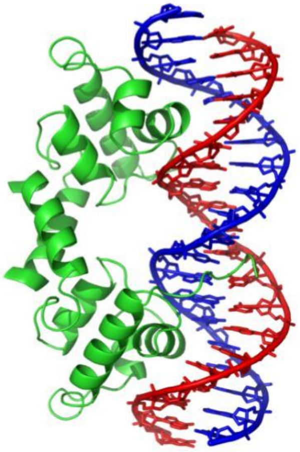
Improving fragment assembly protein structure prediction

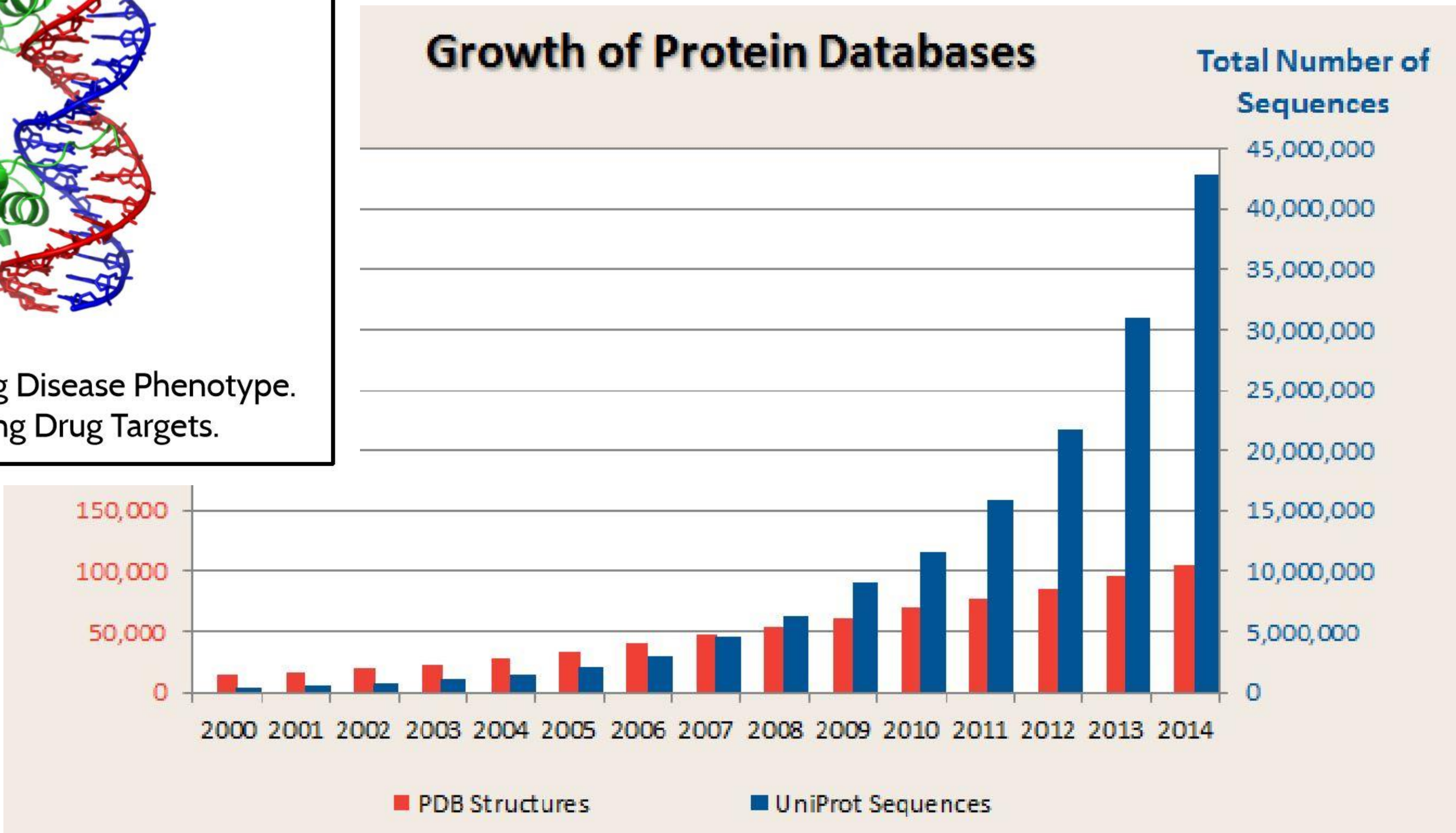
Charlotte Deane
Department of Statistics
Oxford University

Why predict protein structures?

Functional characterization



-Understanding Disease Phenotype.
-Identifying Drug Targets.

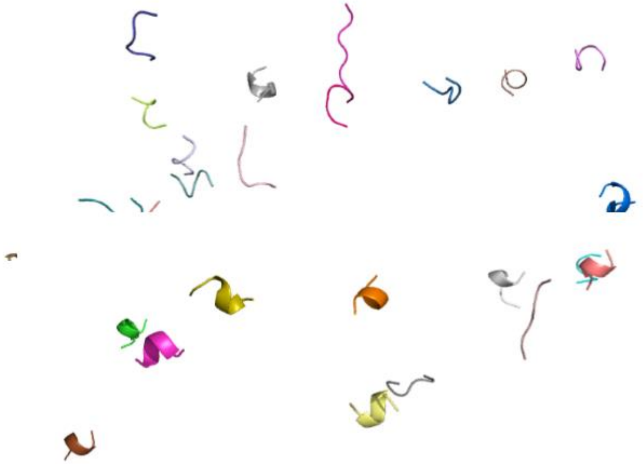


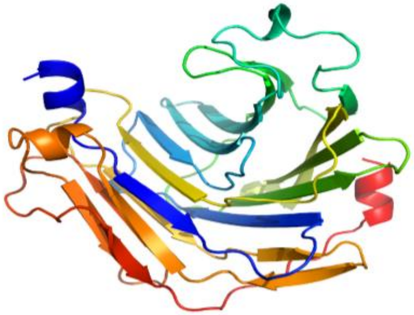
Structure prediction methods

- Template-based methods:
  - Comparative modelling (or Homology modelling):
    - There exists a protein with clear homology.
    - Uses sequence-based techniques to identify a template. – Protein Threading/Fold recognition:
    - There exists a protein of similar fold (analogy).
- Template-free methods:
  - Novel fold prediction

Fragment assembly – Protein structure prediction

RPRTAFSSEQLARLKREFNENR
YLTERRRQQLSSELGLNEAQIKI
WFQNKRAKI





Fragment assembly – Protein structure prediction

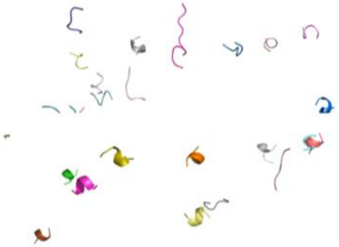
Where for any given position, there are multiple pieces that can fit in it...

Where the pieces got mixed up with pieces from another puzzle...

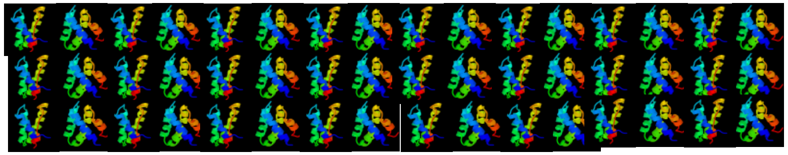
Where some pieces are missing...

And where you cannot look at the box to check how it is supposed to look like...

How does it work?



- Energy function
  - Usually from a Bayesian treatment of residue distributions in known protein structures sometimes combined with physics based energy terms
  - Pair potential terms, Solvation potentials terms, Steric terms, Long-range hydrogen bonding, compactness term
  - Predicted contacts from co-evolution methods
- Use a Monte Carlo search procedure
  - Move set based on fragments of protein structures
- Generate thousands of decoys

