proteinformatics

Homology modeling

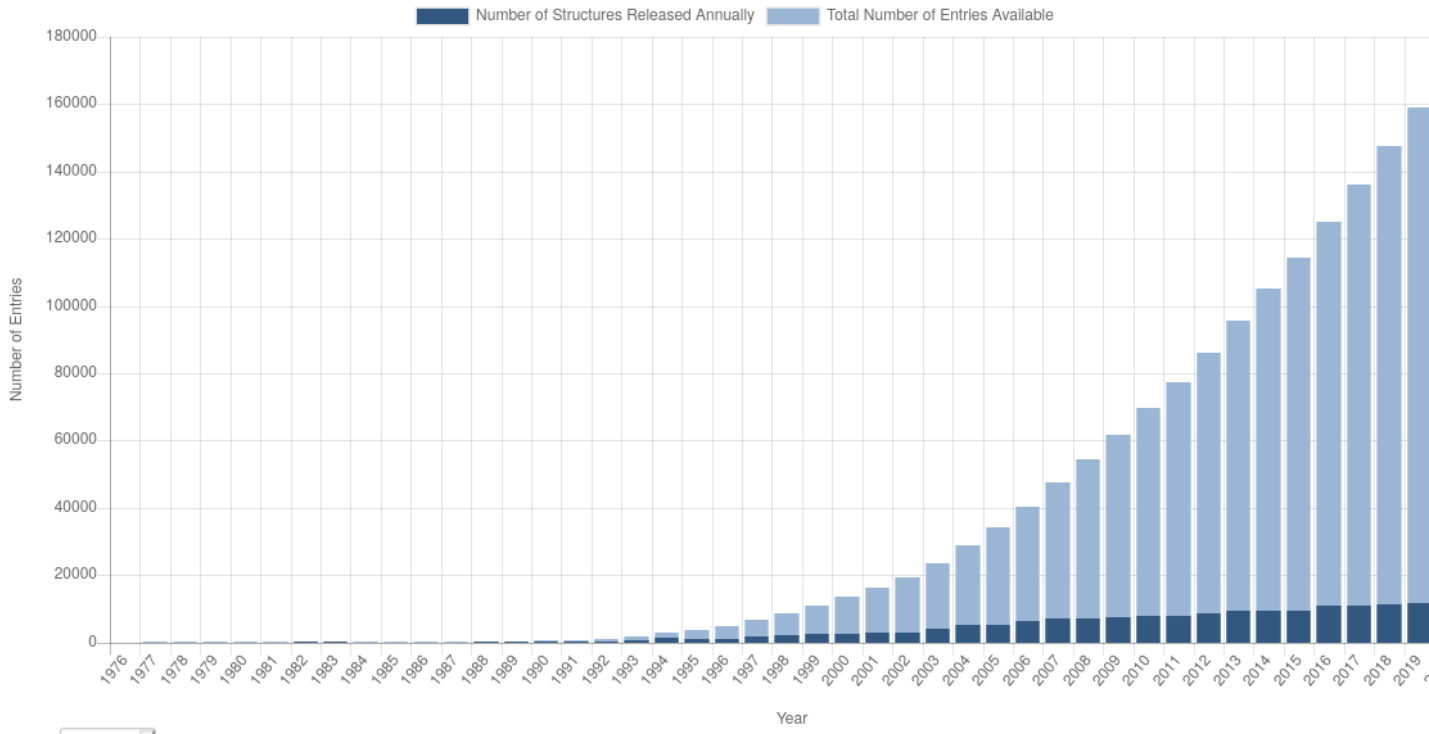
Tutorial

Dr. René Staritzbichler

2020

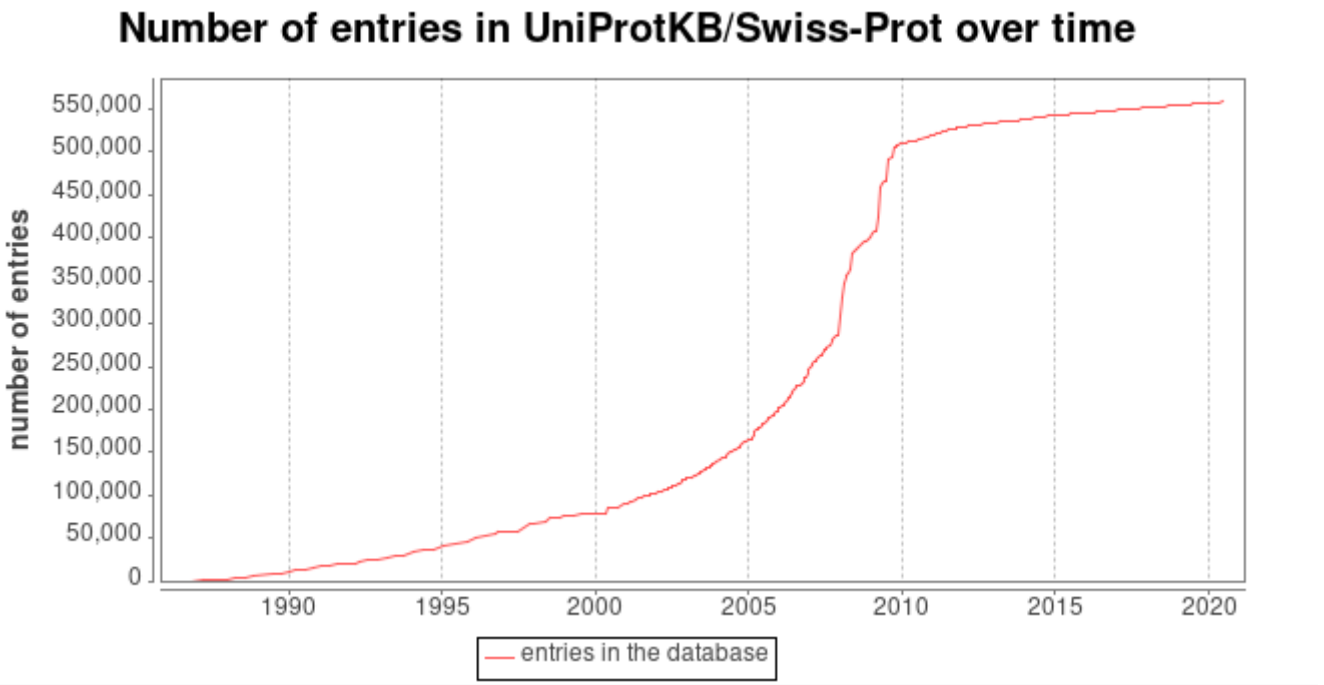
Known structures

- Protein Data Bank (PDB): 167k entries
  - X-ray
  - Cryo EM
  - NMR



Known sequences

- Anotated swissprot: 562k



Homology as bridge

- For many sequences it is possible to find a very similar (homologue) protein in the PDB
- Proteins with similar sequences are expected to have similar folds and similar function
- These can be used as template for comparative modeling

Homology modeling

1) Find template
2) Perform Alignment of query with template
3) Thread aligned positions onto template structure
4) Energy minimization of initial model

1) Find templates

- Search NCBI BLAST
- Select protein version
- Copy & paste sequence of 3sn6.pdb chain R
- Select PDB as database (compare size to ‘nr’)
- Submit

1) Results of BLAST

- BLAST returns list ranked by score
- Check coverage / percentage identical
- A low e-value indicates a high confidence (values below 1e-20 are considered reliable)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Chain A, Beta-2 Adrenergic Receptor [Homo sapiens]	755	755	88%	0.0	99.45%	2R4R_A
☑	Chain A, Beta-2 adrenergic receptor [Homo sapiens]	750	750	88%	0.0	99.18%	3KJ6_A
