

Biologically inspired de novo protein structure prediction

Charlotte Deane
Department of Statistics
Oxford University

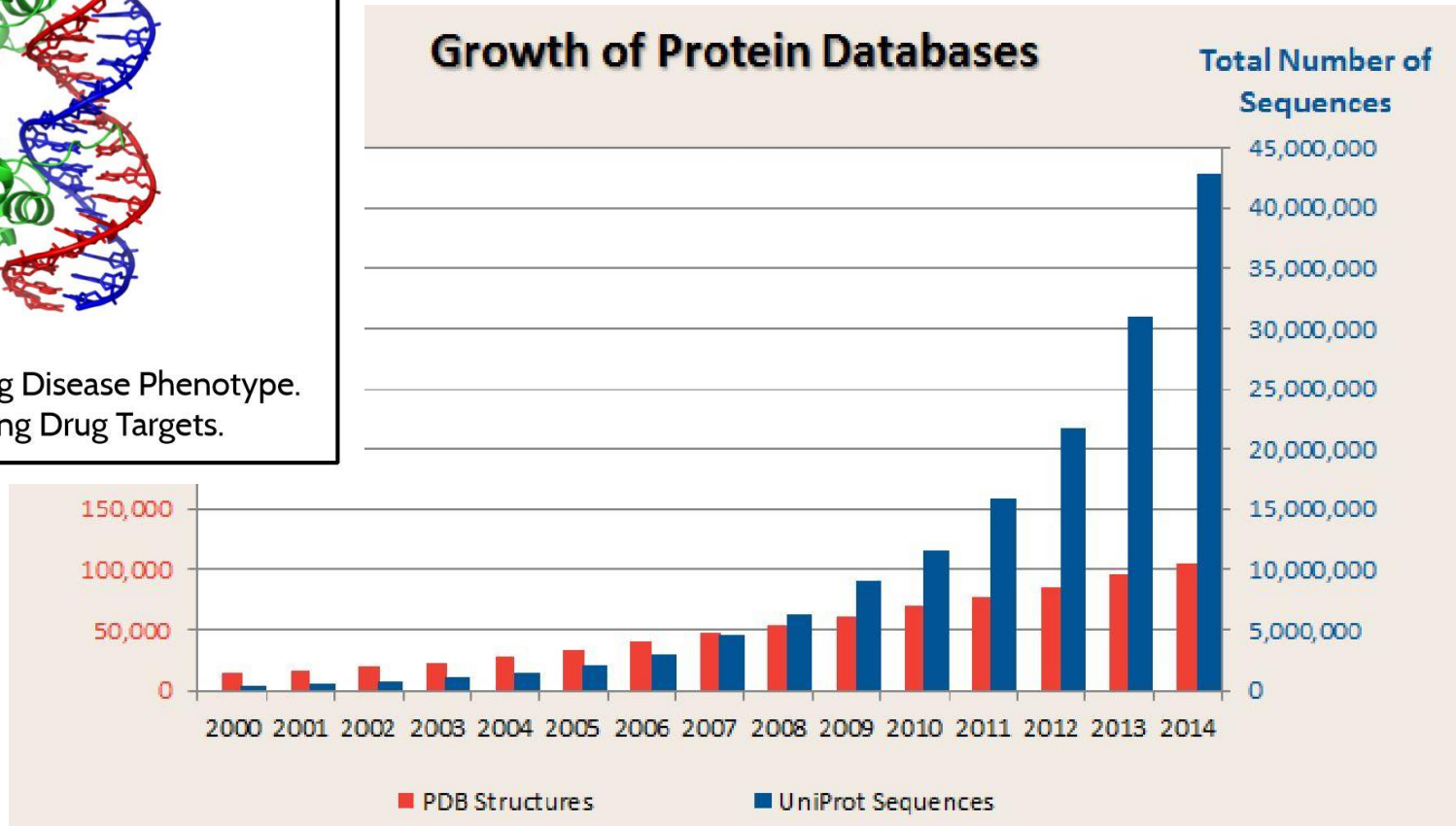
Why predict protein structures?