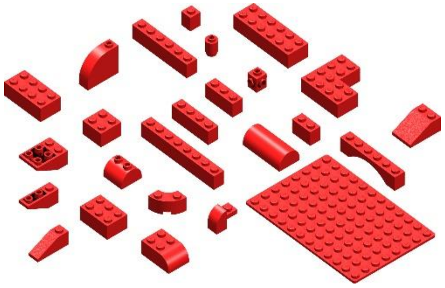


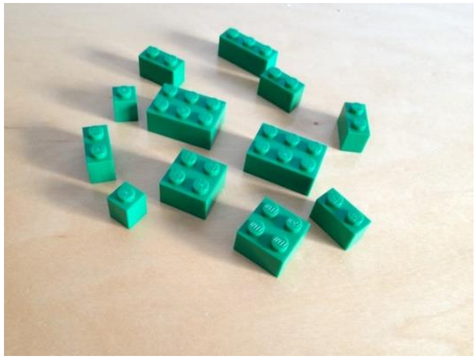
- Select a final answer

Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library

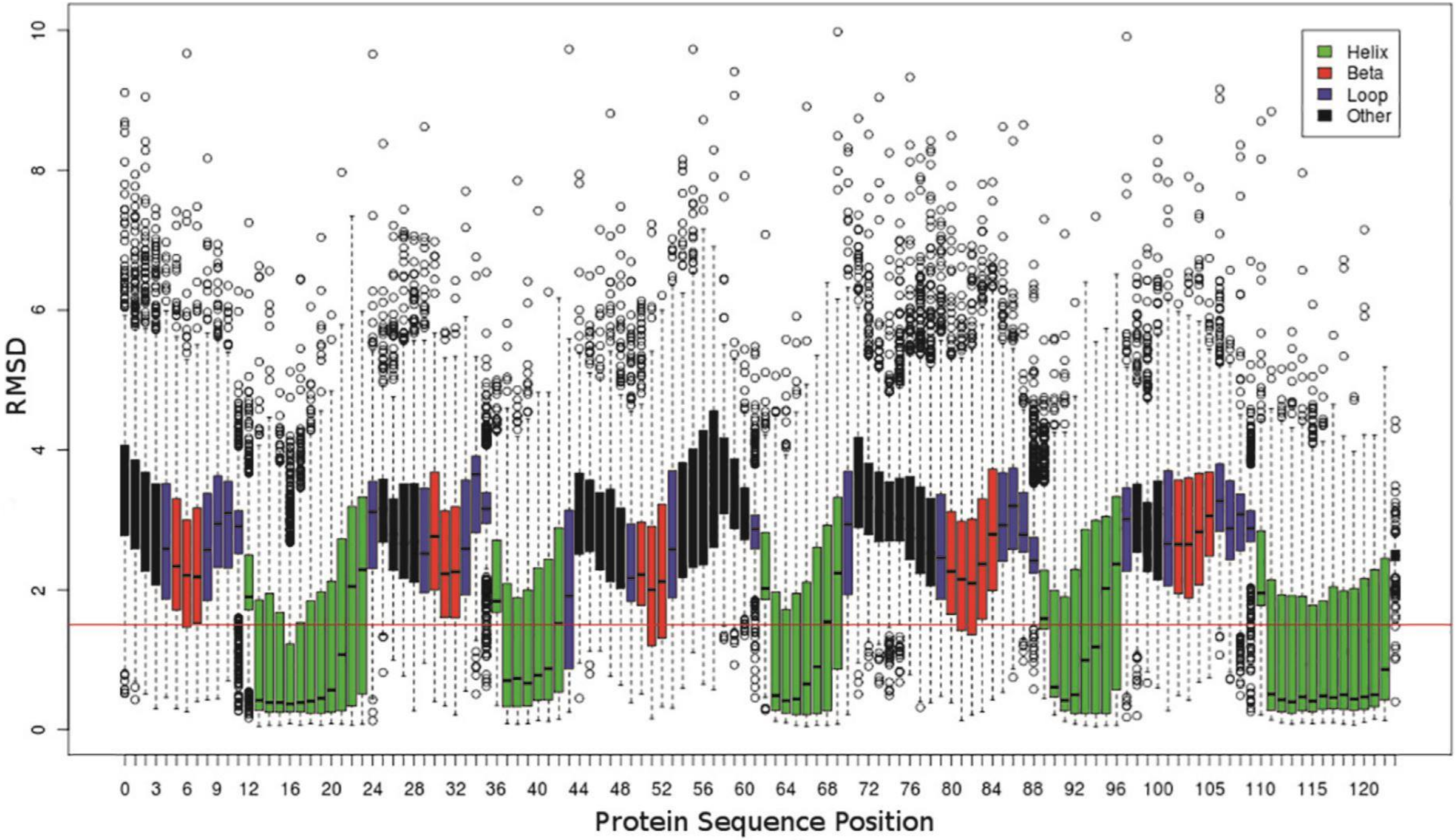
DREFGWTYPACDEFLMNGHIKLMNPQRSTVWY………………





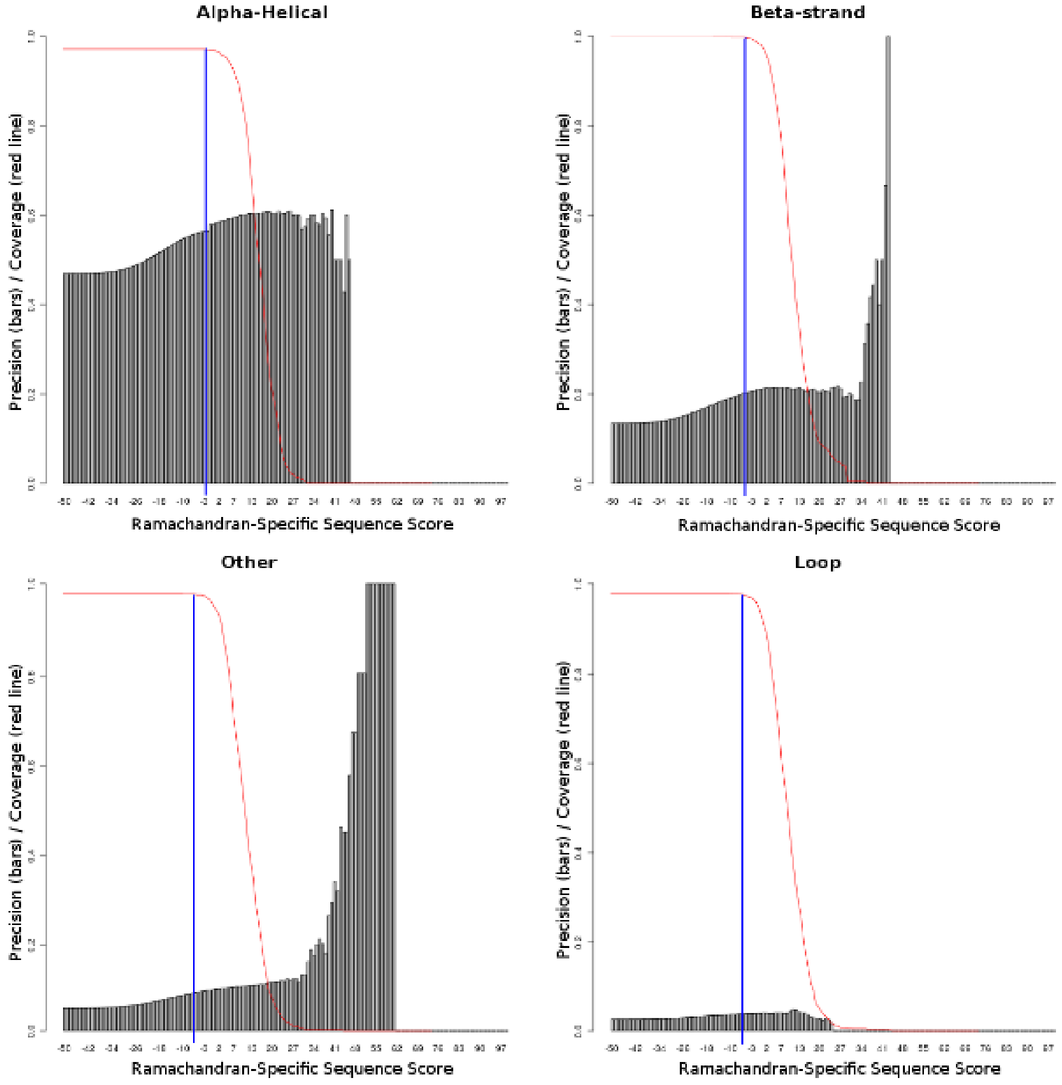
Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library



Ways to improve Fragment assembly

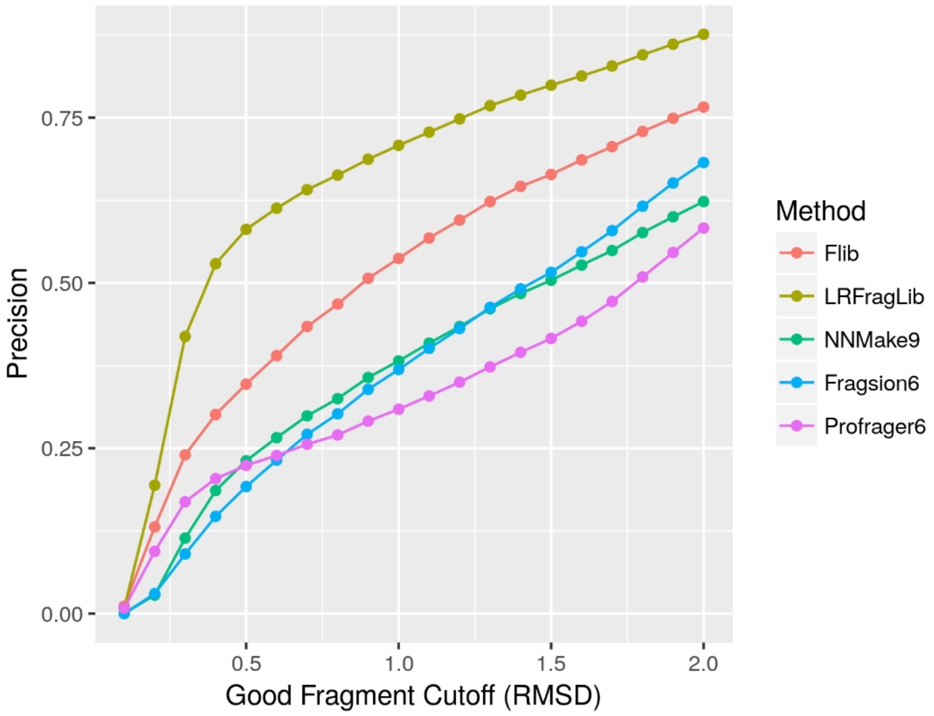
- Consider secondary structure when assessing your fragment library

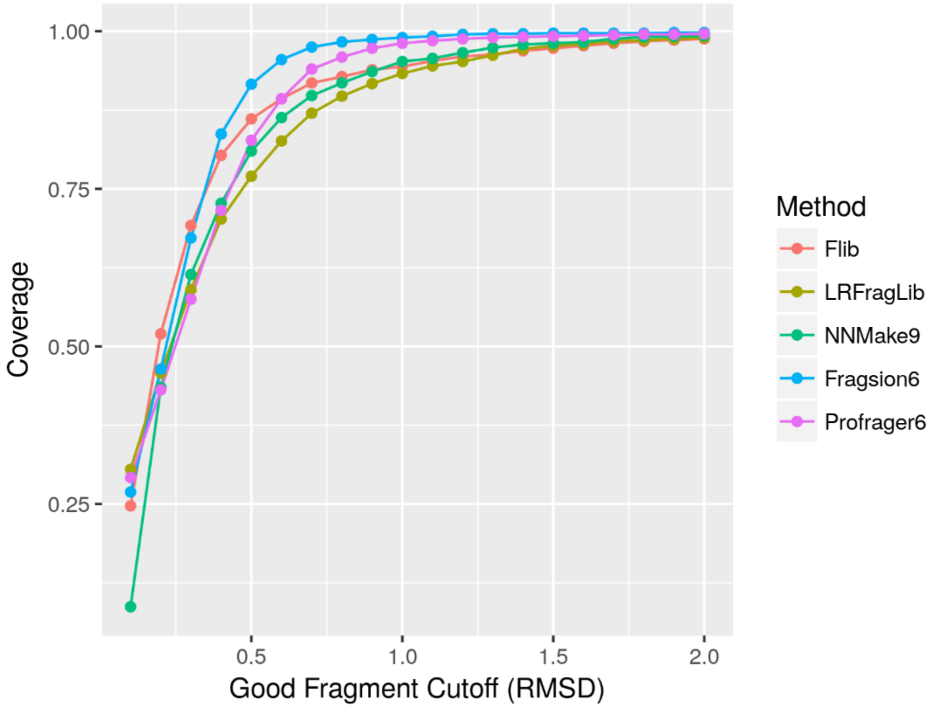


Oliveira et al Plos One (2015)

Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library

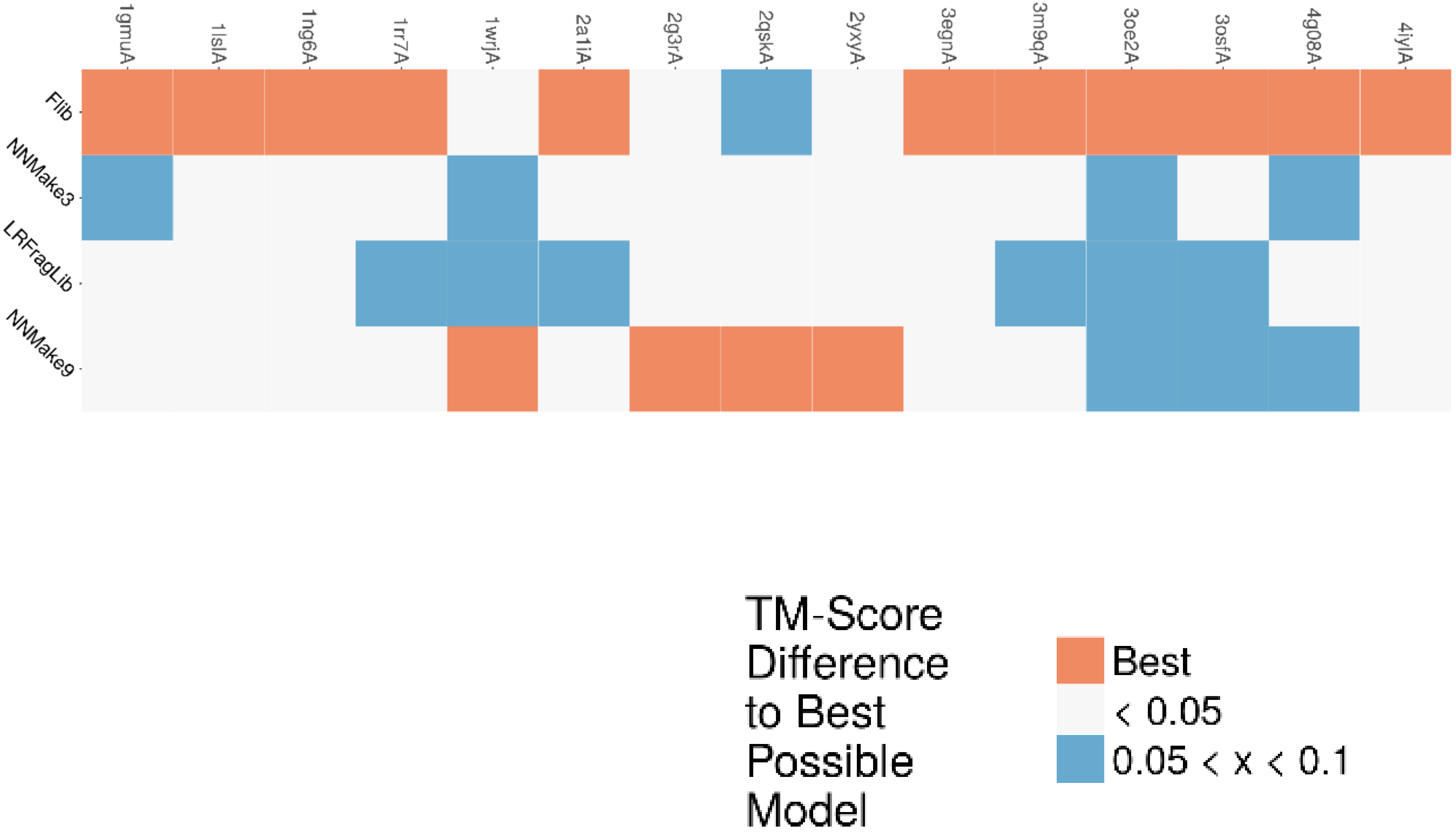




NNMAKE – Gront et al (2011)
FLIB – Oliveira et al (2015)
LRFragLib – Wang et al (2016)
Fragsion – Bhattacharya et al (2016)
Profrager – Santos et al (2015)

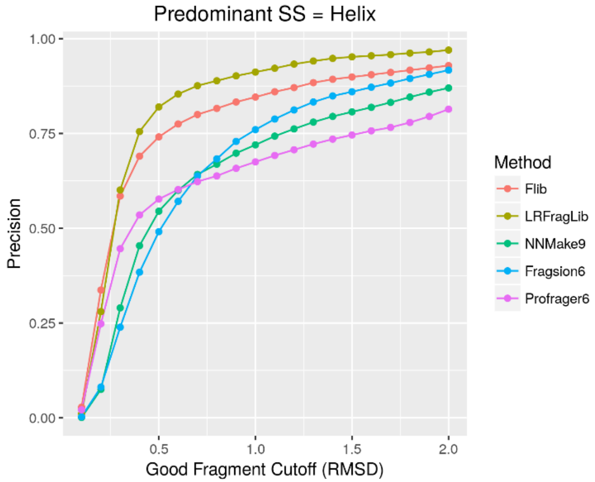
Ways to improve Fragment assembly

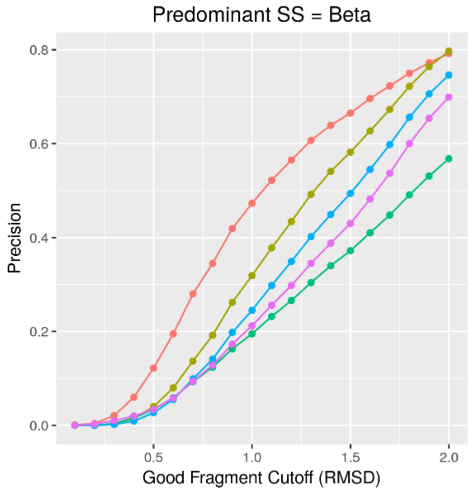
- Consider secondary structure when assessing your fragment library

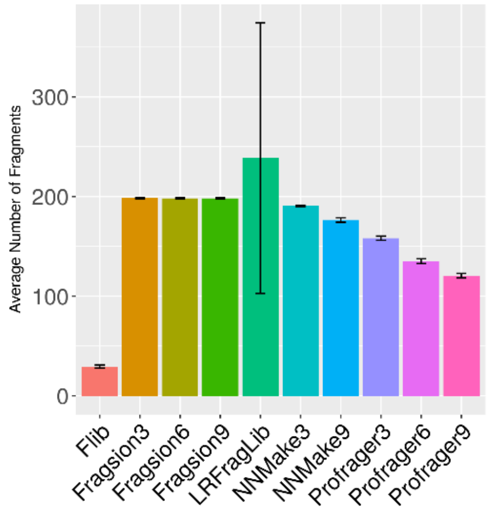


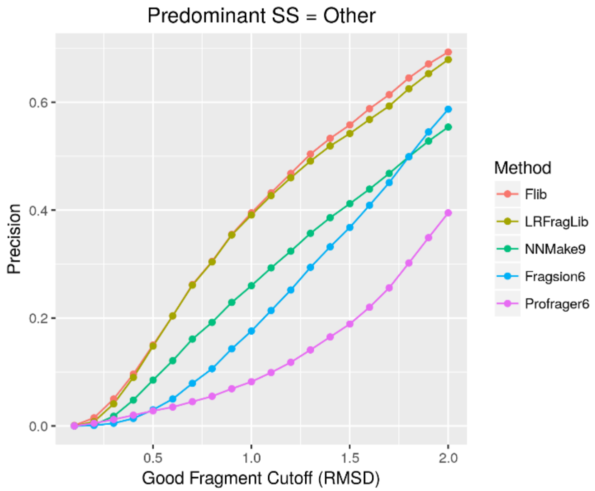
Ways to improve Fragment assembly

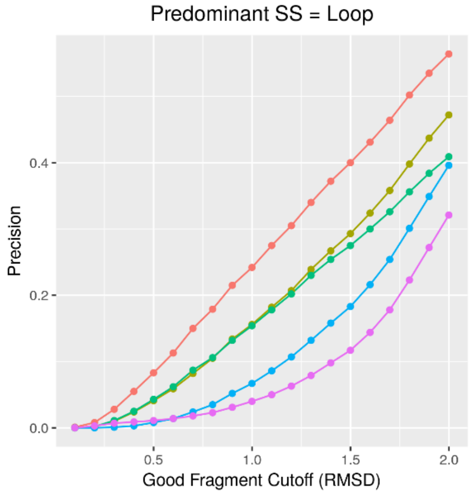
- Consider secondary structure when assessing your fragment library

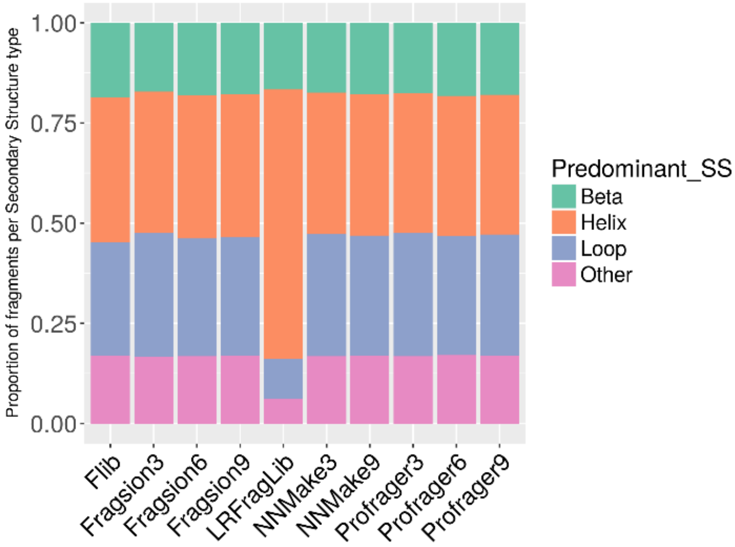

Predominant SS = Helix


Predominant SS = Beta



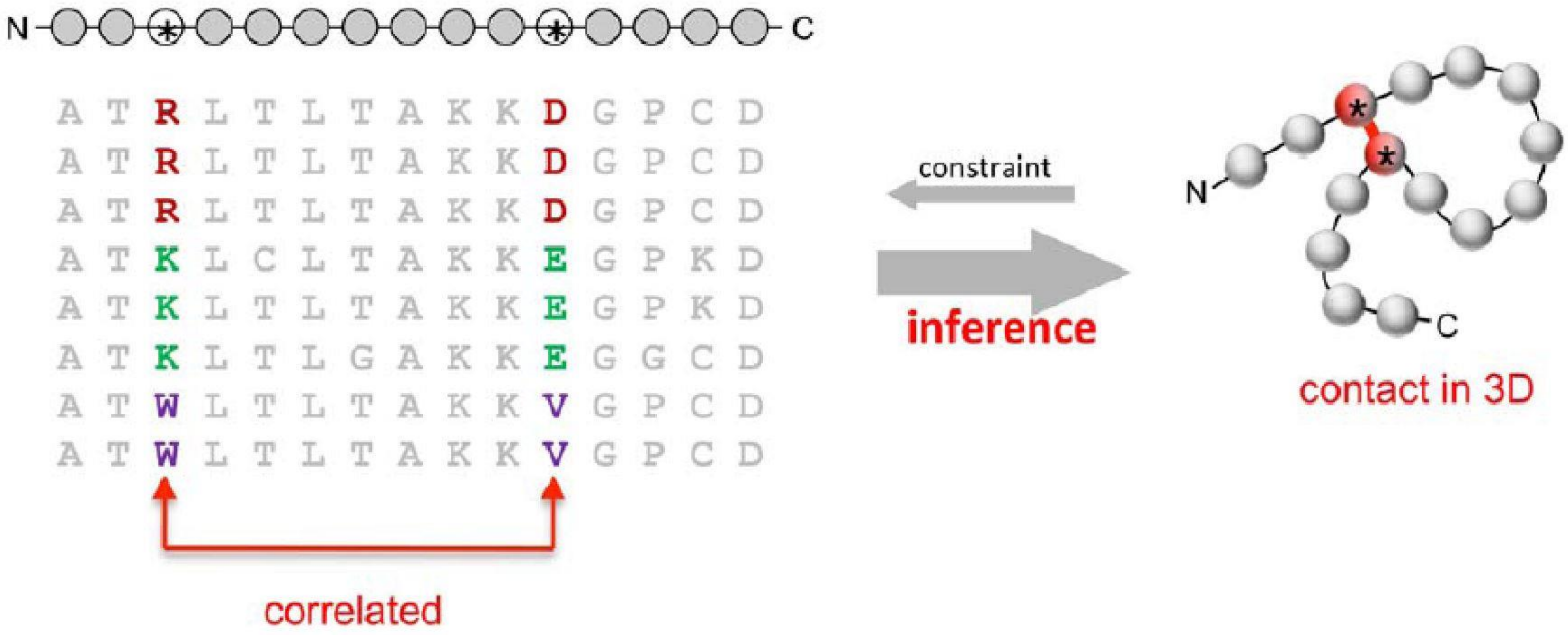

Predominant SS = Other


Predominant SS = Loop



Ways to improve Fragment assembly

- Use contact predictions



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Two residues that mutate in a correlated fashion (co-evolve) are inferred to share spatial proximity.

Oliveira et al Bioinformatics (2016)

Improving co-evolution contact prediction

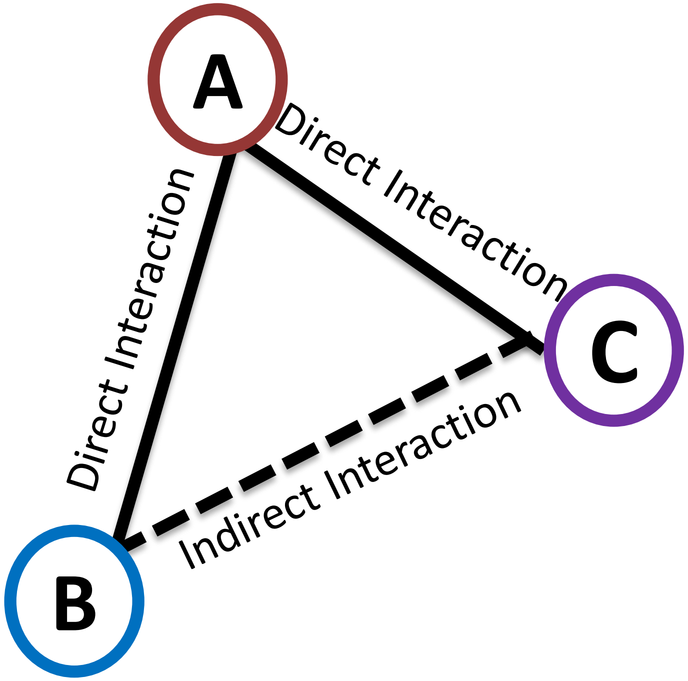
Correlation in amino acid substitution may arise from direct as well as indirect interactions.

Need to use the information of all columns in the multiple sequence alignment when ascertaining the correlation between two individual columns

Mean Field Direct Coupling Analysis

Estimate the inverse covariance matrix to assign a score to residue pairs

Learn the direct couplings as parameters of a Probabilistic Graphical Model (Markov random field) by maximizing its pseudo-likelihood.



Methods

- Test set - 3458 proteins

• FreeContact	Kajan,L. et al. (2014)
• PSICOV	Jones,D.T. et al. (2012)
• CCMPred	Seemayer,S. et al. (2014)
• Bbcontacts	Andreani and Soding (2015)
• metaPSICOV stage 1	Jones,D.T. et al. (2014)
• metaPSICOV stage2	Jones,D.T. et al. (2014)
• metaPSICOV HB	Jones,D.T. et al. (2014)
• GREMLIN	Kamisetty et al. (2013)

Oliveira et al (2016)

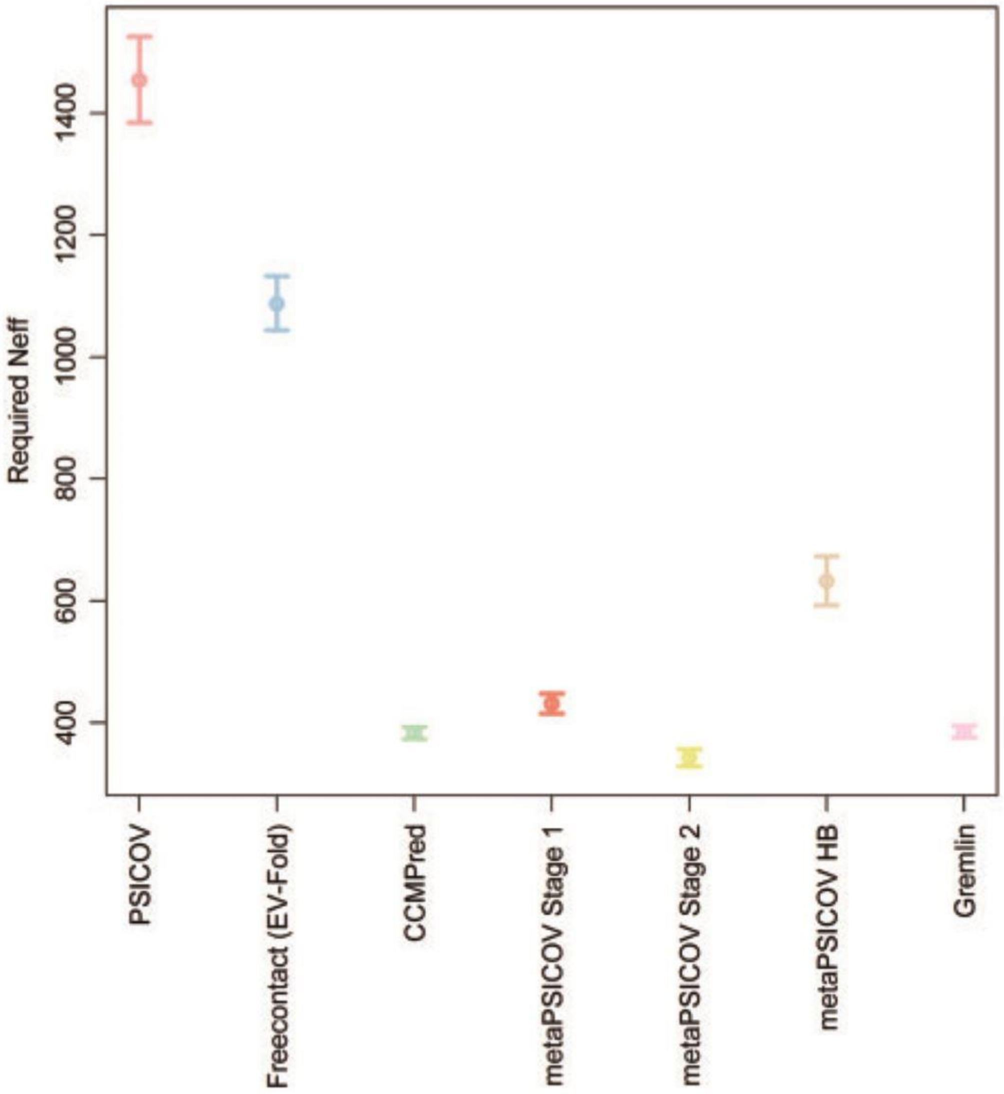
Contact definition

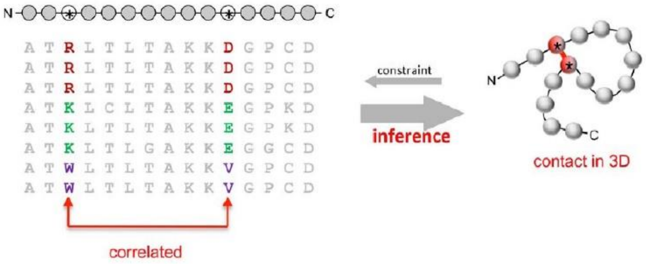
- Two protein residues are defined to be in contact if their C-βs (C-αs for Glycine) are less than 8 A apart

- Contacts between residues being less than five residues apart and are not considered
- A short-range contact between residues i and j is defined when $5 \leq |i - j| \geq 23$.
- A long range contact is defined when $|i - j| > 23$

Jones et al (2012)
Marks et al (2011)

How many sequences do you need in the multiple sequence alignment?

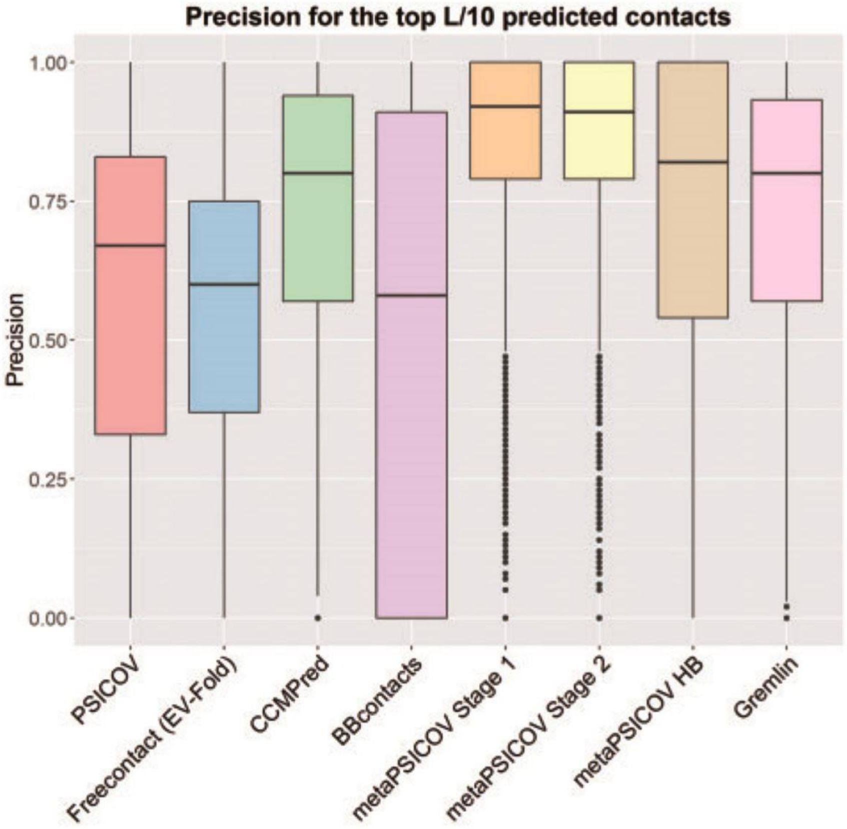




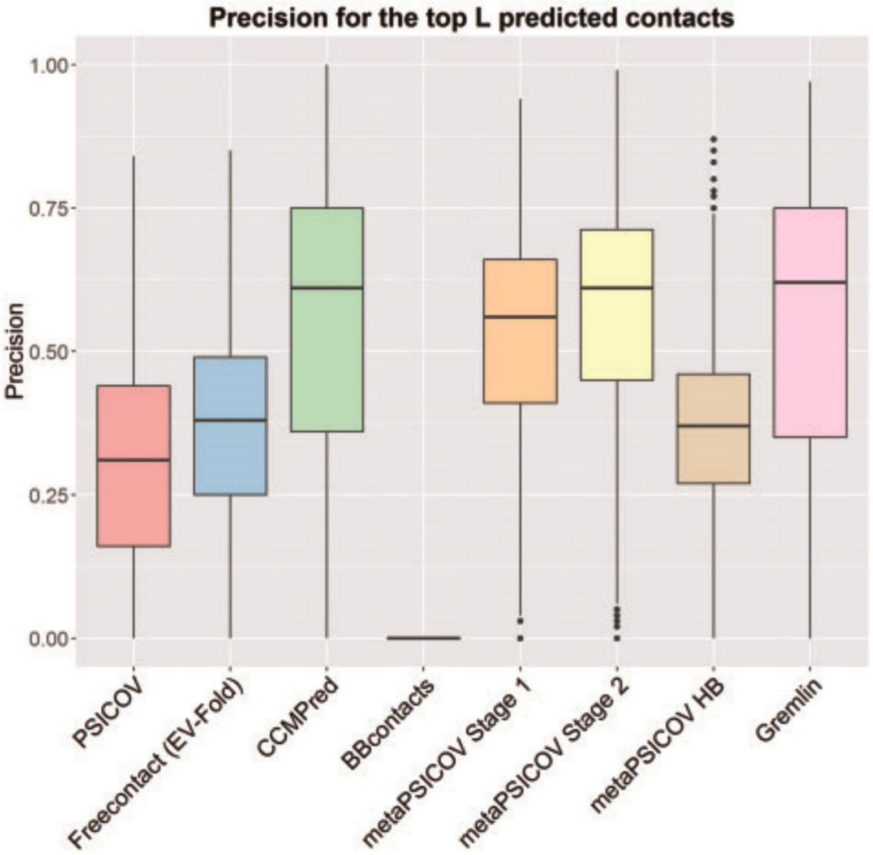
```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Oliveira et al Bioinformatics (2016)

How accurate are the methods?



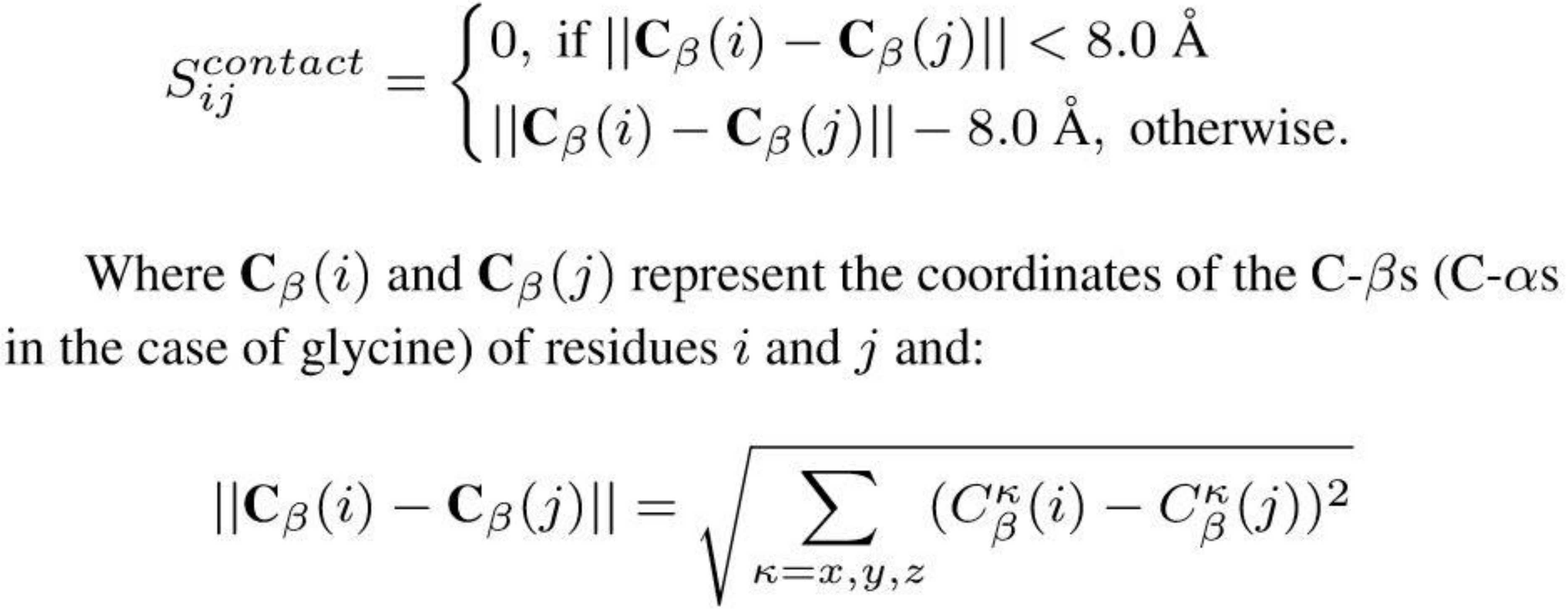
Precision for the top L/10 predicted contacts



Precision for the top L predicted contacts

Oliveira et al Bioinformatics (2016)

Putting co-evolutionary contacts into protein structure prediction

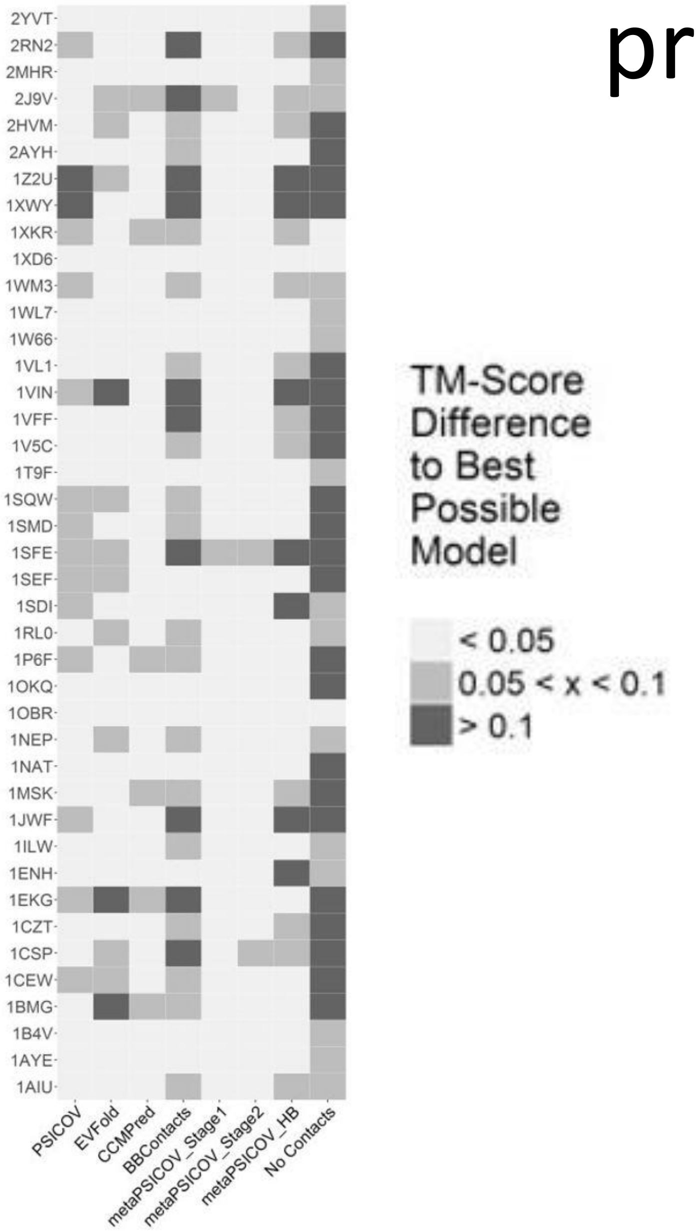


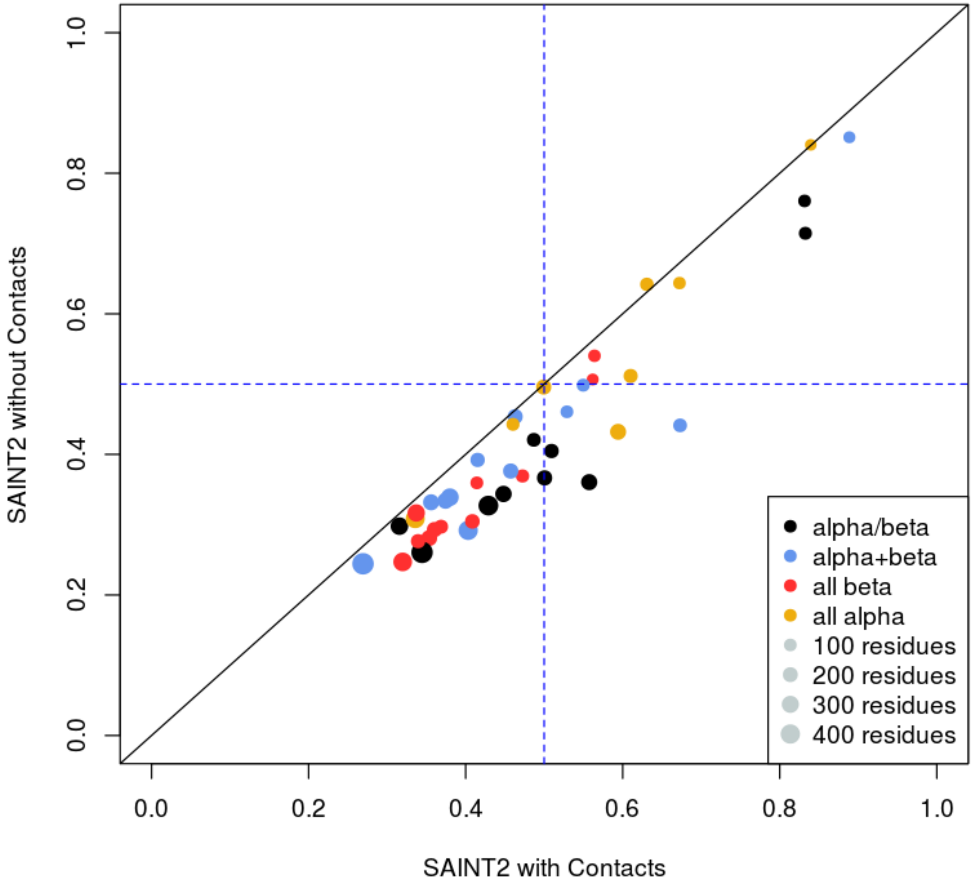
$$S_{ij}^{contact} = \begin{cases} 0, \text{ if } ||\mathbf{C}_\beta(i) - \mathbf{C}_\beta(j)|| < 8.0 \text{ Å} \\ ||\mathbf{C}_\beta(i) - \mathbf{C}_\beta(j)|| - 8.0 \text{ Å}, \text{ otherwise.} \end{cases}$$

Where $\mathbf{C}_\beta(i)$ and $\mathbf{C}_\beta(j)$ represent the coordinates of the C-βs (C-αs in the case of glycine) of residues i and j and:

$$||\mathbf{C}_\beta(i) - \mathbf{C}_\beta(j)|| = \sqrt{\sum_{\kappa=x,y,z} (C_\beta^\kappa(i) - C_\beta^\kappa(j))^2}$$

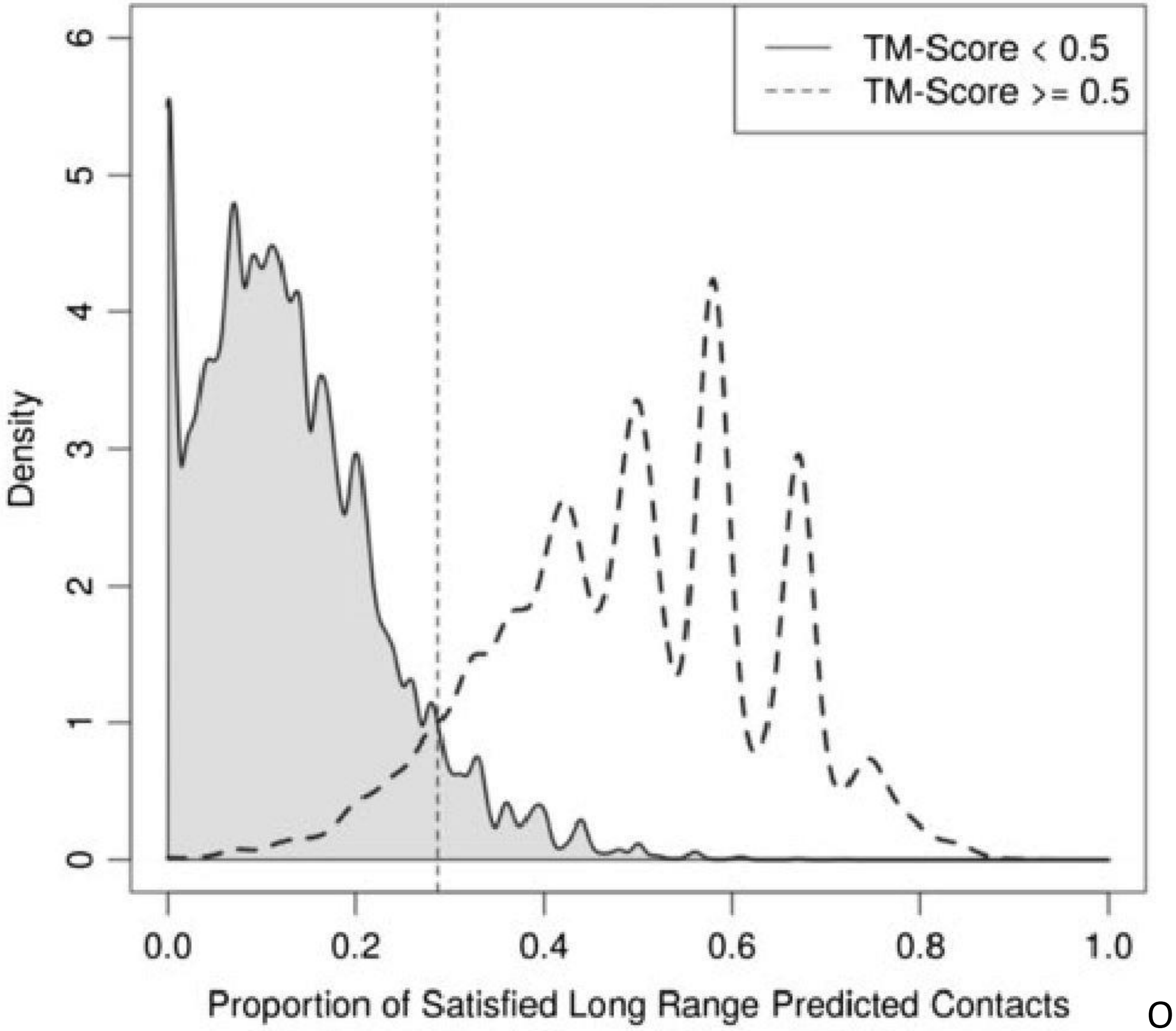
How do they influence structure prediction?





Oliveira et al Bioinformatics (2016)

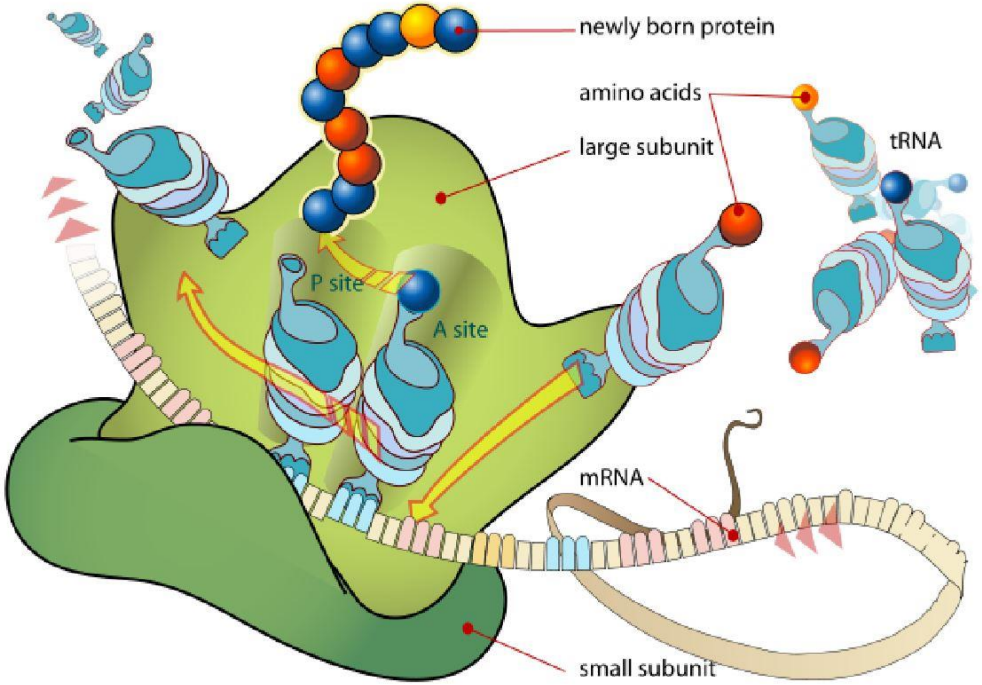
Using co-evolution contacts to identify good models



Oliveira et al Bioinformatics (2016)

Ways to improve Fragment assembly

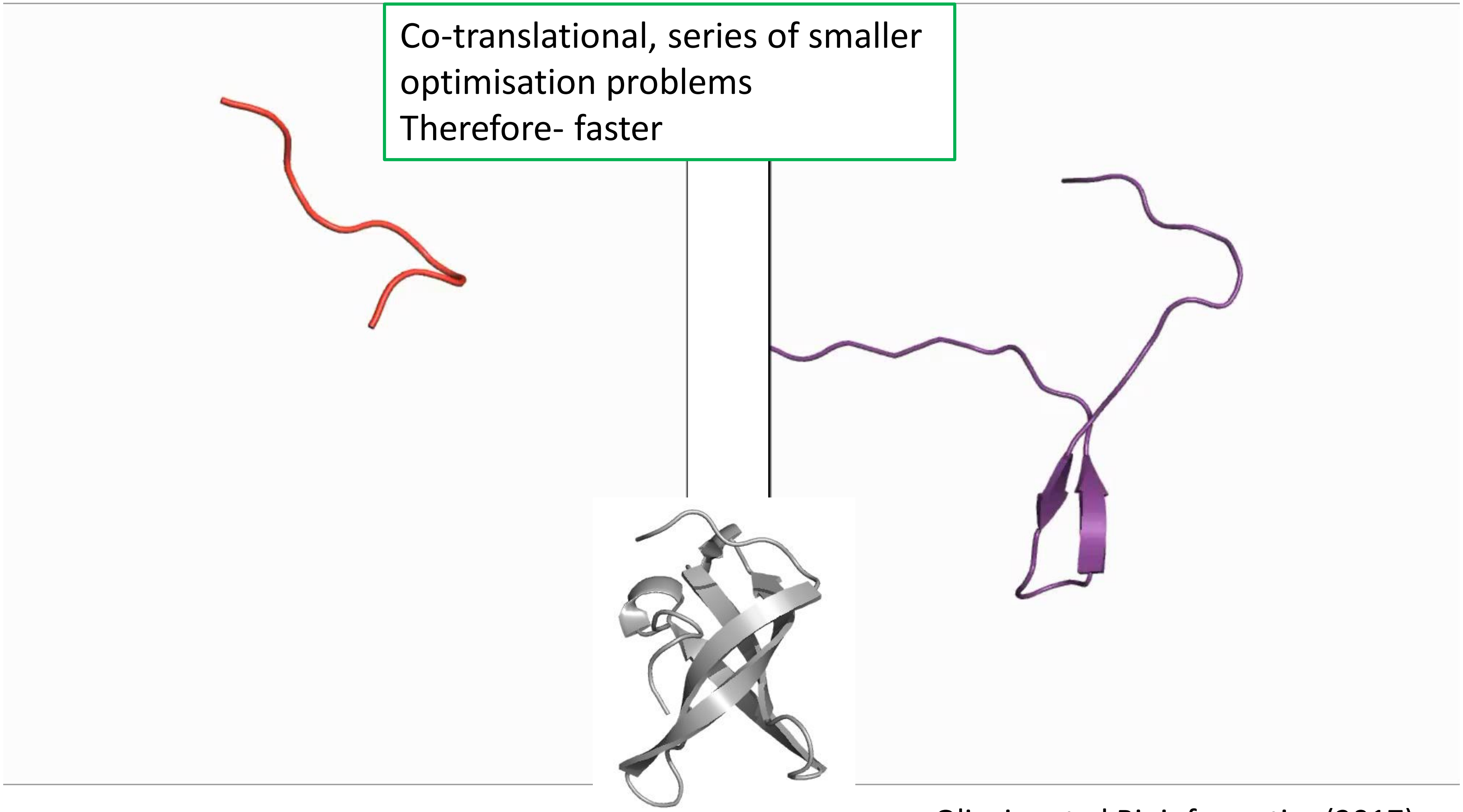
- Improve your search strategy



There is a hypothesis that proteins begin to fold as they are being synthesized. This is known as cotranslational protein folding.

Oliveira et al Bioinformatics (2017)

Improving the search: Cotranslational protein structure prediction



Co-translational, series of smaller optimisation problems
Therefore- faster

Oliveira et al Bioinformatics (2017)

Number of decoys required

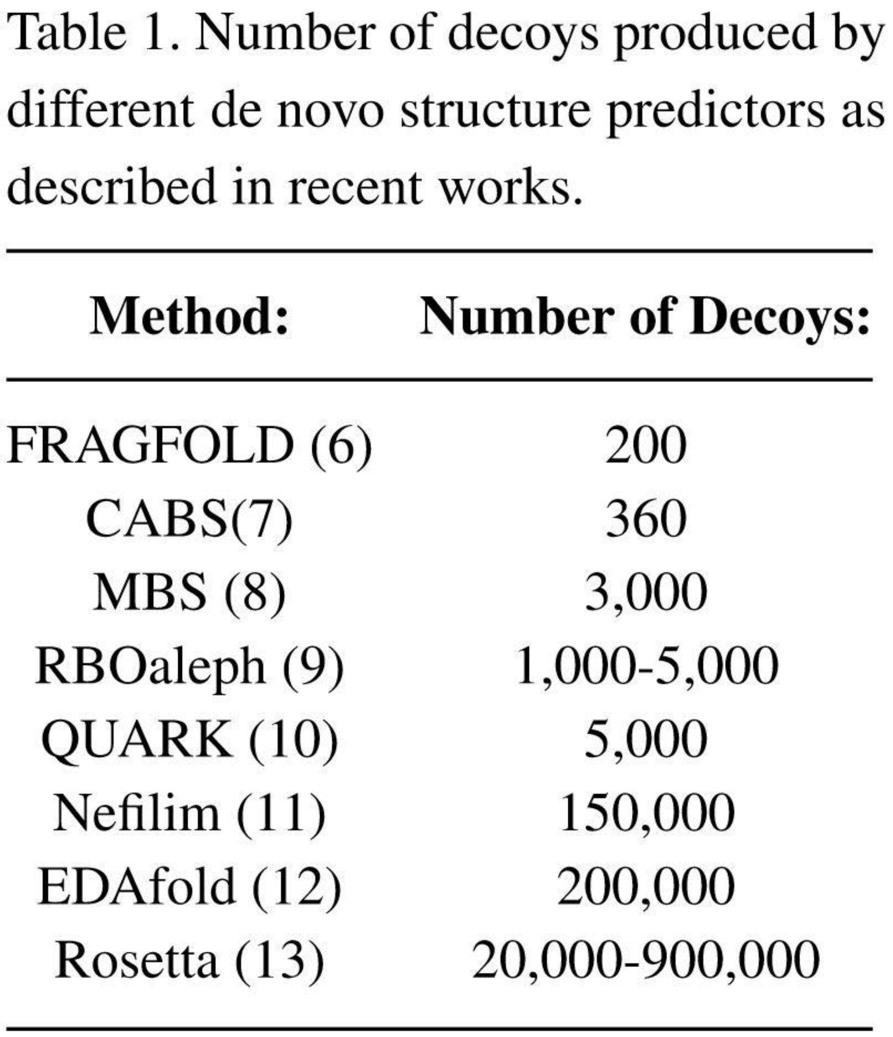
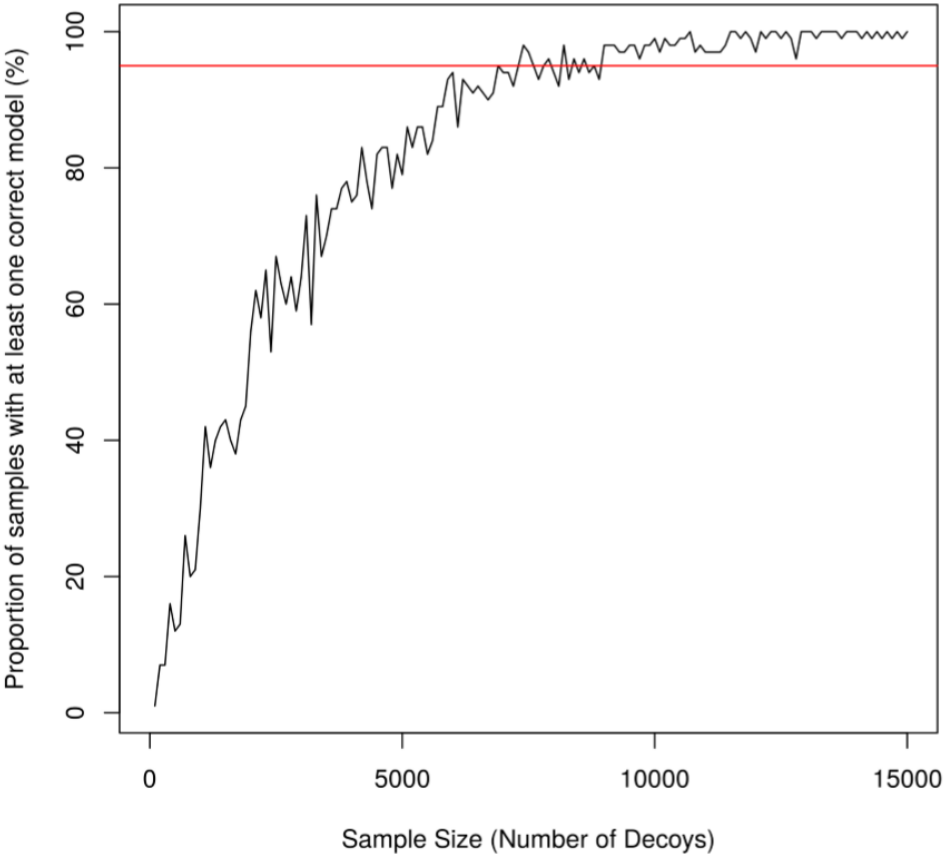


Table 1. Number of decoys produced by different de novo structure predictors as described in recent works.

Method:	Number of Decoys:
FRAGFOLD (6)	200
CABS(7)	360
MBS (8)	3,000
RBOaleph (9)	1,000-5,000
QUARK (10)	5,000
Nefilim (11)	150,000
EDAfold (12)	200,000
Rosetta (13)	20,000-900,000

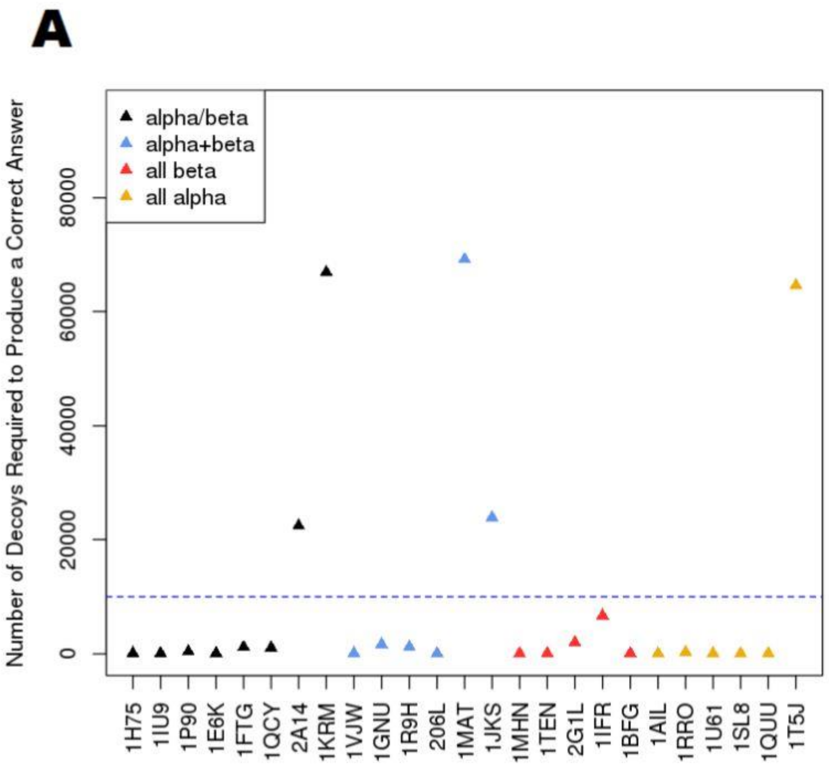
Oliveira et al Bioinformatics (2017)

Number of decoys required



Oliveira et al Bioinformatics (2017)

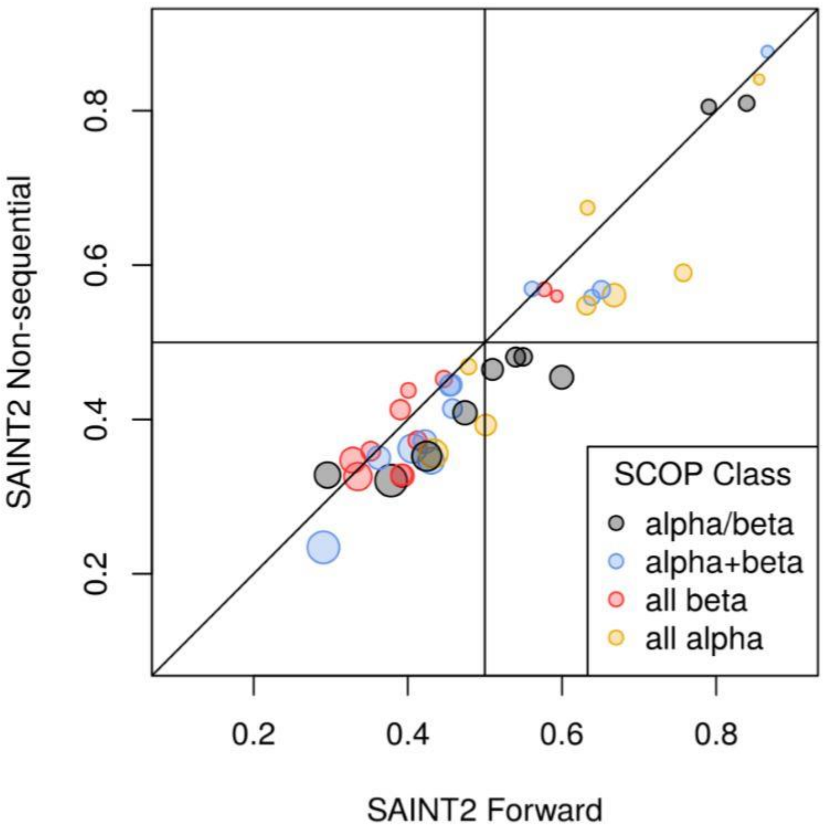
Number of decoys required

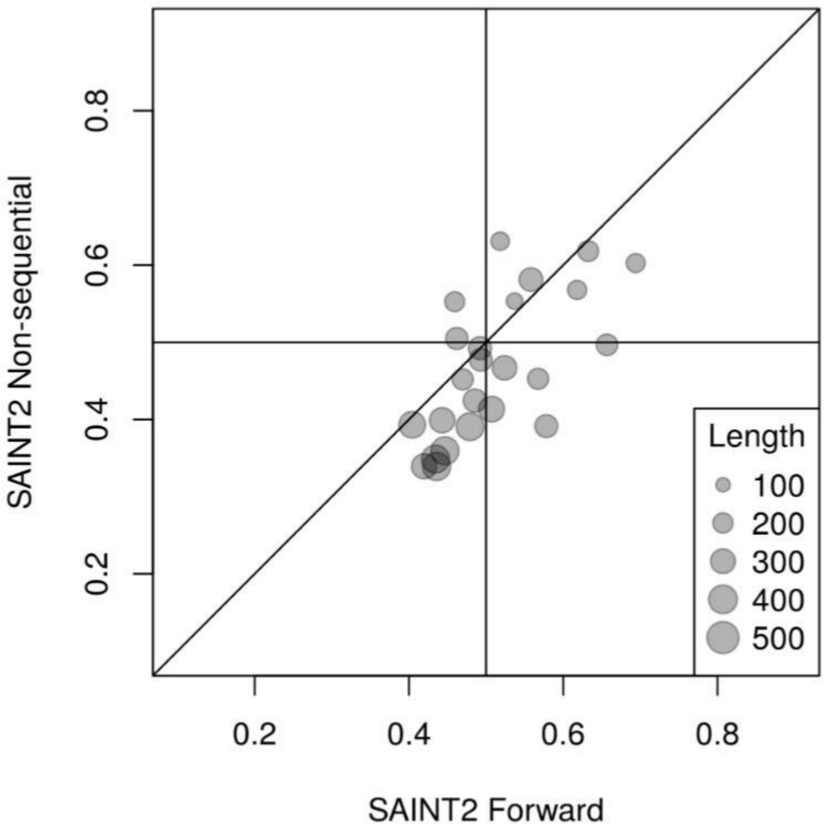


- Number decoys to get a correct answer ~10,000
- Number of decoys to get best answer ~20,000
- Not dependent on protein length (if length <250)

Oliveira et al Bioinformatics (2017)

SAINT2 Cotranslational in action





Oliveira et al Bioinformatics (2017)

Improving the search: Cotranslational protein structure prediction

- Most current de novo structure prediction methods randomly sample protein conformations
  - Require large amounts of computational resource

- SAINT2 uses a sequential sampling strategy, suggested by biology
  - SAINT2 requires ~10,000 decoys to produce a good answer fewer than most other methods suggest

- Sequential sampling improves speed
  - 1.5 to 2.5 times faster than non-sequential prediction.

- SAINT2 sequential produces better models

- SAINT2 sequential a pseudo-greedy search strategy that reduces computational time of de novo protein structure prediction and improves accuracy

Oliveira et al Bioinformatics (2017)

ACKNOWLEDGEMENTS

evotec

Roche

NOVARTIS
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH

MedImmune
A member of the AstraZeneca Group

SGC

MRC Medical Research Council

diamond

gsk GlaxoSmithKline

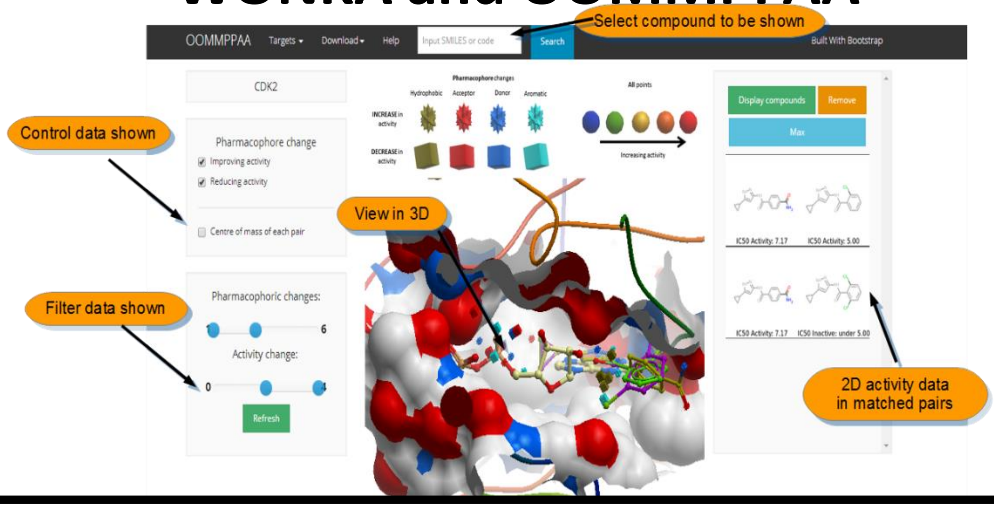
EPSRC
Pioneering research and skills

ucb

e-Therapeutics plc

BBSRC
bioscience for the future

WONKA and OOMMPPAA



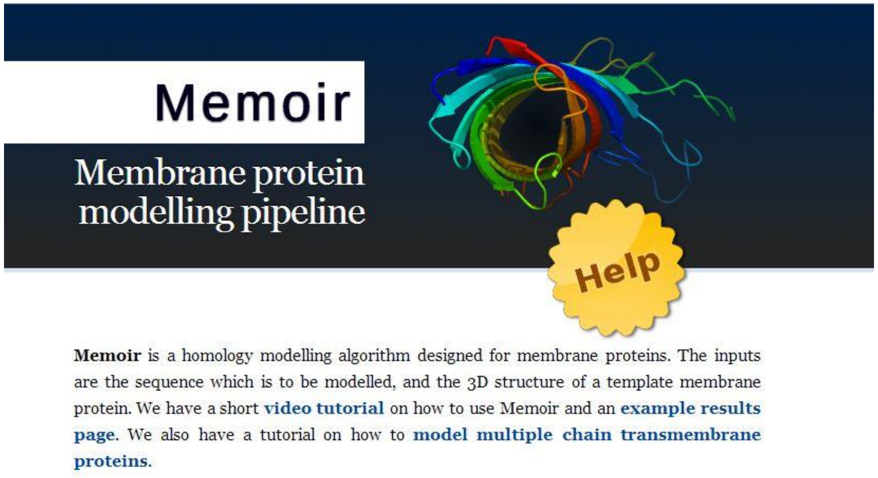
Select compound to be shown

Control data shown

Filter data shown

View in 3D

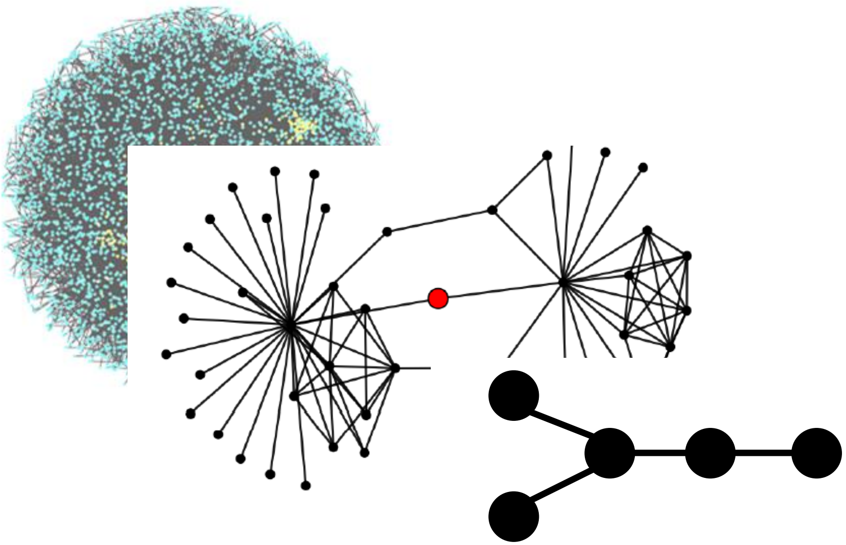
2D activity data in matched pairs

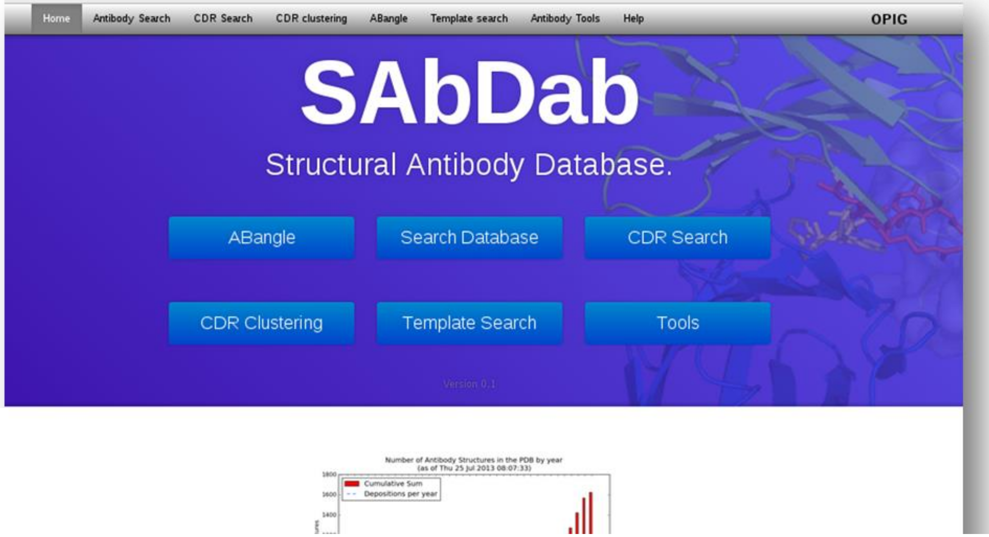


Memoir is a homology modelling algorithm designed for membrane proteins. The inputs are the sequence which is to be modelled, and the 3D structure of a template membrane protein. We have a short **video tutorial** on how to use Memoir and an **example results page**. We also have a tutorial on how to **model multiple chain transmembrane proteins**.

http://www.stats.ox.ac.uk/proteins/resources

NetEMD




Home
Antibody Search
CDR Search
CDR clustering
ABangle
Template search
Antibody Tools
Help
OPIG
SAbDab
Structural Antibody Database.
ABangle
Search Database
CDR Search
CDR Clustering
Template Search
Tools
Version 0.1
Number of Antibody Structures in the PDB by year
(as of Thu 25 Jul 2013 08:07:33)
Cumulative Sum
Depositions per year
1800
1600
1400