☑	Chain A, Beta-2 Adrenergic Receptor [Homo sapiens]	706	706	82%	0.0	99.71%	2R4S_A
☑	Chain R, Endolysin,Beta-2 adrenergic receptor [Escherichia virus T4]	695	695	81%	0.0	99.11%	3SN6_R

1) Inspect template

- Scroll down list until you see Lysozyme entries
- Lysozyme was fused to 3sn6 for crystallization
- Here, it is counterproductive
- Notice both E value and Identity!
- Download sequence of first Lysozyme
- Paste its sequence with the one of 3sn6 to AlignMe (fast mode)
- Select subrange (1 row is 60 AA) not containing lysozyme
- Repeat BLAST using that range (input parameter)
- No Lysozyme entries should appear

1) Select template

- Generally one of the top hits should be selected as template
- However, we want to use an active conformation of the receptor
- There are not many structures of active receptors in the PDB yet
- This would require extensive (!) research to exclude all inactive structures
- Instead, we will simply use 3dqb as template

2) AlignMe

- http://www.bioinfo.mpg.de/AlignMe/
- Sequence to sequence alignment
- Enter query and template sequence (3sn6 and 3dqb)
- Compare fast mode with PS (may take some time, come back later)
- Add to PS: ‘show detailed alignment parameter options’
  - Select scale: HWvH, triangular window, size 13, weight 0 (for comparing plots with fast mode)