Functional characterization



- Understanding Disease Phenotype.
- Identifying Drug Targets.

Growth of Protein Databases

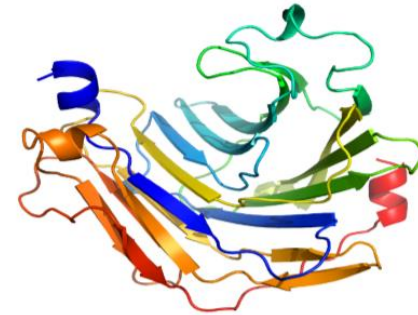
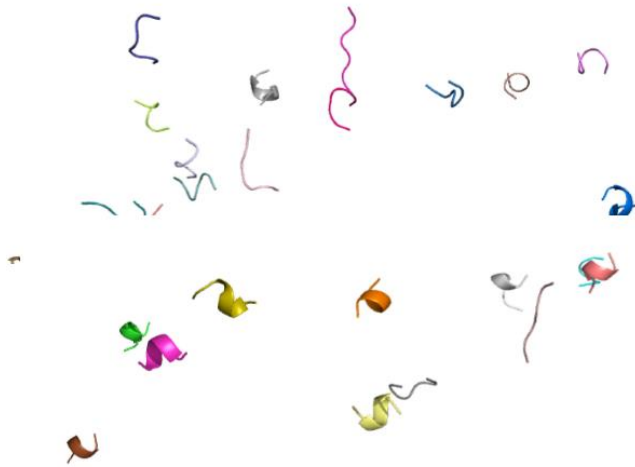


Structure prediction methods

- 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...



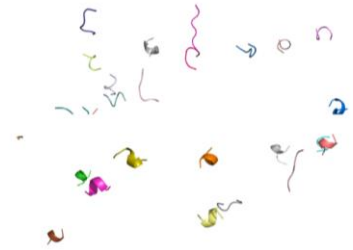
Where the pieces got mixed up with pieces from another puzzle...

Where some pieces are missing...

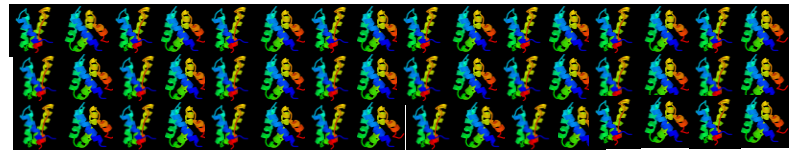


And where you cannot look at the box to check how it is supposed to look like...

How does it work?



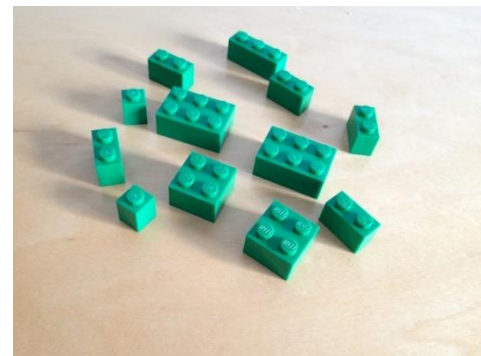
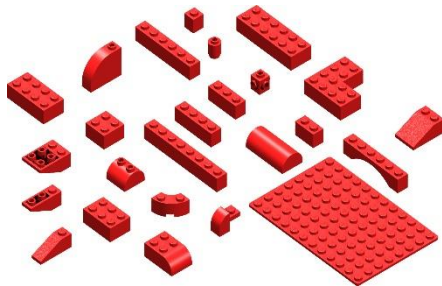
- Energy function
 - Usually from a Bayesian treatment of residue distributions in known protein structures sometimes combined with physics based energy terms
 - Pair potential terms, Solvation potentials terms, Steric terms, Long-range hydrogen bonding, compactness term
 - Predicted contacts from co-evolution 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

