When homology modeling does not work?

Jon K. Lærdahl,
Structural Bioinformatics

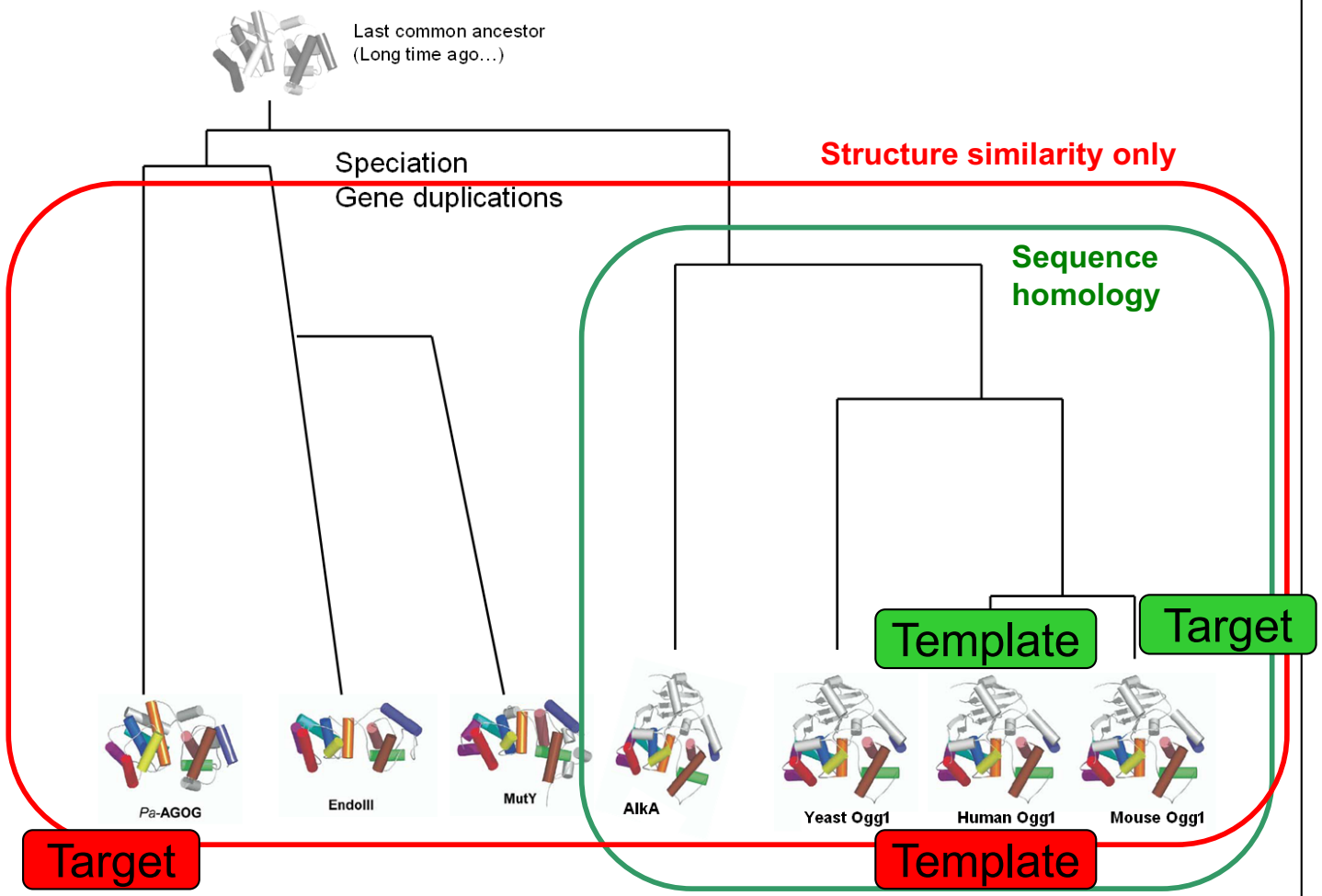
Homology modeling

Jon K. Lærdahl,
Structural Bioinformatics

Start with a protein sequence (tar

Do sequence search in all "PDB

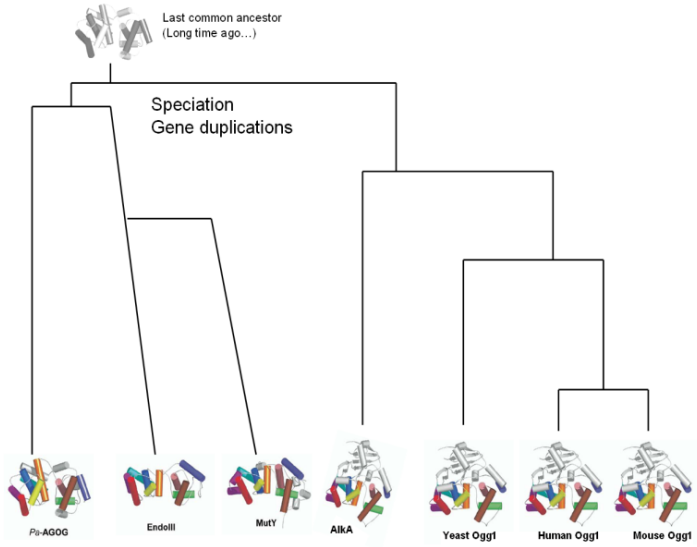
1. Template selecti
   - Find template sequences
2. Correct alignmer
   - Use the best M
   - Correct placer and deletions
3. Backbone model
4. Model loops and
   - Rotamer libra
   - Loop modelin or *ab initio* me
5. Refine and optim
6. Validate and che



Threading

Jon K. Lærdahl,
Structural Bioinformatics

- Same as “fold recognition”
- Prediction of the structural fold of a protein sequence by ”fitting” the sequence onto structures from a structural database
- Secondary structure prediction is important to choose the best template candidates
- Calculate energy and other parameters for all possibilities
- Choose the best fold, for example the one with the lowest energy
- Detects structural similarity in the absence of sequence similarity