3) Modeller

- https://salilab.org/modeller
- The classic tool, implemented as python library
- Preparation:
  i. Convert alignment
  ii. Adjust script

3.i) Convert alignment

- Open editor and translate alignment of AlignMe into this format:
- Measure length:

```
$ echo –n SEQUENCE | sed 's/–//g' | wc
```

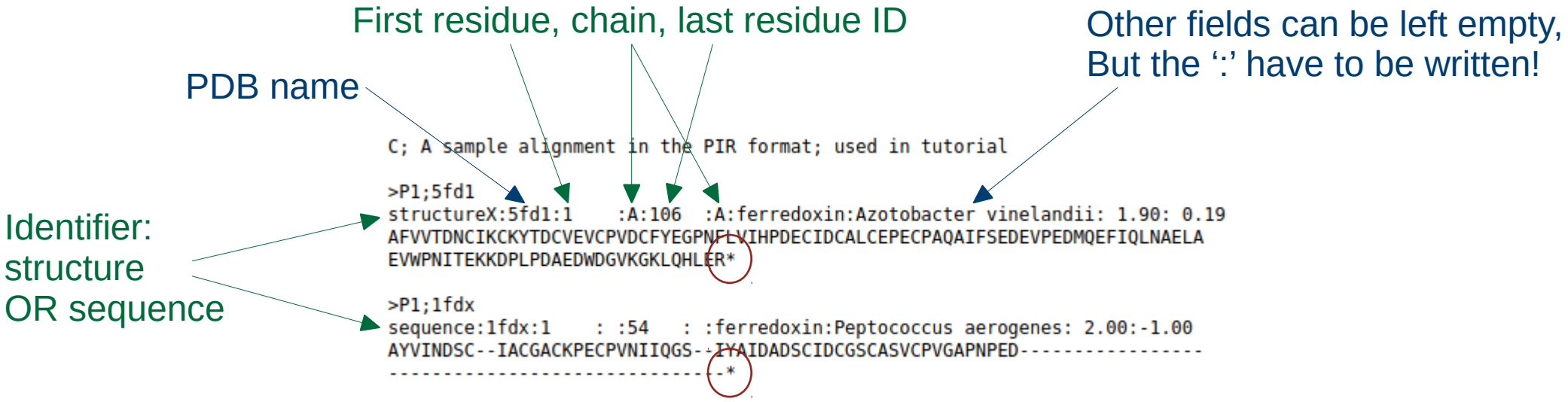
“echo” returns the string, ‘-n’ without newline
“sed ‘s/FROM/TO/g’ ” replaces all occurrences of FROM with TO
“wc” word count

```
C; A sample alignment in the PIR format; used in tutorial

>P1;5fd1
structureX:5fd1:1    :A:106  :A:ferredoxin:Azotobacter vinelandii: 1.90: 0.19
AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAIFSEDEVPEDMQEFIQLNAELA
EVWPNITEKKDPLPDAEDWDGVKGKLQHLER*

>P1;1fdx
sequence:1fdx:1    : :54   : :ferredoxin:Peptococcus aerogenes: 2.00:-1.00
AYVINDSC--IACGACKPECPVNIIQGS--IYAIDADSCIDCGSCASVCPVGAPNPED----------------
--------------------------------*
```

3.i) Convert alignment



3.ii) Adjust Modeller script

```
# Comparative modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']

a = automodel(env,
              alnfile  = 'alignment.ali',     # alignment filename
              knowns   = '5fd1',              # codes of the templates
              sequence = '1fdx')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual comparative modeling
```

3.ii) Adjust Modeller script

Directory where PDB file is located

File names

Notice plural

Number of models

```
# Comparative modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']

a = automodel(env,
              alnfile  = 'alignment.ali',     # alignment filename
              knowns   = '5fd1',              # codes of the templates
              sequence = '1fdx')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual comparative modeling
```

4) Refinement

- Modeller returns minimized models
- However, these can be further optimized
  - Rosetta
  - MD

(both are subject to individual sessions)

Quality

- Most crucial step is the alignment
- Quality of model can be improved by improving alignment
- Worth spending some weeks with alignment !

Common applications

- Fixing structure from PDB for MD
- Many proteins (e.g. GPCRs) have multiple states, but not structures for all states
- Some proteins have no structural data available

(sorted by complexity)

Alternative approaches

- https://swissmodel.org
- Rosetta comparative modeling (next week)