Bioinformatik

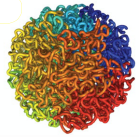
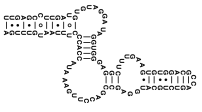
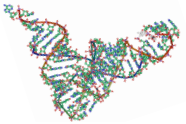
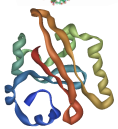
Wintersemester 2019 / 2020, Uni Potsdam

Stefanie Hartmann

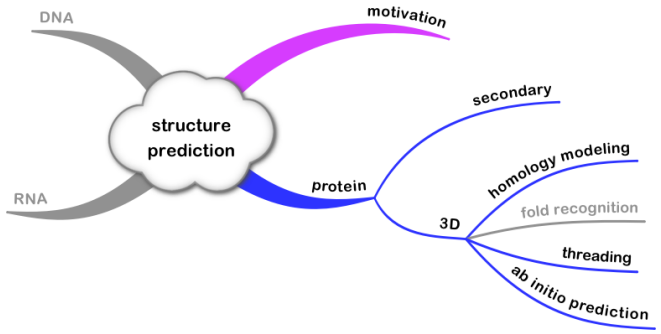
Protein structure prediction

Dec 20, 2019

Sequence, structure, function

	primary	secondary	tertiary
DNA	ATCTCGATCCTG		
RNA	AUCUCGAUCCUG		
protein	TIVFPAKFRRVL		

Overview



Determination of sequence & structure

DNA sequence

- Sanger sequencing, NGS technologies
- genomes, protein-coding genes, non-coding RNAs

protein sequence

- mass spectrometry, Edman degradation, …
- computational translation of DNA sequence

protein 3D structure

- mostly X-ray crystallography, NMR spectroscopy, cryo-electron microscopy

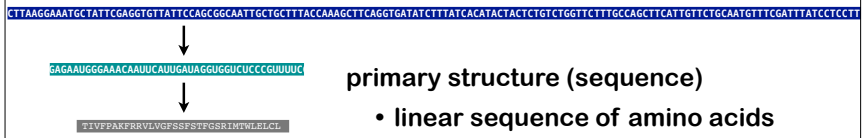
Determined sequences & structures

NCBI	SRA	3.6	Petabytes
	GenBank	210,000,000	sequences
	WGS	722,000,000	sequences
UniProt	SwissProt	560,000	sequences
	TrEMBL	137,000,000	sequences
PDB	*Homo sapiens*	50,000	structures
	Escherichia coli	8,200	structures
	Mus musculus	7,000	structures
	S. cerevisiae	4,000	structures
	...		
	Total	140,000	structures

Prediction of structure

- many more sequences than structures available
  - experimental determination of structure: difficult, time-consuming, expensive
- early bioinformatics endeavors
  - infer a structure based on sequence information
    - intrinsic information
    - comparative approaches
- the only practical solution?

Protein: sequence, structure, function



primary structure (sequence)

- linear sequence of amino acids

secondary structure

- hydrogen bonds between backbone NH and CO groups
- α-helix, β-strand (sheet), loop (coil, turn)

tertiary structure

- folded secondary structure elements, compact & organized

quaternary structure

Protein: sequence, structure, function

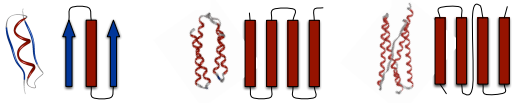
domain

- functional / structural element (e.g., Calcium-binding domain)

super-secondary structure

- specific geometric arrangements & combinations of α-helices and β- sheets

fold



- general arrangement and connectivity of secondary structure elements in a protein

example: CATH

- protein structures from PDB: split up into domains

C	class	architectures with similar content of secondary structure (e.g., mainly alpha or beta, mixed, etc)
A	architecture	topologies that share a roughly similar spatial arrangement of secondary structures
H	homologous superfamily	domains that share a clear common ancestor
T	topology	homologous superfamilies that share the same fold; no clear evidence for evolutionary relationship