    - GenThreader
    - Phyre2
    - Fugue
- May only be used to generate a *rough model*
- Threading does *not* give accurate models!
- May be used to detect remote homologs
    - No hits with BLAST or PSI-BLAST?
    - *Try threading!*
- “BLAST will give you the protein family”
- “Threading will give you the protein superfamily”
- Might argue: threading is more useful for detecting homology than for generating 3D structures



Jon K. Lærdahl,
Structural Bioinformatics

Threading/Fold recognition

```
>Unknown_protein
MPARALLPRRMGHRTLASTPALWASIPCPRSELRLDLVL
PSGQSFRWREQSPAHWSGVLADQVWTLTQTEEQLHCTVY
RGDKSQASRPTPDELEAVRKYFQLDVTLAQLYHHWGSVD
SHFQEVAQKFQGVRLLRQDP
GMVERLCQAFGPRLIQLDDV
LRKLGLGYRARYVSASARAI
EAHKALCILPGVGTKVADCI
IAQRDYSWHPTTSQAKGPSP
WAQATPPSYRCCSVPTCANP
RWGTLDKEIPQAPSPPFPTS
KARHPQIKQSVCTTRWGGGY
```

What is the structure of this?

Phyre2

Email	jonkl@medisin.uio.no
Description	ALKBH1___
Date	Sun Nov 16 15:24:04 GMT 2014
Unique Job ID	adf2998f1014a31f
Sequence	MGKMAAAVGS ... Download FASTA
Job Expiry	29 days Renew for 30 days

