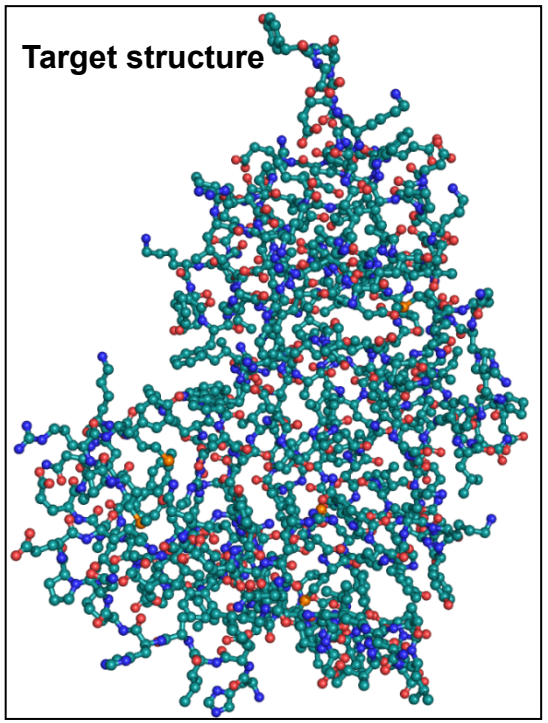
– Loop modeling
or *ab initio* met

5. Refine and optimi

6. Validate and chec

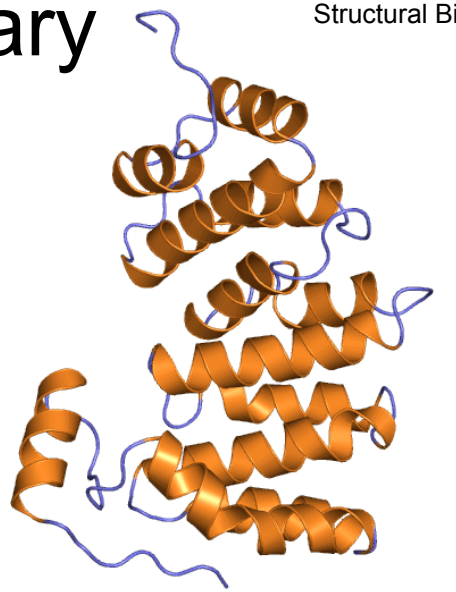






Jon K. Lærdahl,
Structural Bioinformatics

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

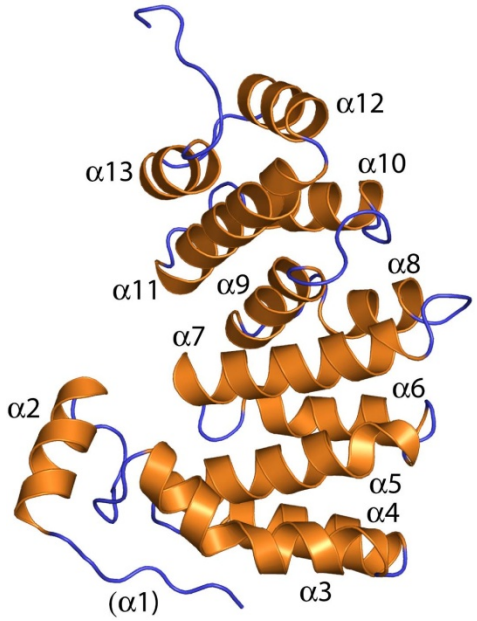
Homology model databases:
- Modbase (automatic modeling with Modeller)
- SWISS-MODEL Repository (automatic modeling with Swiss-Model)

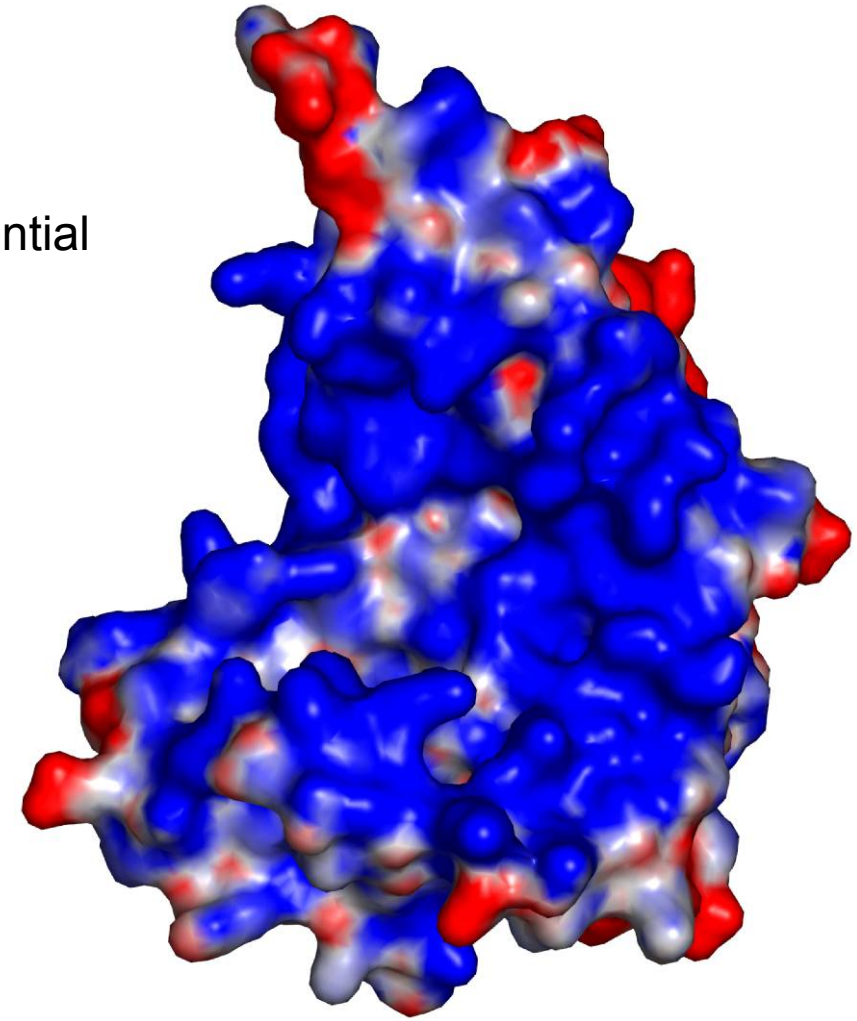
Structural bioinformatics

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: