FACULTY OF SCIENCE
Charles University

Klara Hlouchova
Research Group
Synthetic biology | Faculty of Science, Charles University

Carolina Rocha

Klara Hlouchova Research group

Department of Cell Biology
Faculty of Science
Charles University

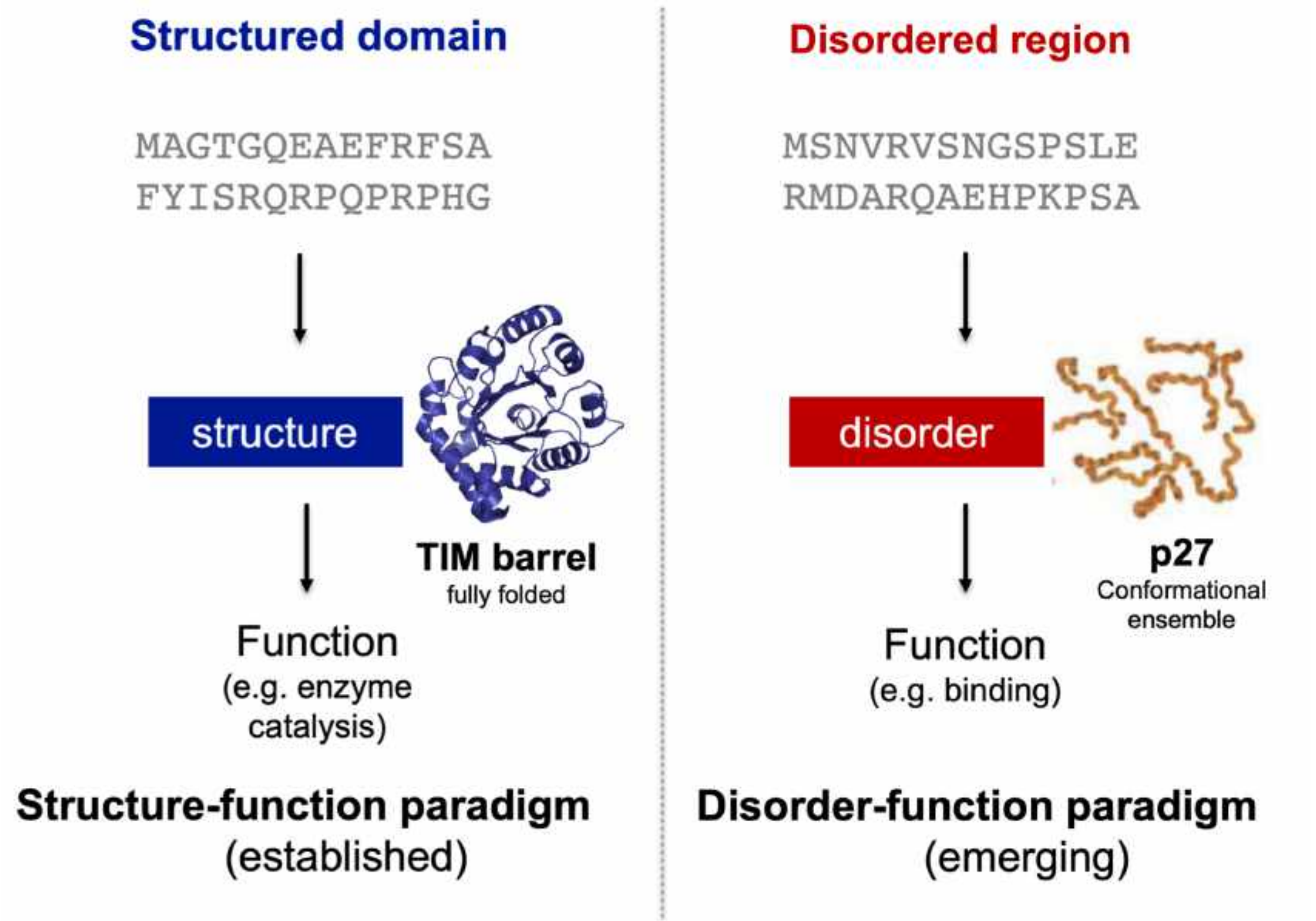
Correspondences between protein sequence and structure

Basic concepts	10 min
- Sequence-structure relationship	
- Protein folding prediction	
- Intrinsic disorder prediction	
Structure and disorder prediction	15 min
- Template-based structure prediction	
- Intrinsically disordered regions prediction	
Q&A	5 min

Sequence-structure relationship

In the 1960s, Christian Anfinsen postulated that the unique three-dimensional structure of a protein is determined by its amino acid sequence (sequence–structure–function paradigm). However, intrinsically disordered proteins and regions does not conform to this postulate. Disordered regions contribute to protein function and do not fold into a defined tertiary structure.

Babu M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human Disease. Biochemical Society Transactions. 2016. 44:1185–1200. doi: 10.1042/BST20160172



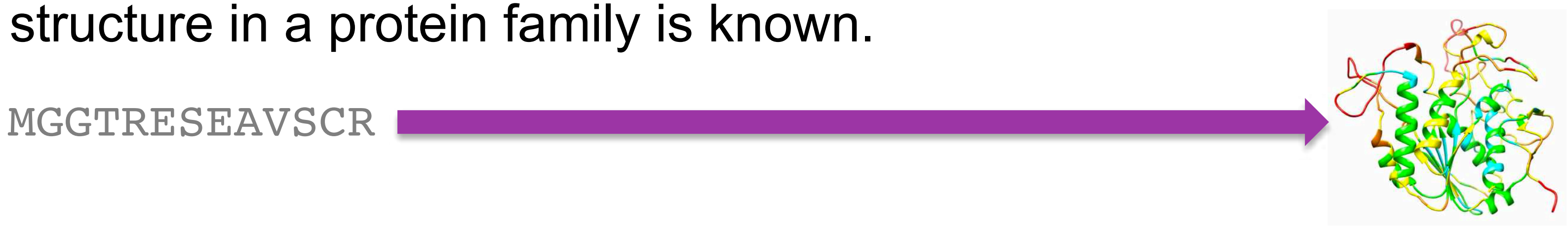
Adapted from van der Lee et al. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. 2014. 114: 6589−6631. doi: 10.1021/cr400525m

Is it possible to predict protein structure?

Yes, there are two main approaches:

❖ Template-free (or *de novo or ab initio)*

They do not use any known structures. Useful when not a single structure in a protein family is known.



MGGTRESEAVSCR

Rosetta Commons

The hub for Rosetta modeling software

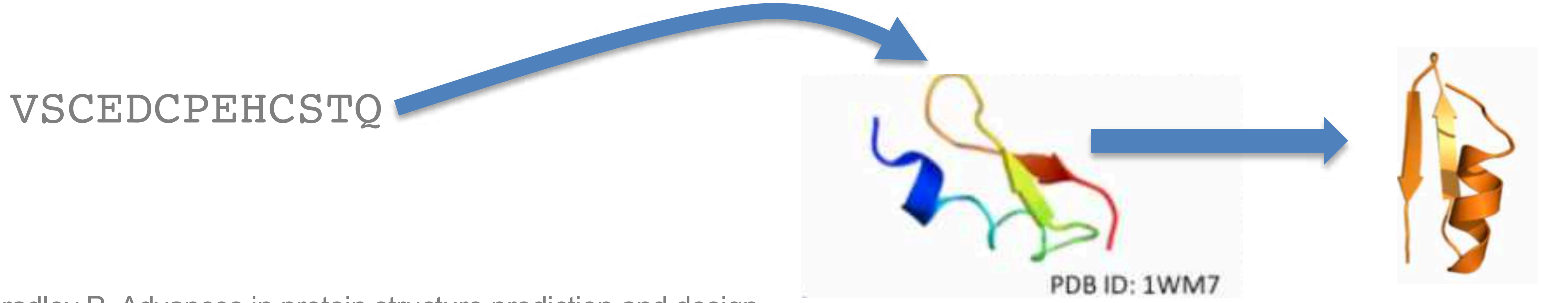
doi: 10.1016/S0076-6879(04)83004-0

AlphaFold

doi:10.1038/s41586-019-1923-7

❖ Template-based (or homology-modeling)

Use the similarity to another protein whose three-dimensional structure is known.



VSCEDCPEHCSTQ

HHpred

doi: 10.1093/nar/gki408

I-TASSER

Protein Structure & Function Predictions

doi: 10.1093/nar/gkv342

Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology. 2019. 20:681–697.

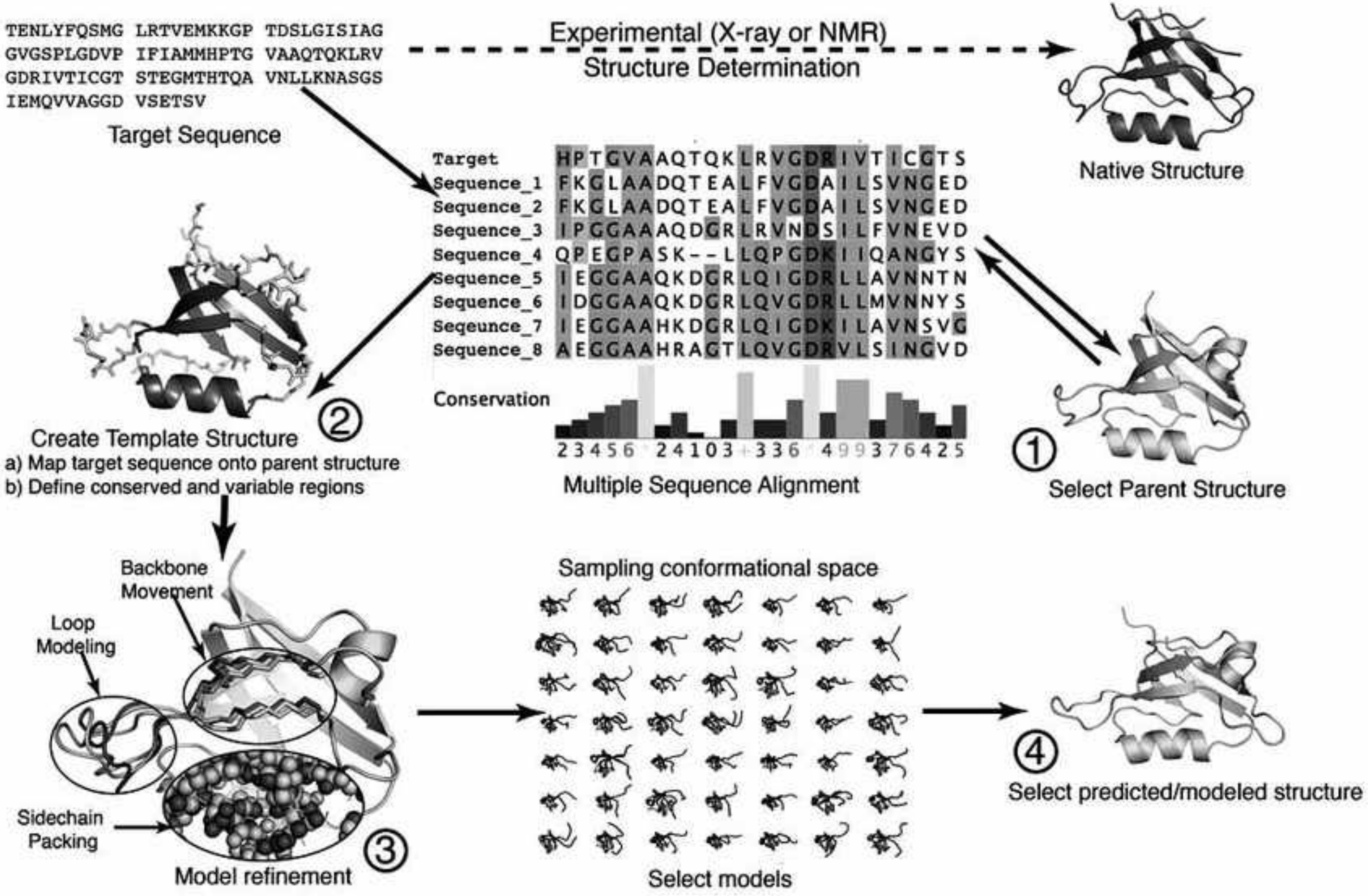
Dorn M, Barbachan M, Buriol L, Lamb L. Three-dimensional protein structure prediction: Methods and computational strategies. Comput Biol Chem. 2014. 53PB:251-276.doi: 10.1016/j.compbiolchem.2014.10.001.

Template-based (or homology model building)

The steps in standard template-based modelling involve:

1) *Selection of a suitable structural template (known structure).*
2) *Alignment of the target sequence to the template structure.*
3) *Model refinement and molecular modelling to account for mutations, insertions and deletions present in the target-template alignment.*
4) *Select your modeled structure.*

Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology. 2019. 20:681–697.

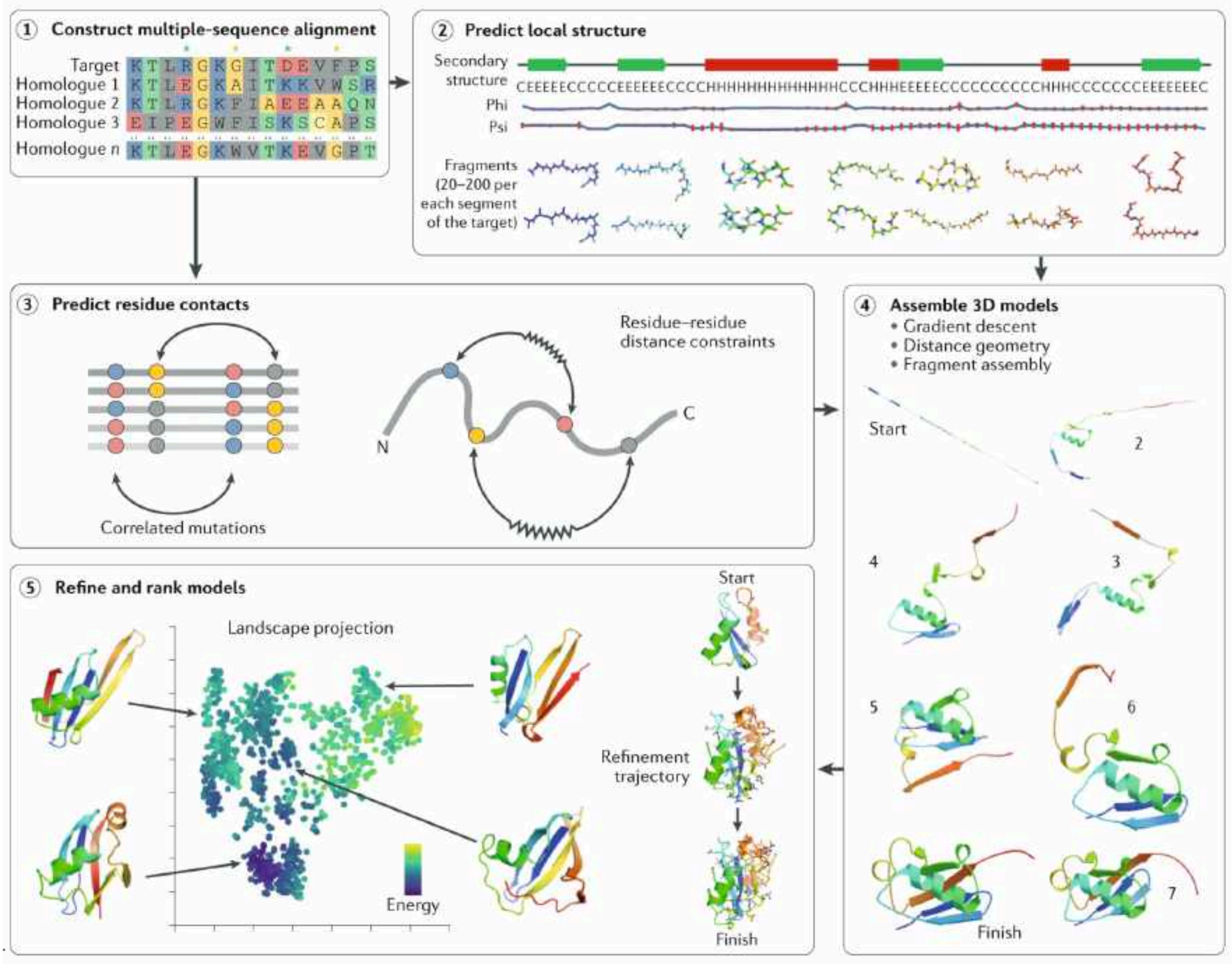

TENLYFQSMG LRTVEMKKGP TDSLGISIAG
GVGSPLGDVP IFIAMMHPTG VAAQTQKLRV
GDRIVTICGT STEGMTHTQA VNLLKNASGS
IEMQVVAGGD VSETSV
Target Sequence
Experimental (X-ray or NMR)
Structure Determination
Native Structure
Target HPTGVAAQTQKLRVGDRIVTICGTS
Sequence_1 FKGLAADQTEALFVGDAILSVNGED
Sequence_2 FKGLAADQTEALFVGDAILSVNGED
Sequence_3 IPGGAAAQDGRLRVNDSILFVNEVD
Sequence_4 QPEGPASK--LLQPGDKIIQANGYS
Sequence_5 IEGGAAQKDGRLQIGDRLLAVNNTN
Sequence_6 IDGGAAQKDGRLQVGDRLLMVNNYS
Seqeunce_7 IEGGAAHKDGRLQIGDKILAVNSVG
Sequence_8 AEGGAAHRAGTLQVGDRVLSINGVD
Conservation
23456 24103 336 499376425
Multiple Sequence Alignment
1
Select Parent Structure
2
Create Template Structure
a) Map target sequence onto parent structure
b) Define conserved and variable regions
Backbone Movement
Loop Modeling
Sidechain Packing
3
Model refinement
Sampling conformational space
Select models
4
Select predicted/modeled structure

Xiaotao Q, Rosemarie S, Ryan D, Jerry Ti. A Guide to Template Based Structure Prediction. Current Protein & Peptide Science. 2009. 10:270-285. doi: : 10.2174/138920309788452182

What if there´s a lack of any structural template (known structure)?

The strategy of template-free (de novo) folding prediction is:

1) Construction of a multiple-sequence alignment of the target protein and related sequences.
2) The sequences of the target and its homologues are then used to predict local structural features, such as secondary structure and backbone torsion angles.
3) The alignments are also useful to predict residue–residue contacts.
4) These predicted features guide the process of building 3D models of the target protein structure.
5) The models will be refined, ranked and compared with one another to select the final predictions.



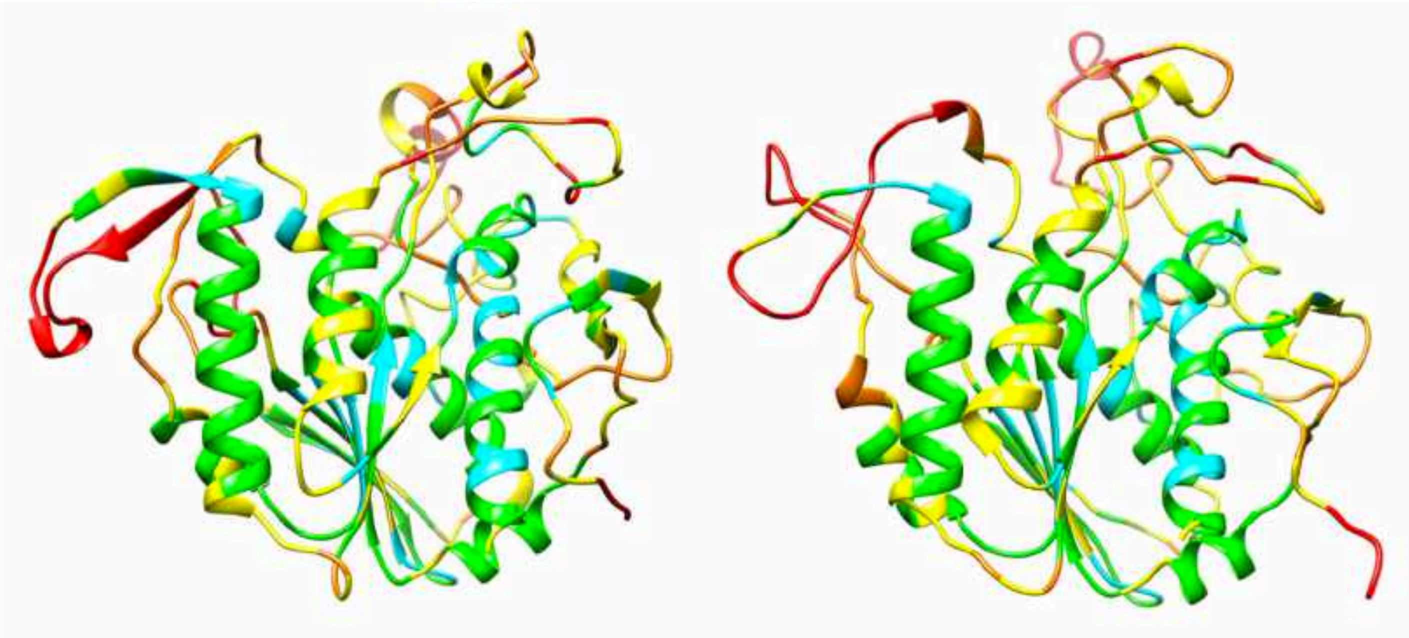
1 Construct multiple-sequence alignment
Target KTLRGKGITDEVFPS
Homologue 1 KTLEGKAITKKVWSR
Homologue 2 KTLRGKFIAEEAAQN
Homologue 3 EIPEGWFISKSCAPS
Homologue n KTLEGKWVTKEVGPT
2 Predict local structure
Secondary structure
CEEEEECCCCEEEEEEECCCCHHHHHHHHHHHHHCCCHHHEEEEECCCCCCCCCCHHHCCCCCCCEEEEEEEC
Phi
Psi
Fragments (20–200 per each segment of the target)
3 Predict residue contacts
Correlated mutations
N
C
Residue–residue distance constraints
4 Assemble 3D models
• Gradient descent
• Distance geometry
• Fragment assembly
Start
2
3
4
5
6
7
Finish
5 Refine and rank models
Landscape projection
Energy
Start
Refinement trajectory
Finish

Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology. 2019. 20:681–697.

CASP (Critical Assessment of Structure Prediction)

Created in 1994, CASP is a community wide experiment to determine and advance the state of the art in modeling protein structure from amino acid sequence.

The most recent CASP13 (2018) saw a dramatic progress in template-free modeling by using deep learning techniques to predict inter-residue distances. *With the proviso that there are an adequate number of sequences known for the protein family, the new methods essentially solve the long-standing problem of predicting the fold topology of monomeric proteins.*



X-ray structure of Xylan acetyltransferase (unknown for the participants in CASP)

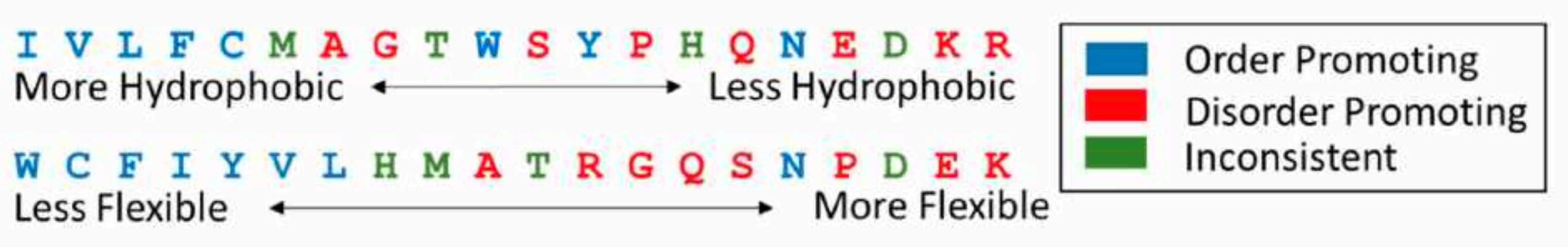
Most accurate CASP model by template-free modeling

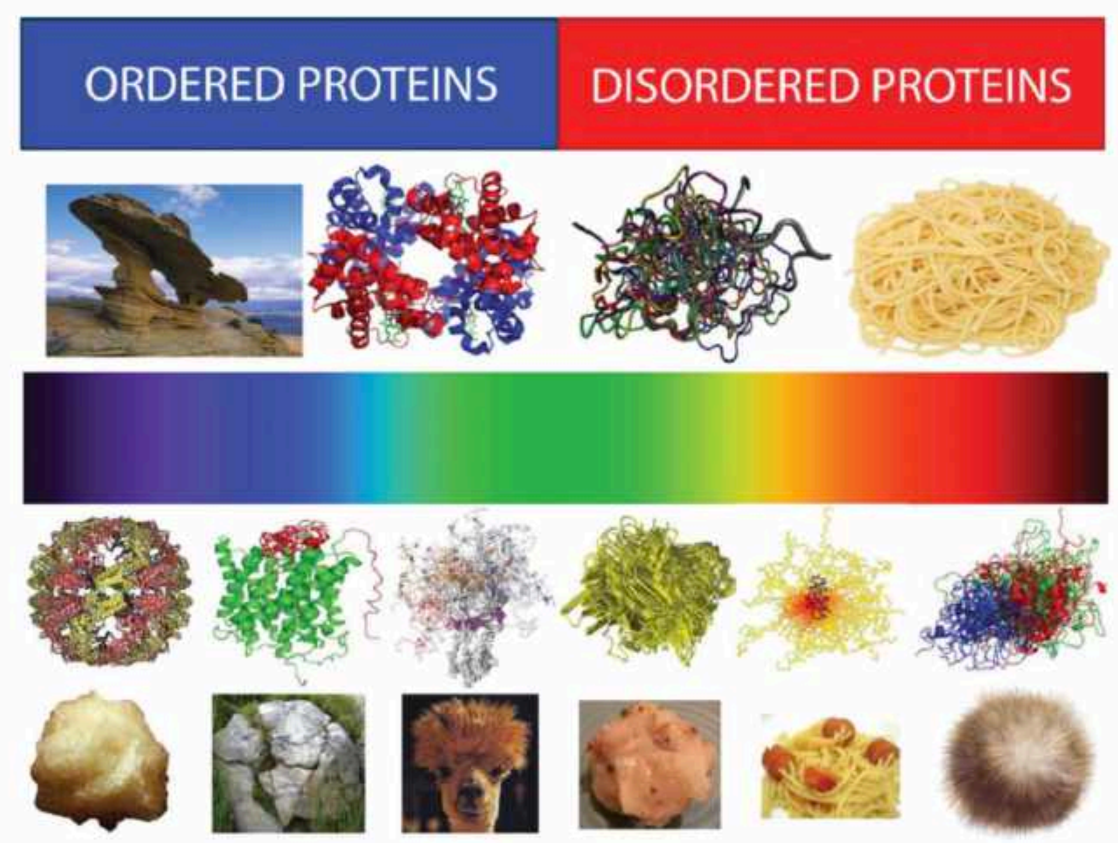
Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. 2019. doi:10.1002/prot.25823.

Is it achievable to predict intrinsic disorder?

The distinct sequence features that are present in IDPs and IDRs allow the construction of sequence based rules that can facilitate high performance disorder prediction.

DeForte S, Uversky V. Order, Disorder, and Everything in Between. Molecules. 2016. 21,1090. doi:10.3390/molecules21081090.





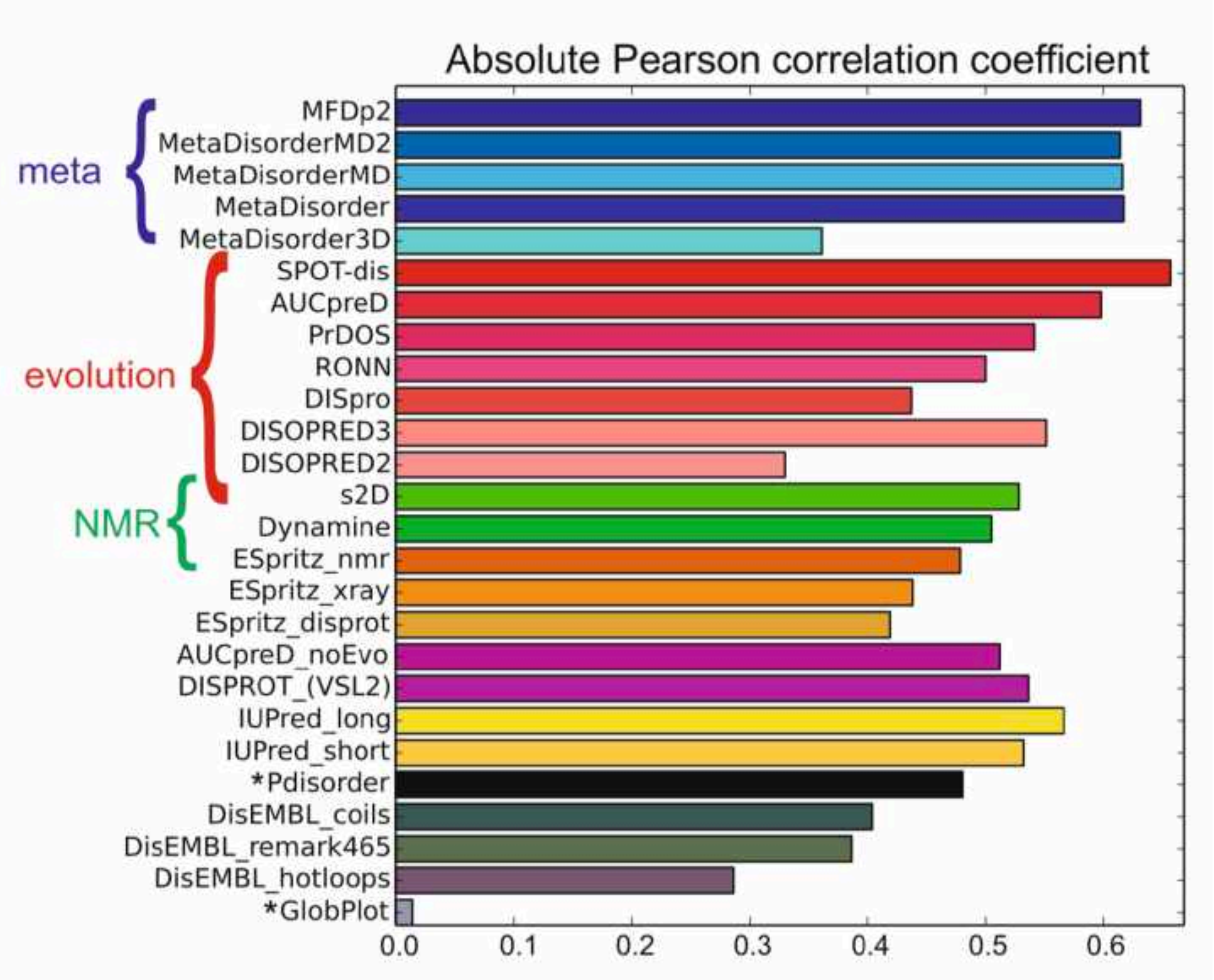
Uversky V. A decade and a half of protein intrinsicdisorder: Biology still waits for physics. PROTEIN SCIENCE .2013. 22:693-724. doi:10.1002/pro.2261

Strategies of Intrinsic Disorder Prediction

Three general prediction strategies currently exist:

Meta-predictors	Combine several individually successful disorder prediction methods have been developed more recently, resulting in increases in prediction accuracy
Machine learning	Use of training sets for machine learning. For instance: unresolved residues in X-ray structures; linear support vector machines (SVMs) trained on PSI-BLAST sequence profiles surrounding unresolved residues.
Sequence properties	Estimation residue interaction energies. Sequences with lower predicted pairwise interaction energies are considered more likely to be disordered due to a lack of stabilizing contacts.

van der Lee et al. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. 2014. 114: 6589−6631. doi: 10.1021/cr400525m



Ranking of disorder prediction methods according to the absolute Pearson linear correlation coefficient between estimated disorder probability and Z-score.

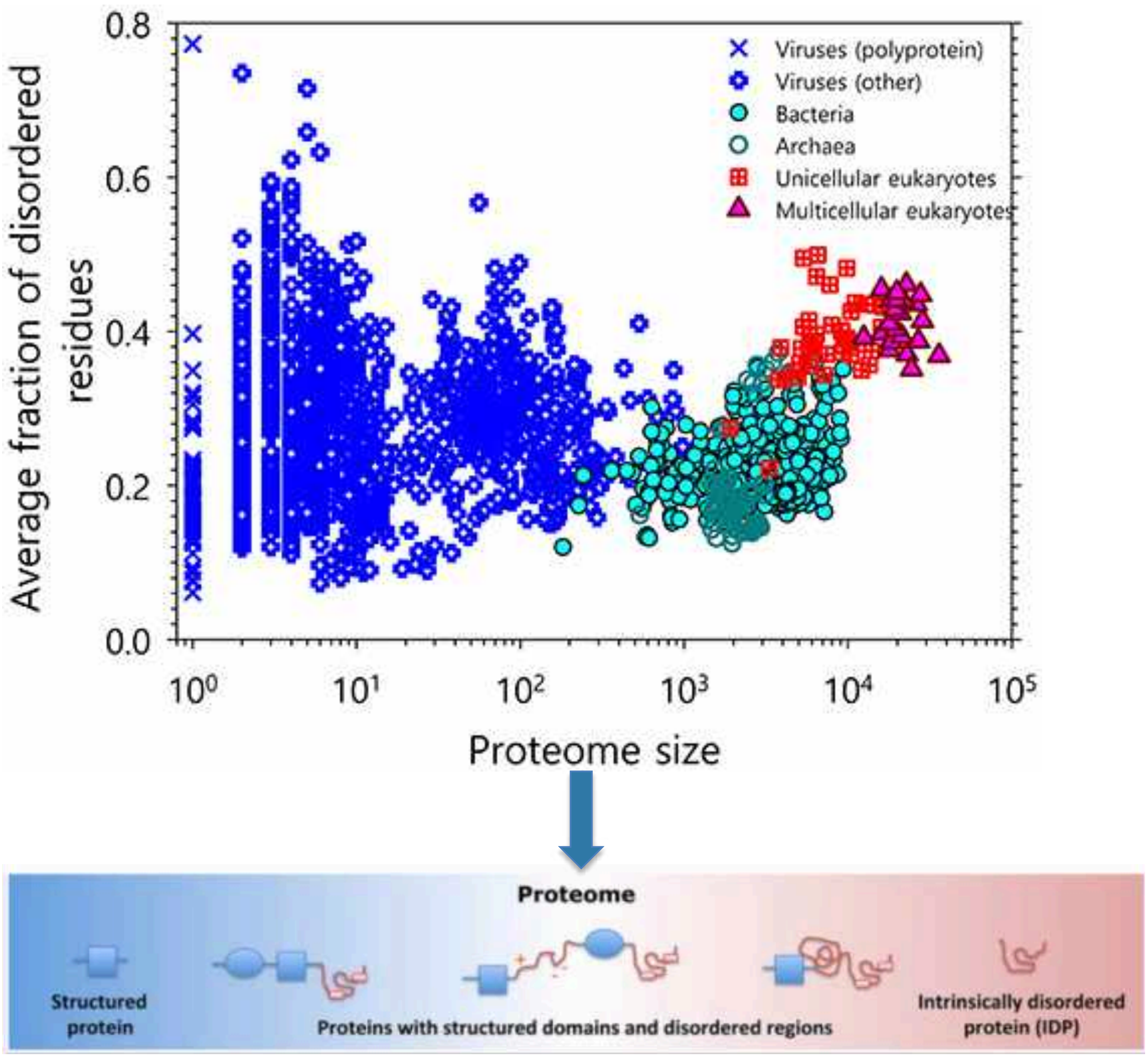
Nielsen J, Mulder F. Quality and bias of protein disorder Predictors. Scientific Reports. .2019. 9:5137. doi.org/10.1038/s41598-019-41644-w

How common is disorder in Biology?

Disorder predictors were pivotal for the establishment of the IDP field.

30% of human proteome is composed of intrinsically disordered proteins and IDRs.

The bioinformatic analysis of the proteomes of organisms of the three domains of life, Bacteria, Archaea and Eukarya revealed the presence of disordered proteins and regions in all known proteomes.



Uversky V. Introduction to Intrinsically Disordered Proteins (IDPs). Chem. Rev. 2014. 114:6557−6560. doi.org/10.1021/cr500288y.
Uversky V. Intrinsically Disordered Proteins and Their “Mysterious” (Meta) Physics. Front. Phys. 2019 .doi.org/10.3389/fphy.2019.00010.
Adapted from van der Lee et al. Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. 2014. 114: 6589−6631. doi: 10.1021/cr400525m

CAID (Critical Assessment of Intrinsic protein Disorder)

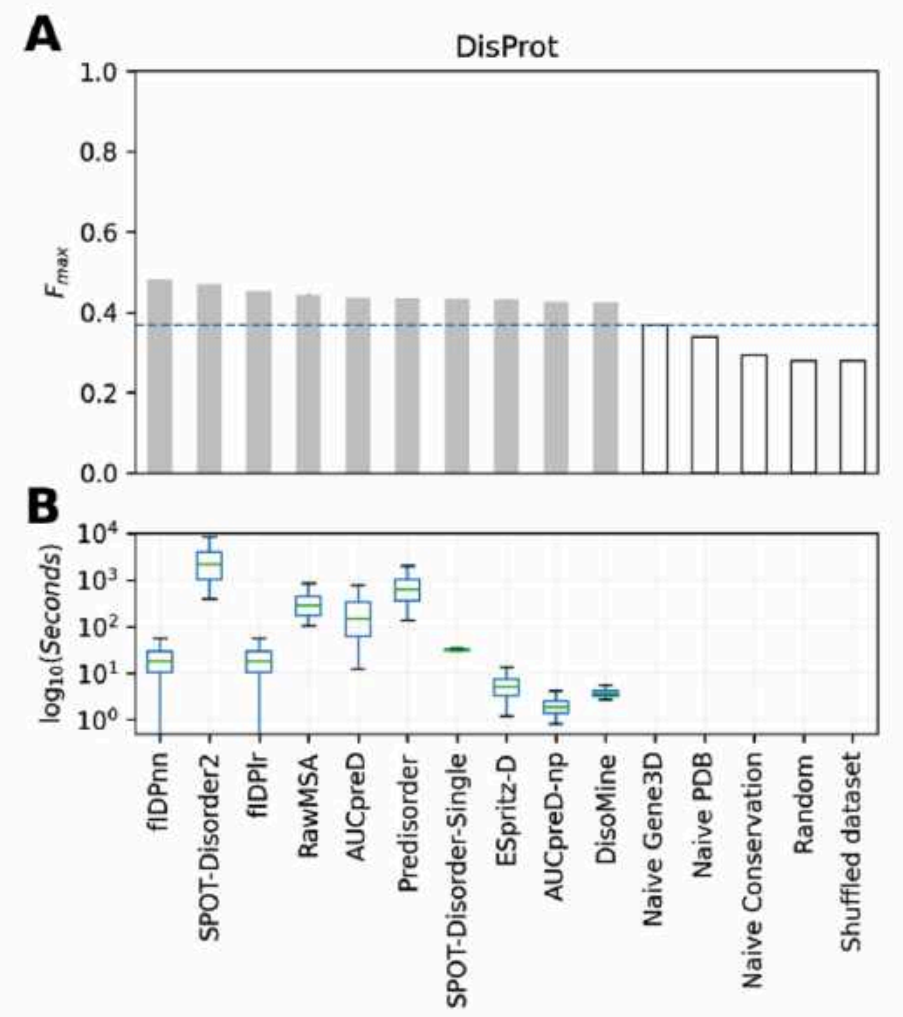
Although disorder prediction accuracy was evaluated also by CASP, in 2018 was created CAID, a community wide experiment to determine and advance the state of the art in the detection of intrinsically disordered residues form the amino acid sequence.

CAID has two main prediction categories:

i) intrinsic structural disorder

ii) binding sites found in IDRs (known as MoRFs, SLIMs or LIPs).

Necci M, Piovesan D, CAID Predictors , DisProt Curators . Tosatto S. Critical Assessment of Protein Intrinsic Disorder Prediction. doi:10.1101/2020.08.11.245852



Performance of predictors for the top ten best ranking methods (A) and the distribution of execution time per-target (B).

Adapted from Necci M, Piovesan D, CAID Predictors , DisProt Curators . Tosatto S. Critical Assessment of Protein Intrinsic Disorder Prediction. doi:10.1101/2020.08.11.245852

Amino acid CompOsitioN in protein disorder

How to design disorder?

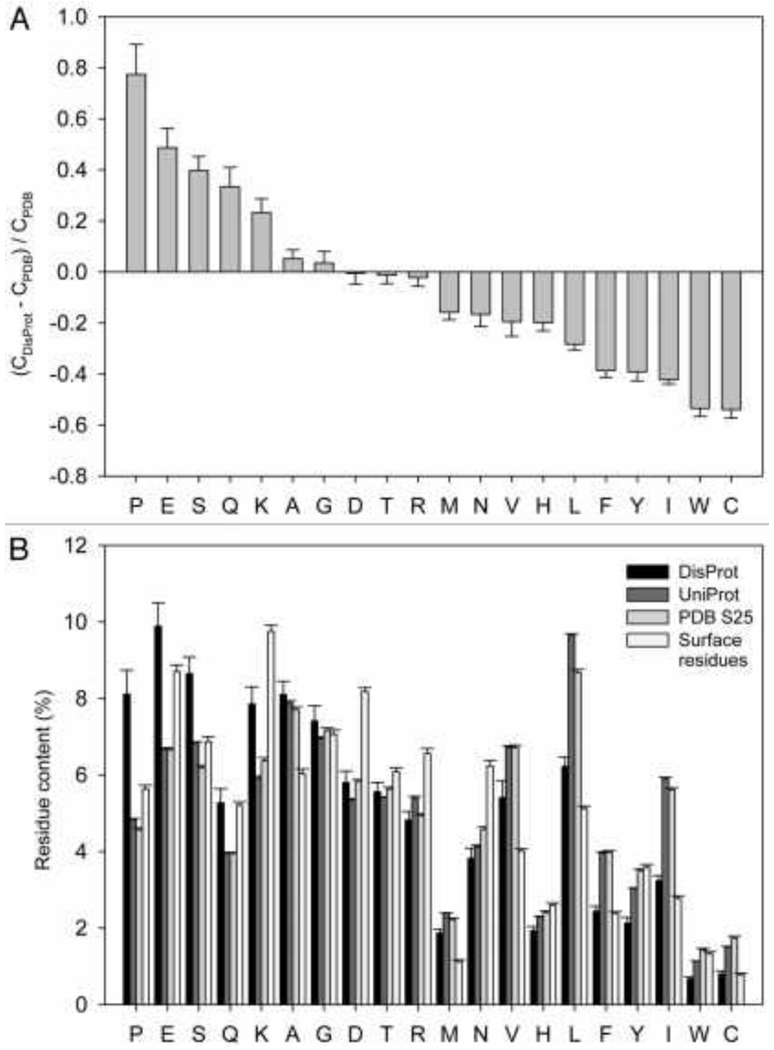
- low complexity domains
- chains with defined amino acid ratios

What if we want to design many disordered sequences in one tube?

- library approach
- combinatorial design of composition-centric libraries

How to design such a library?

- mixtures of specific degenerate codons
- one DNA library – many IDP coding templates



Vymětal,J. *et al.* (2019) Sequence versus composition: What prescribes IDP biophysical properties? *Entropy*, **21**, 1–8.
Uversky, Vladimir N. "The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins." *Intrinsically disordered proteins* 1.1 (2013): e24684.

CoLiDe – combinatorial library design

https://github.com/voracva1/CoLiDe

Input – amino acid composition and length of the library
Output – degenerate nucleotide string for combinatorial library synthesis

length of library proteins

codons which will NOT appear in the resulting library (e.g. ATG, AAA...)

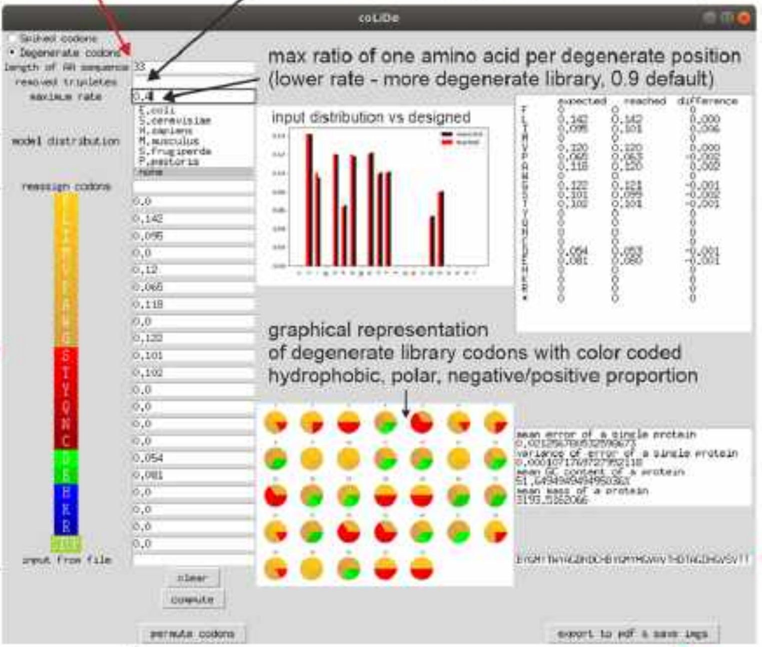
type of codons for degenerate library construction

organismal codon bias options

assign occurency to specific codon (e.g AAA 0.3 - codon AAA will appear in 30% of library)

amino acid distribution manual input (values are normalized to sum to 1 automatically)

amino acid distribution file input (format in manual)



mean error of a single protein
0,00212567...
variance of error of a single protein
0,00010717...
mean GC content of a protein
51,6494949494950363
mean mass of a protein
3193,5352066

max ratio of one amino acid per degenerate position (lower rate - more degenerate library, 0.9 default)

graphical representation of degenerate library codons with color coded hydrophobic, polar, negative/positive proportion

statisics on single amino acids input vs designed

overall designed library statistics

output degenerate string (corresponds to pie chart library representation)

shuffle codon arrangement (shuffles picture and output degenerate string)

generate PDF and text output file with all the statistics and degenerate template string

Tretyachenko V. *et al.*, in press *Bioinformatics,* 2020