Download zip of all results

Top model

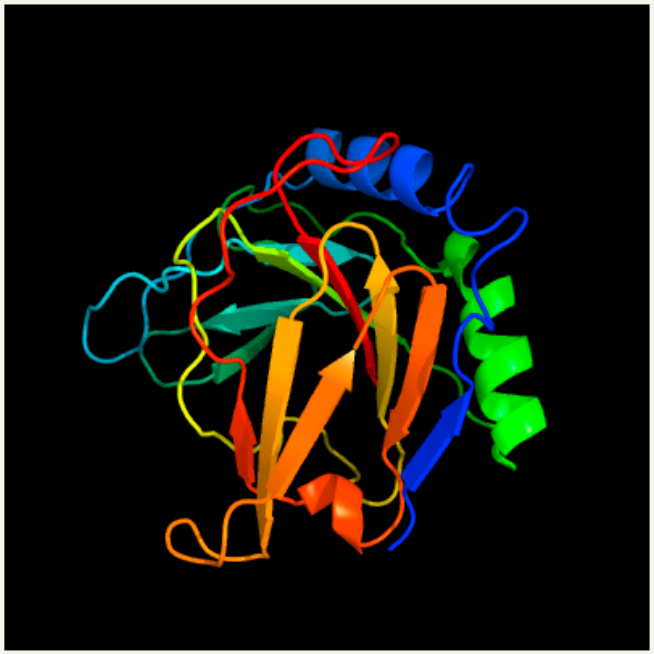


Image coloured by rainbow N → C terminus

Model dimensions (Å): **X**:42.352 **Y**:42.991 **Z**:41.673

Model (left) based on template d2fdia1

Top template information

Fold: Double-stranded beta-helix
Superfamily: Clavaminate synthase-like
Family: AlkB-like

Confidence and coverage

Confidence:	100.0%	Coverage:	50%

194 residues (50% of your sequence) have been modelled with 100.0% confidence by the single highest scoring template.

Additional confident templates have been detected (see Domain analysis) which cover other regions of your sequence.

259 residues (67%) could be modelled at >90% confidence using multiple-templates.

You may wish to try resubmitting your sequence in "intensive" mode to model more of your sequence.

3D viewing

Interactive 3D view in JSmol

For other options to view your downloaded structure offline see the FAQ

Jon K. Lærdahl,
Structural Bioinformatics

Threading/Fold recognition

Quickphyre results for job AlkD_____ - Windows Internet Explorer

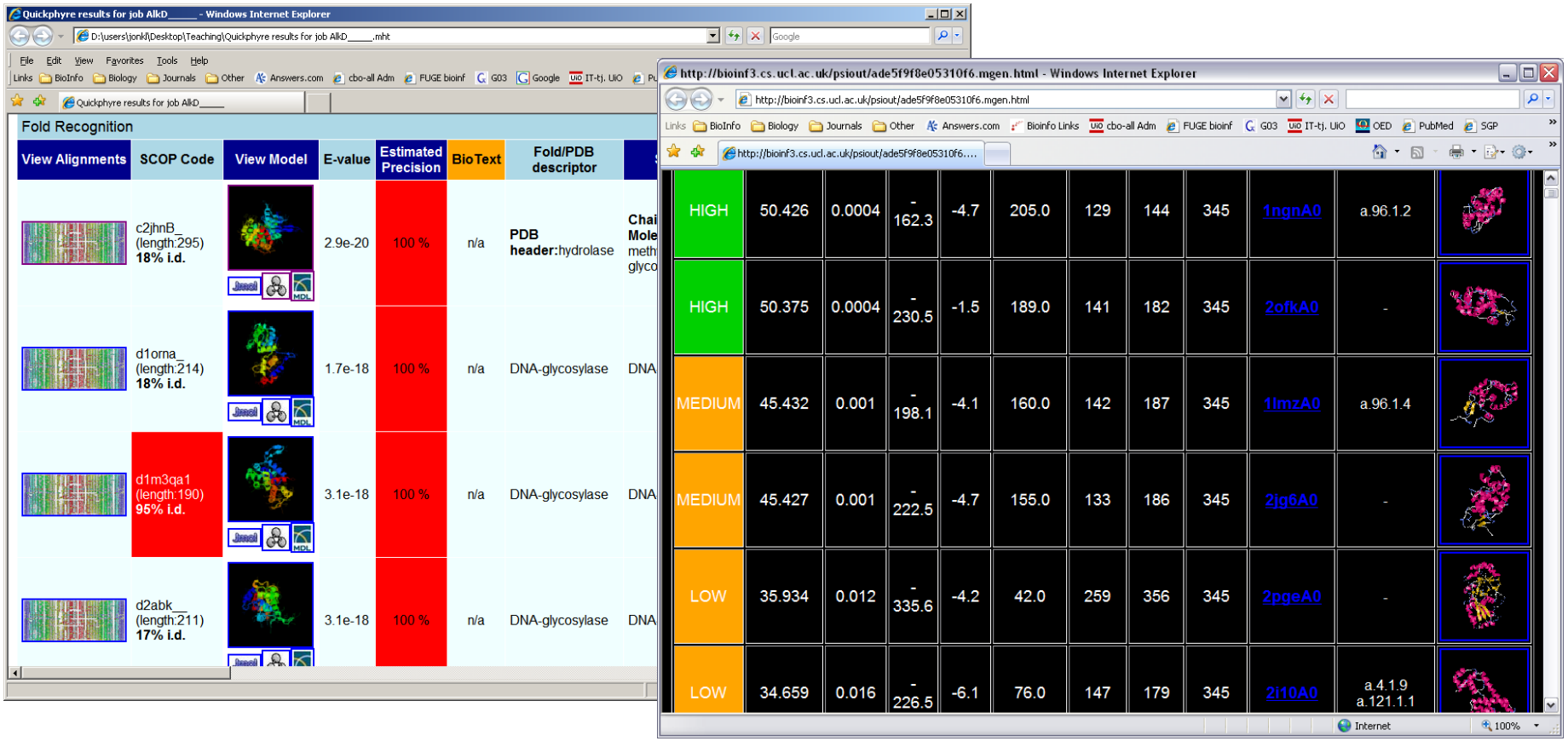
D:\users\jonkl\Desktop\Teaching\Quickphyre results for job AlkD_____.mht

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily	Family
	c2jhnB_ (length:295) **18% i.d.**		2.9e-20	100 %	n/a	**PDB header:**hydrolase	**Chain:** B: **PDB Molecule:**3-methyladenine dna-glycosylase;	**PDB Title:** 3-methyladenine dna-glycosylase from archaeoglobus fulgidus
	d1orna_ (length:214) **18% i.d.**		1.7e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	Endonuclease III
	d1m3qa1 (length:190) **95% i.d.**		3.1e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	DNA repair glycosylase, 2 C-terminal domains
	d2abk__ (length:211) **17% i.d.**		3.1e-18	100 %	n/a	DNA-glycosylase	DNA-glycosylase	Endonuclease III

Jon K. Lærdahl,
Structural Bioinformatics

Threading/Fold recognition



View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor
	c2jhnB_ (length:295) **18% i.d.**		2.9e-20	100 %	n/a	**PDB header:**hydrolase
	d1orna_ (length:214) **18% i.d.**		1.7e-18	100 %	n/a	DNA-glycosylase
	d1m3qa1 (length:190) **95% i.d.**		3.1e-18	100 %	n/a	DNA-glycosylase
	d2abk__ (length:211) **17% i.d.**		3.1e-18	100 %	n/a	DNA-glycosylase

HIGH	50.426	0.0004	-162.3	-4.7	205.0	129	144	345	1ngnA0	a.96.1.2
HIGH	50.375	0.0004	-230.5	-1.5	189.0	141	182	345	2ofkA0	-
MEDIUM	45.432	0.001	-198.1	-4.1	160.0	142	187	345	1lmzA0	a.96.1.4
MEDIUM	45.427	0.001	-222.5	-4.7	155.0	133	186	345	2jg6A0	-
LOW	35.934	0.012	-335.6	-4.2	42.0	259	356	345	2pgeA0	-
LOW	34.659	0.016	-226.5	-6.1	76.0	147	179	345	2i10A0	a.4.1.9 a.121.1.1

As for other bioinformatics methods:

- Precision might be overestimated
- The results might be completely wrong
- If several independent tools give the same result it is much more likely to be correct
- *Use several tools!*

3D structure prediction – Summary 1

Jon K. Lærdahl,
Structural Bioinformatics

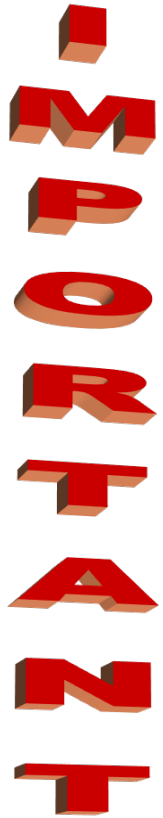
3 methods:

Ab initio modeling == *De novo* modeling

Knowledge-based modeling:

Homology modeling
Threading == Fold recognition

Your results will be predictions: They must be checked with experiments!

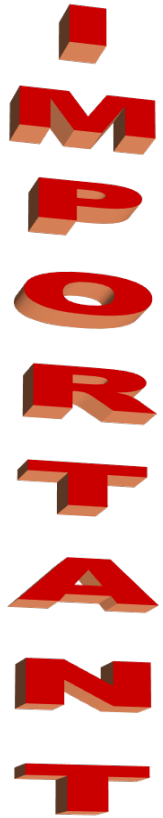


3D structure prediction - Summary

Jon K. Lærdahl,
Structural Bioinformatics

Start with target sequence

1. Sequence homology to protein that has structure in PDB (better than 20-30% sequence identity) → Homology modeling
2. No good hit with sequence searching:
   - Fold recognition/threading might give correct fold
3. No results from fold recognition/threading:
   - You *might* try *ab initio* folding, but the result will most likely be very unreliable



Homology models can be of good quality and might be useful for:

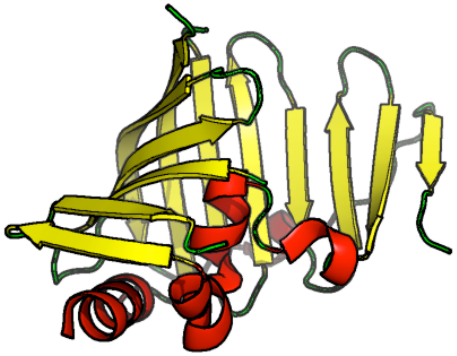
- Docking two or more proteins together
- Designing drugs
- Identifying active sites and amino acids for generating mutant proteins, etc.

Fold recognition/threading might give the protein overall fold and possibly indicate function

If the fold is that of a helical cytokine → Your protein is also possibly a helical cytokine

How to find a homolog…

- Try blast (will find close homologs) or similar
  - Protein search will find more remote homologs than nucleotide search
- Then try psi-blast (will find less close homologs, that still have some sequence similarity)
- If you known the 3D structure of your query protein, use Dali or similar and search in PDB – will find remote homologs if template structure is known
- Else, try fold recognition – might find homologs if template structure is known
- Homology can tell you about function, structure, etc.



“Modeling” or “Experiment”

Jon K. Lærdahl,
Structural Bioinformatics

nature chemical biology

ARTICLE

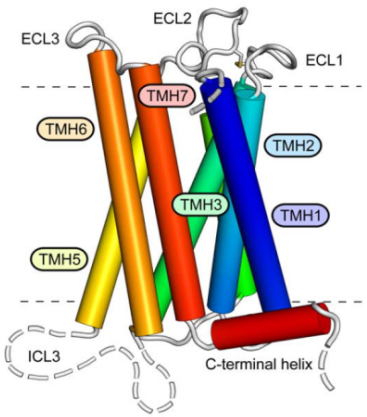
PUBLISHED ONLINE: 18 SEPTEMBER 2011 | DOI: 10.1038/NCHEMBIO.662

Ligand discovery from a dopamine D_3 receptor homology model and crystal structure

Jens Carlsson[1,5], Ryan G Coleman[1,5], Vincent Setola[2,5], John J Irwin[1], Hao Fan[1,3,4], Avner Schlessinger[1,3,4], Andrej Sali[1,3,4], Bryan L Roth[2*] & Brian K Shoichet[1*]

G protein-coupled receptors (GPCRs) are intensely studied as drug targets and for their role in signaling. With the determination of the first crystal structures, interest in structure-based ligand discovery increased. Unfortunately, for most GPCRs no experimental structures are available. The determination of the D_3 receptor structure and the challenge to the community to predict it enabled a fully prospective comparison of ligand discovery from a modeled structure versus that of the subsequently released crystal structure. Over 3.3 million molecules were docked against a homology model, and 26 of the highest ranking were tested for binding. Six had affinities ranging from 0.2 to 3.1 μM. Subsequently, the crystal structure was released and the docking screen repeated. Of the 25 compounds selected, five had affinities ranging from 0.3 to 3.0 μM. One of the new ligands from the homology model screen was optimized for affinity to 81 nM. The feasibility of docking screens against modeled GPCRs more generally is considered.

GPCRs are a large family of membrane proteins that are critical for signal transduction. They have been a major focus of pharmaceutical research and are the primary targets of almost 30% of approved drugs[1]. All of these drugs were discovered without the aid of receptor structures by classical ligand-based medicinal chemistry. Accordingly, many of these drugs reflect their origins as mimics of natural signaling molecules. The determination of the first drug-relevant GPCR structures in the last 4 years[2–4] has opened up opportunities for structure-based discovery



CASP: Critical Assessment of Techniques for Protein Structure Prediction
- Benchmarking of structure prediction tools

Jon K. Lærdahl,
Structural Bioinformatics

STOP RUSSIA'S INVASION IN UKRAINE

Protein Structure Prediction Center

Menu

- Home
- FORCASP Forum
- PC Login
- PC Registration
- CASP Experiments
  - CASP ROLL
  - CASP11 (2014)
    - Home
    - Targets
      - Target List
      - Target Submission
  - CASP10 (2012)
  - CASP9 (2010)
  - CASP8 (2008)
  - CASP7 (2006)
  - CASP6 (2004)
  - CASP5 (2002)
  - CASP4 (2000)
  - CASP3 (1998)
  - CASP2 (1996)
  - CASP1 (1994)
- Initiatives
- Data Archive
- Local Services
- Proceedings
- Feedback
- Assessors
- People
- Community Resources

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been ten previous CASP experiments. The eleventh experiment will start in May 2014. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012) | CASP11 (2014)

Raw data for the experiments held so far are archived and stored in our data archive.

In November 2011 we have opened a new rolling CASP experiment for all-year-round testing of ab initio modeling methods:

CASP ROLL

Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. CASP proceedings include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on this web page. The main numerical measures used in evaluations are described in the papers [1], [2]. The latter paper also contains explanations of data handling procedures and guidelines for navigating the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

CASP ROLL | CASP 10 | CASP 11

FORCASP "no more dead trees"

Discussion Forum

Message Board

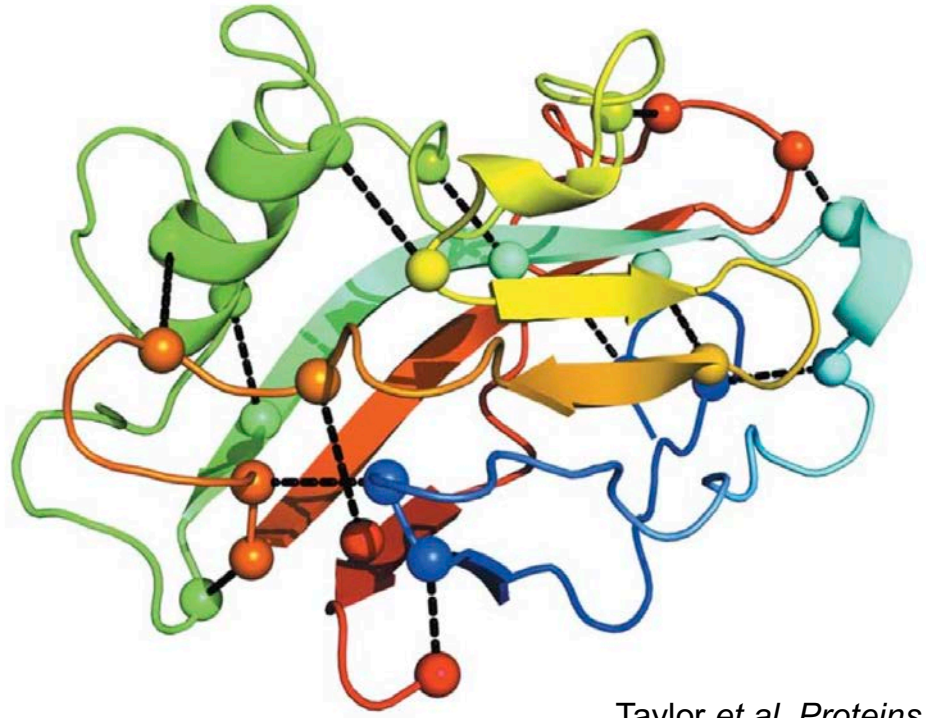
CASP11 registration opens March 31
Dear CASP Participants, Exiting news: new CASP experiment is just around the corner! We hope that you are full of enthusiasm and anxiety (as we are) and have your computers greased and warmed up. ...

Resuming CASP ROLL
Dear CASPers, Best regards for all of you in the New Year! Hoping that you had good rest after the CASP10 experiment and meeting, we are resuming CASP ROLL with two new targets later this week. ...

Predictors meeting in Gaeta
Dear CASP10 Participants, On the last day of the Meeting we will have our regular Predictors get-together. In advance, I would like to ask you to send in any comments regarding the CASP process in ...

Contact assisted methods

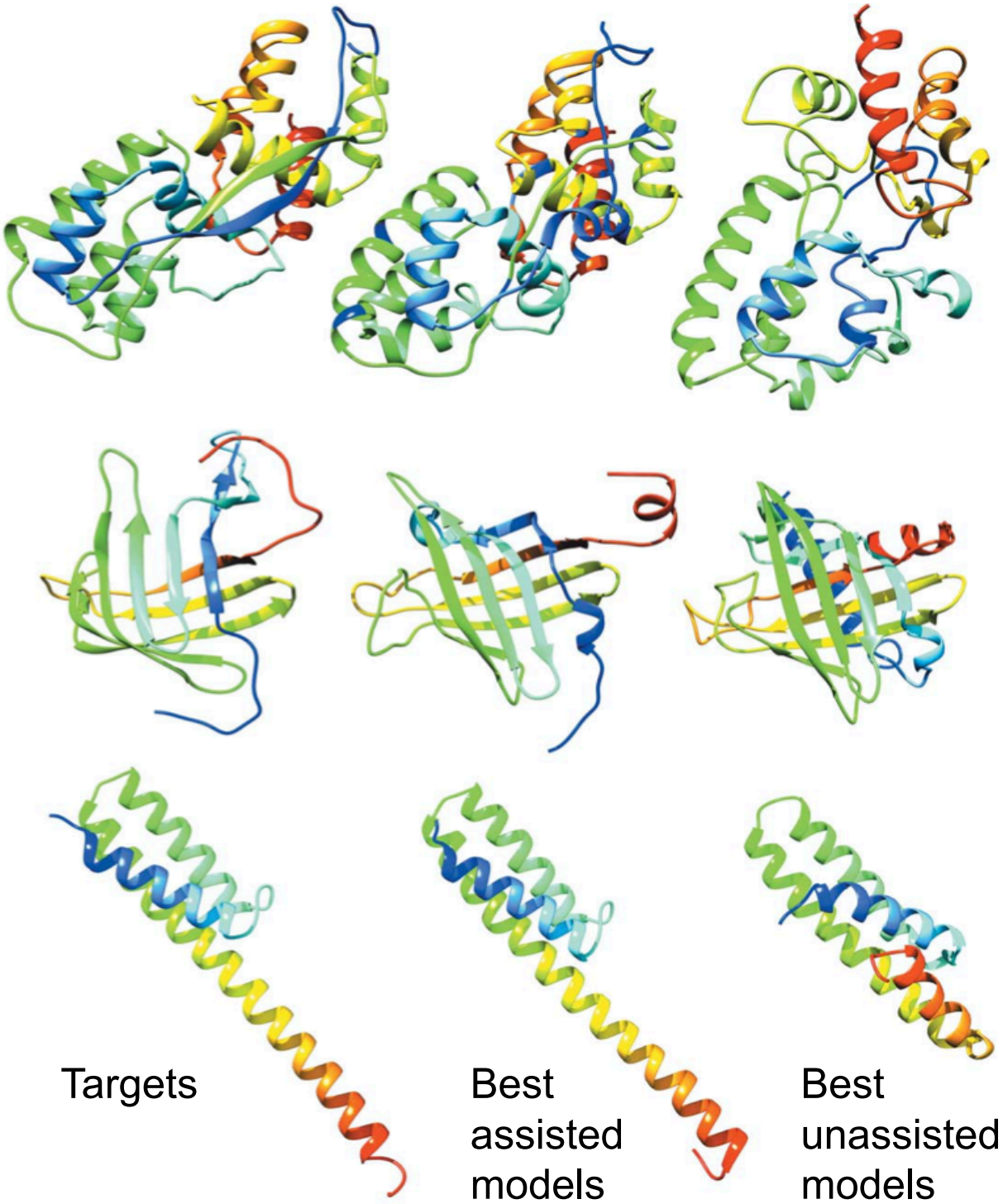
- Use the methods we have discussed, but in addition, information on residues that are close together in 3D space



Taylor *et al. Proteins* **82**, 84 (2014)

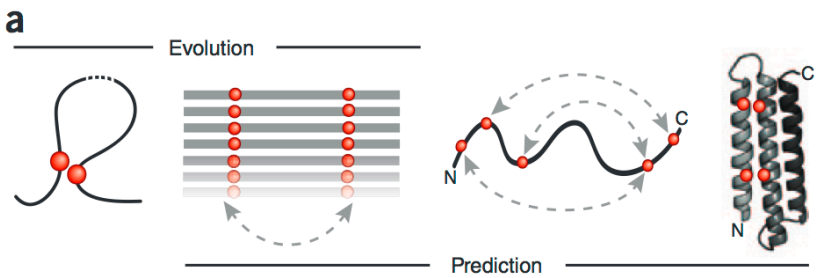
Contact assisted methods

Taylor *et al. Proteins* **82**, 84 (2014)

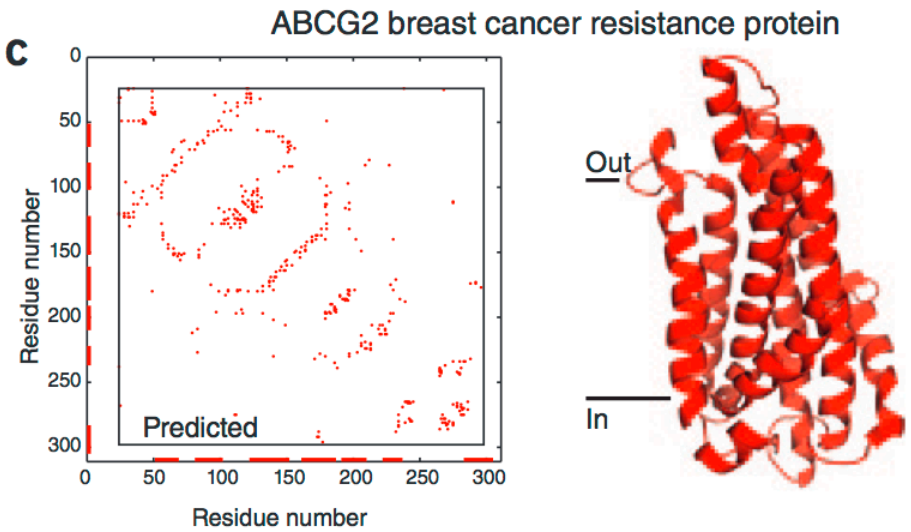


Contact assisted methods

- Co-evolution/co-variation from MSAs



Marks *et al. Nat. Biotech.* **30**, 1072 (2012)



- Experimental methods
  - NMR
  - Crosslinkers & MS
- Promising field!

End

jonkl@medisin.uio.no