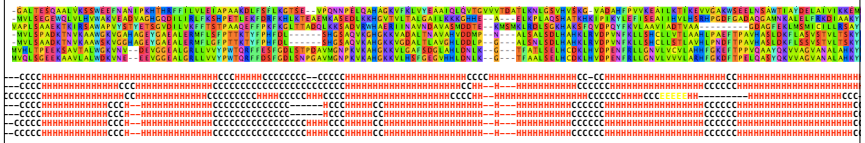
Protein secondary structure

tertiary (or quaternay) structure

- function!

secondary structure

- simpler description of structure
- important for classifying proteins
- useful for comparing proteins / protein structure
- important for predicting 3D structures



Protein secondary structure

1. assign secondary structure from a given 3D structure
   - use 3D coordinates for automatic assignment
   - most commonly used approaches are based on hydrogen bonding patterns: DSSP, STRIDE

→ atomic-level resolution coordinates →

```
LINEVLANGMQPYVTLFHWDVPQALEDEYRGFLGRNIVDDFR
STTHHHHHHHHHHHHHHHHHTTTEEEEEEHHHHHHHHHTSSTT
```

2. predict secondary structure from a primary sequence
   - different approaches & programs exist!

```
LINEVLANGMQPYVTLFHWDVPQALEDEYRGFLGRNIVDDFR  →
LINEVLANGMQPYVTLFHWDVPQALEDEYRGFLGRNIVDDFR
STTHHHHHHHHHHHHHHHHHTTTEEEEEEHHHHHHHHHTSSTT
```


Protein secondary structure prediction

- analyze & learn from known protein structures
- more data: better prediction results

First generation methods

- statistical methods
- single-residue propensities

Second generation methods

- statistical methods
- consider adjacent residues

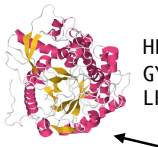
Third generation methods

- use evolutionary information
- use neural networks, HMMs

```
helix
  ↑
  R

          helix
            ↑
CWNHHKAI R HIDTVRAY

          helix
            ↑
CWNHHKAI R HIDTVRAY

CWNHHKAV R HIDTVRAY
CWNHHEEI R HIDTVRAY
AWNHHEQI K HIDTVRAY
AIRHHPQI K RVDTVRAY
```

Prediction accuracy

- unknown for proteins without a solved 3D structure!
- evaluate accuracy by using protein sequences for which a 3D structure is available



HLCYLQAAIKDGVRVK
GYFAWVRFGIVDYDNG
LKRRSKFSAHWLKNFL

1. predict

HHHHHHHHHHCCCCEE
EECCCCCEEEEECCCC
CCCCCHHHHHHHHHHH

1. compare

Prediction accuracy

```
      HLCYLQAAIKDGVRVKGYFAWVRFGIVDYDNGLKRRSKFSAHWLKNFL
      HHHHHHHHHHCCCCEEEECCCCCEEEEECCCCCCCCCHHHHHHHHHH
pred1 HHHHHHHHCCCCCCEEEEEECCCCEEEEEECCCCCCCCCHHHHHHHHH  9/50
pred2 HHHHHHHHHHCCCCHHHHCCCCCHHHHHCCCCCCCCCHHHHHHHHHH  9/50
```

measures of accuracy

- overall accuracy (Q_3)
- accuracy for each state
- overlap between observed vs. predicted structures (SOV)

current best predictors: ~82% Q3 accuracy

3D protein structure

- helps to understand the protein's function
- helps to design site-directed mutations & drugs
- can explain binding specificity, antigenic property
- ...

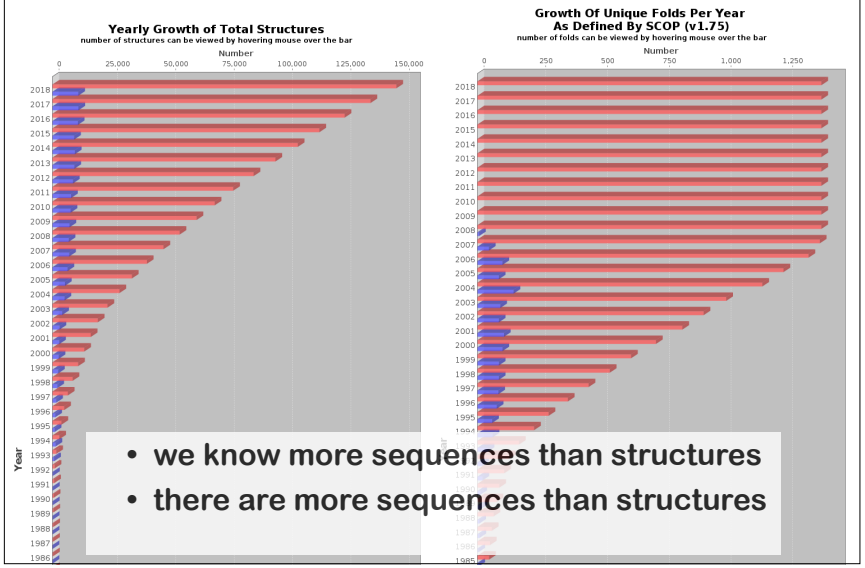
experimental determination of 3D structure

- expensive! time consuming!
- doesn't always work

computational prediction of 3D structure

- can provide useful results
- doesn't always work

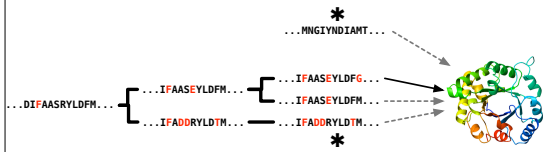
PDB content growth



- we know more sequences than structures
- there are more sequences than structures

Approaches to structure prediction

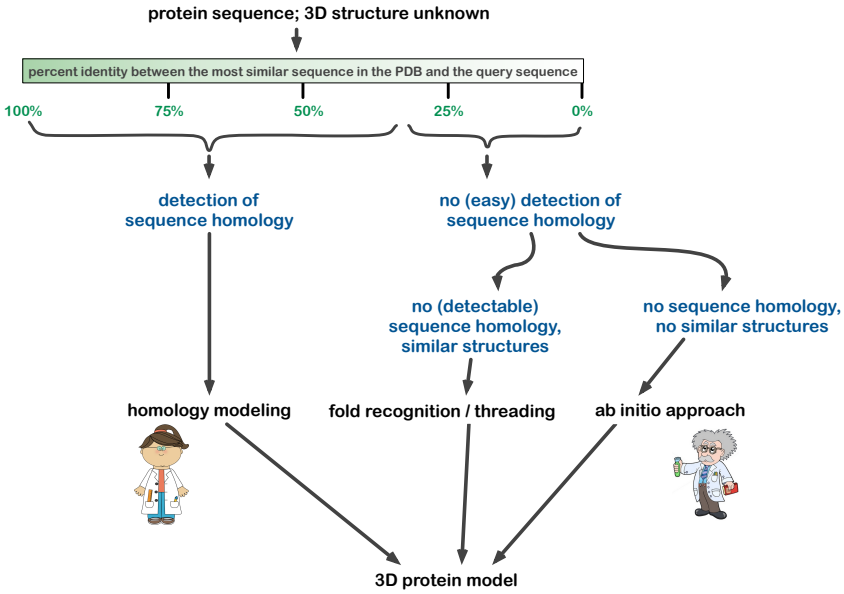
clearly related sequence, same structure



unrelated sequence, same structure

distantly related sequence, same structure

structure not yet available



3D structure prediction

is a sufficiently similar sequence found in PDB?

- YES: use homology (comparative) modeling
- model unknown structure based on known structure

assumption

- different but homologous sequences fold into the same or highly similar structures

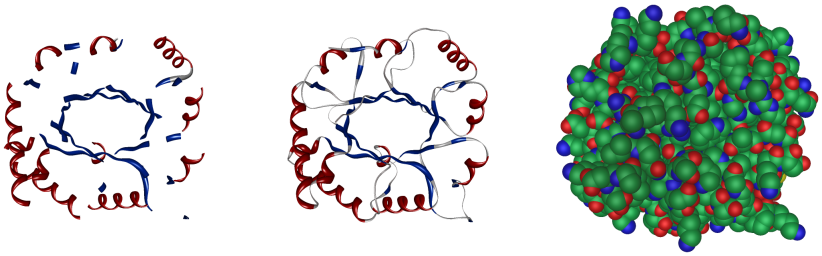
Homology modeling

1. Template recognition
   - use BLAST to compare query sequence to all sequences with a structure in PDB
   - identify homologous, highly similar sequence(s)

2. Alignment
   - optimal alignment of query sequence and homologous sequences with structures
   - structure-aware alignment

```
seq      ...VPLYKHLADLSKSKTSPYVLPVPFLNVLNGGSHAGGALALQEFMIAPTGA-KTFAEALRI...
           ||||||||  ||||| ||V      ||||GGS||GG||  ||F|||| || ||F||||||
1tim     ...QEVHEKLRGWLKTHVSDAVAV--QSRIIYGGSVTGGNCDVDGFLVGGASLKPEFVDIINA...
P00925   ...VPLYQHLADLSKSKTSPYVLPVPFLNVLNGGSHAGGALALQEFMIAPTGA-KTFAEAMRI...
P30575   ...IPLYKHIANISNAKKGKFVLPVPFQNVLNGGSHAGGALAFQEFMIAPTGV-STFSEALRI...
```

Homology modeling

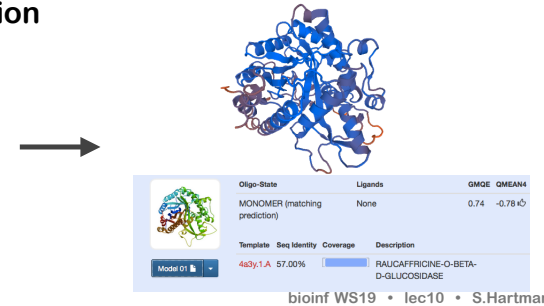


3. Backbone generation
   - copy coordinates from template for conserved regions
4. Loop building
   - based on existing structures or modeled
5. Side chain modeling

Homology modeling

6. Model refinement
   - side chains, backbone, energy calculations, etc
7. Model verification
   - check bond angles, torsion angles, bond lengths
   - check for distribution of polar/apolar residues
8. Model evaluation

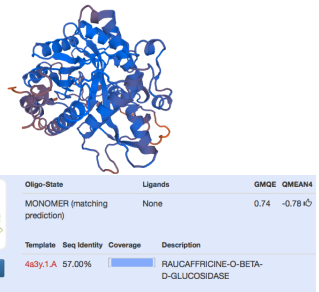
```
MAMQLGSLCAMLLIGFALRNTNAVRTDPPSHCPVL
NRSSFESLVPGFIFGTASAAYQVEGAANEGGRGPS
IWDAYTHNHPERIKDRSNGDIAIDQYHRYKEDVGI
MKNMGLDSYRLSISWSRLLPNGKLSGGVNKEGIEY
YNNLINELLRNGITPFVTLFHWDVPQALVDEYGGF
LSPRIVDHYKDYTELCFKEFGDRIKHWITLNEPYA
VSHHGYAIGIHAPGRCSDWEACLGGDSAIEPYLVT
HNQLLAHASTVKVYKDKYQASQNGVIGITVVSHWI
EPASKSKEDIDAASRYLDFMFGWFMSPLTIGDYPH
SMRHLVGERLPVFTEEQSKLLNGSFDFIGLNYYSA
RYASDFSNDYIAPPSYLTDHRANVTFELNGVPIGP
RGASDWLYVYPEGIYKLLLHTKETYNNPLIYITEN
GIDEFNNPKLSLEEALNDTMRIDYYYHHLCYLQAA
IKDGVRVKGYFAWSVLDNFEWNSGYTVRFGINYVD
YDNGLKRRSKFSAHWLKNFLKNYSGSKKEIRVRVD
DNARDTKAGYEI
```



How well does it work?

seq identity	limiting factor
<25%	template recognition
25-50%	alignment
50-75%	loop modeling
>75%	side chain modeling

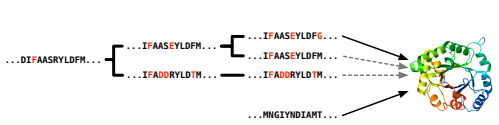
```
MAMQLGSLCAMLLIGFALRNTNAVRTDPPSHCPVL
NRSSFESLVPGFIFGTASAAYQVEGAANEGGRGPS
IWDAYTHNHPERIKDRSNGDIAIDQYHRYKEDVGI
MKNMGLDSYRLSISWSRLLPNGKLSGGVNKEGIEY
YNNLTNELLRNGITPFVTLFHWDVPQALVDEYGGF
LSPRIVDHYKDYTELCFKEFGDRIKHWITLNEPYA
VSHHGYAIGIHAPGRCSDWEACLGGDSAIEPYLVT
HNQLLAHASTVKVYKDKYQASQNGVIGITVVSHWI
EPASKSKEDIDAASRYLDFMFGWFMSPLTIGDYPH
SMRHLVGERLPVFTEEQSKLLNGSFDFIGLNYYSA
RYASDFSNDYIAPPSYLTDHRANVTTELNGVPIGP
RGASDWLYVYPEGIYKLLLHTKETYNNPLIYITEN
GIDEFNNPKLSLEEALNDTMRIDYYYHHLCYLQAA
IKDGVRVKGYFAWSVLDNFEWNSGYTVRFGINYVD
YDNGLKRRSKFSAHWLKNFLKNYSGSKKEIRVRVD
DNARDTKAGYEI
```

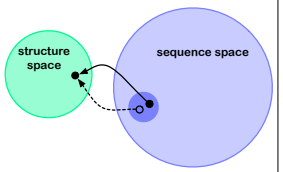


3D structure prediction

is a sufficiently similar sequence found in PDB?

- YES: use homology (comparative) modeling
- NO
  - can I find homologous sequences with more sensitive methods? (sequence-based fold recognition)
  - can I find a similar structure? (threading)

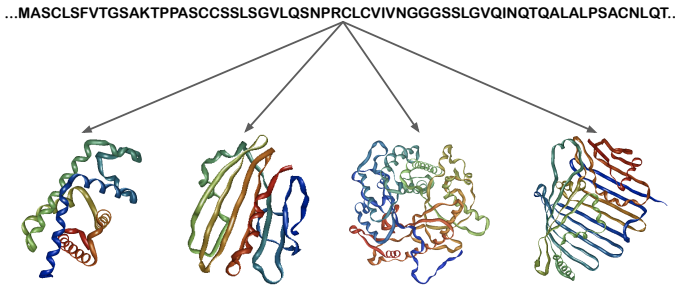




assumption

- structure space is smaller than sequence space
- physical constraints on protein structure

Structure-based approaches: threading



compute compatibility scores

- free energy of the sequence "in" structure A, B, ...?

Structure-based approaches

threading

- align (fit) a sequence to a structure
- compute compatibility between sequence & structure
- if a sequence is highly compatible with a structure, it might fold into that structure

for a given structure

- relate atomic coordinates to potential energy
- use mathematical equations & parameters
  - potential energy functions (force field)
  - based on empirical data

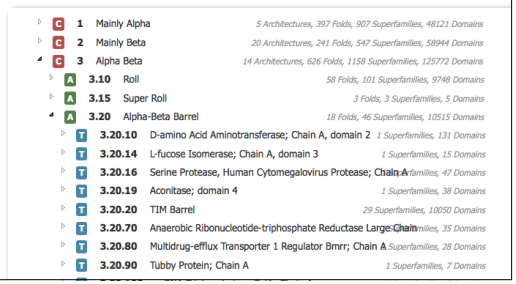
Fold library for threading

fit a sequence to all structures in PDB?

- no!
- one protein with different ligands or mutations
- many proteins with highly similar folds

use a library of non-redundant folds

- CATH topology
- SCOP fold



3D structure prediction

is a sufficiently similar sequence found in PDB?

- YES: use homology (comparative) modeling
- NO
  - was fold recognition or threading successful?
  - NO: use ab initio prediction

assumption & approach

- the native fold of a protein is the one with the lowest free energy
- (in theory) evaluate free energy of all possible conformations for a given sequence

Ab initio structure prediction

thermodynamic hypothesis

- proteins fold by following physical laws
- native conformation: global free energy minimum

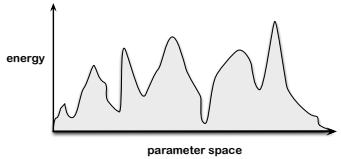
minimizing free energy

- stereochemically allowed / preferred positions and angles of backbone and side chains
- buried hydrophobic surface
- hydrogen bonds between polar residues and other polar residues (or backbone or water)
- electrostatic & Van der Waals interactions
- ...

Ab initio structure prediction

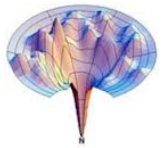
KSARPEYLQKLTYDIVLPADMP

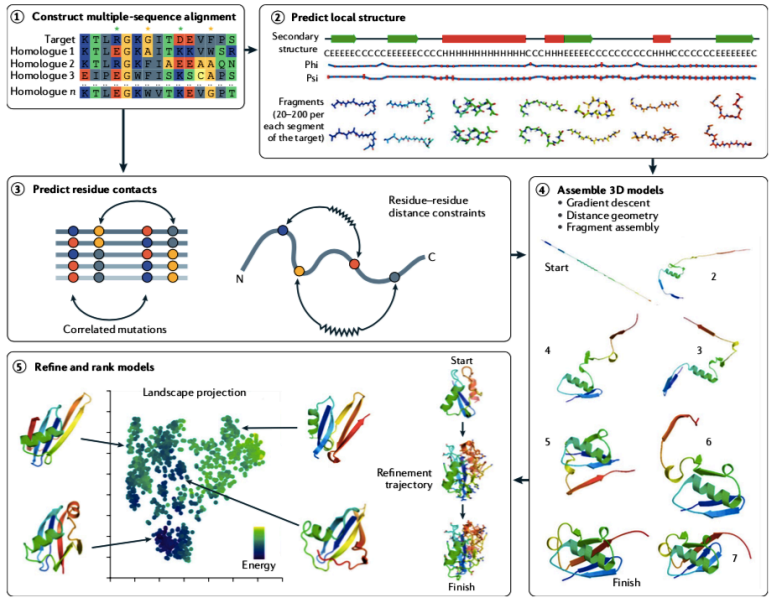
thermodynamic & physicochemical theory



challenges

- number of possible conformations
- computing power
- local minima





How well does it work?

CASP: Critical Assessment of protein Structure Prediction

- every two years, 1994 - current (CASP 13)
- structures about to be solved experimentally
- their sequences are made available
- predictions are submitted ...

  ...and then compared to the solved structures

assessment of:

- structure prediction, structure refinement, ...

http://www.predictioncenter.org

How well does it work?

CASP13: progress!

- increased amount of experimental data
- sequence data for huge protein families
- improved hardware & algorithms
- use of deep learning (deep neural networks)
  - machine learning approach
  - use of training data (sequences & their solved structures)
  - "learn"
  - predict structure of new sequence

How well does it work?

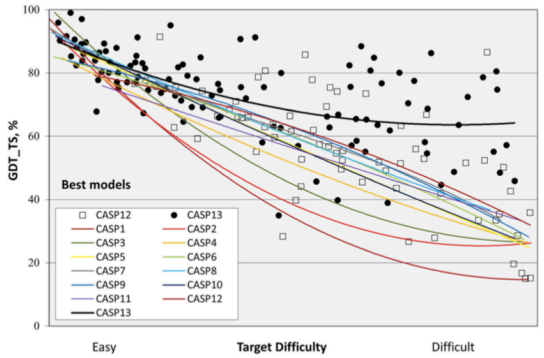


FIGURE 1 Trend lines of backbone accuracy for the best models in each of the 13 CASP experiments. Individual target points are shown for the two most recent experiments. The accuracy metric, GDT_TS, is a multiscale indicator of the closeness of the Cα atoms in a model to those in the corresponding experimental structure. Target difficulty is based on sequence and structure similarity to other proteins with known experimental structures (see Reference 5 for details). There is a striking improvement in model accuracy in CASP13 (top black line), particularly for the more difficult targets

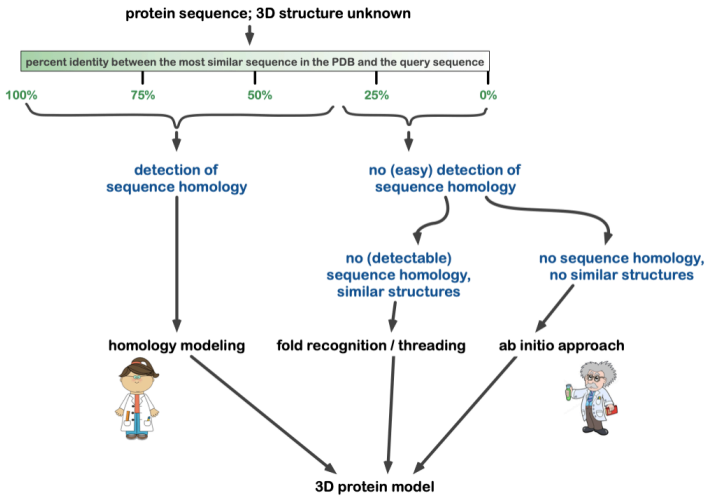
A Kryshtafovych et al., 2019. Critical assessment of methods of protein structure prediction (CASP)—Round XIII

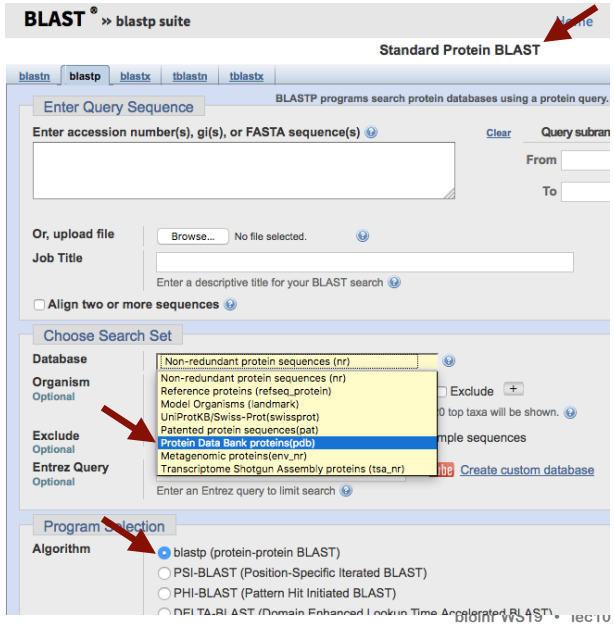
Terms and concepts

- domain, fold, super-secondary structure, CATH, threading, ab initio

- structure elements of proteins
- ~number of sequences vs. structures
- main approaches to predict 3D protein structure: principles, assumptions, applications, limitations
  - homology modeling
  - threading
  - ab initio prediction

Today's computer exercise





3 pools only!

- **1a, 1b**
- **2a**