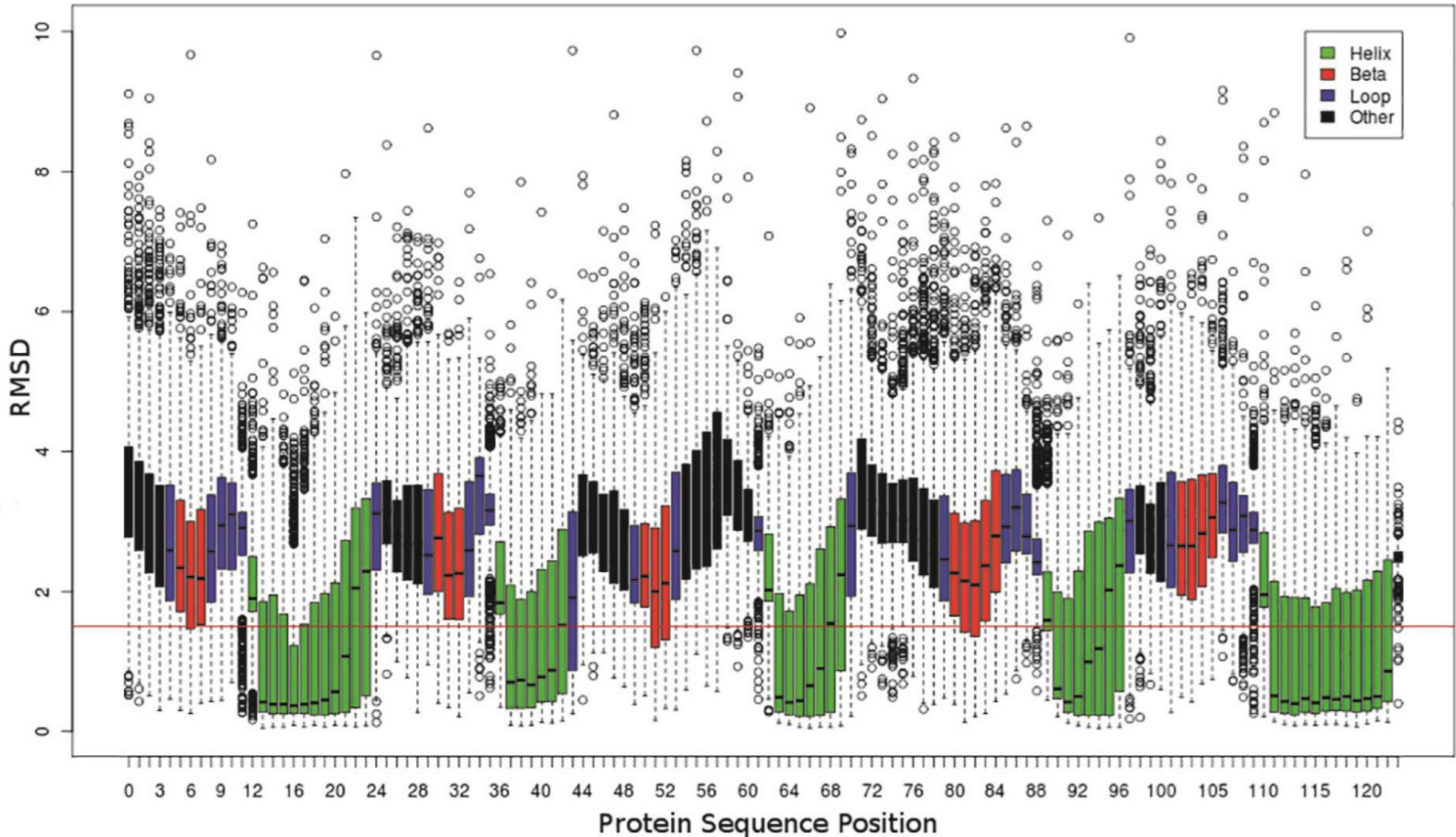
- Consider secondary structure when assessing your fragment library

DREFGW|TY|PACDEF|LMNGHIKLMNPQRSTVWY.....



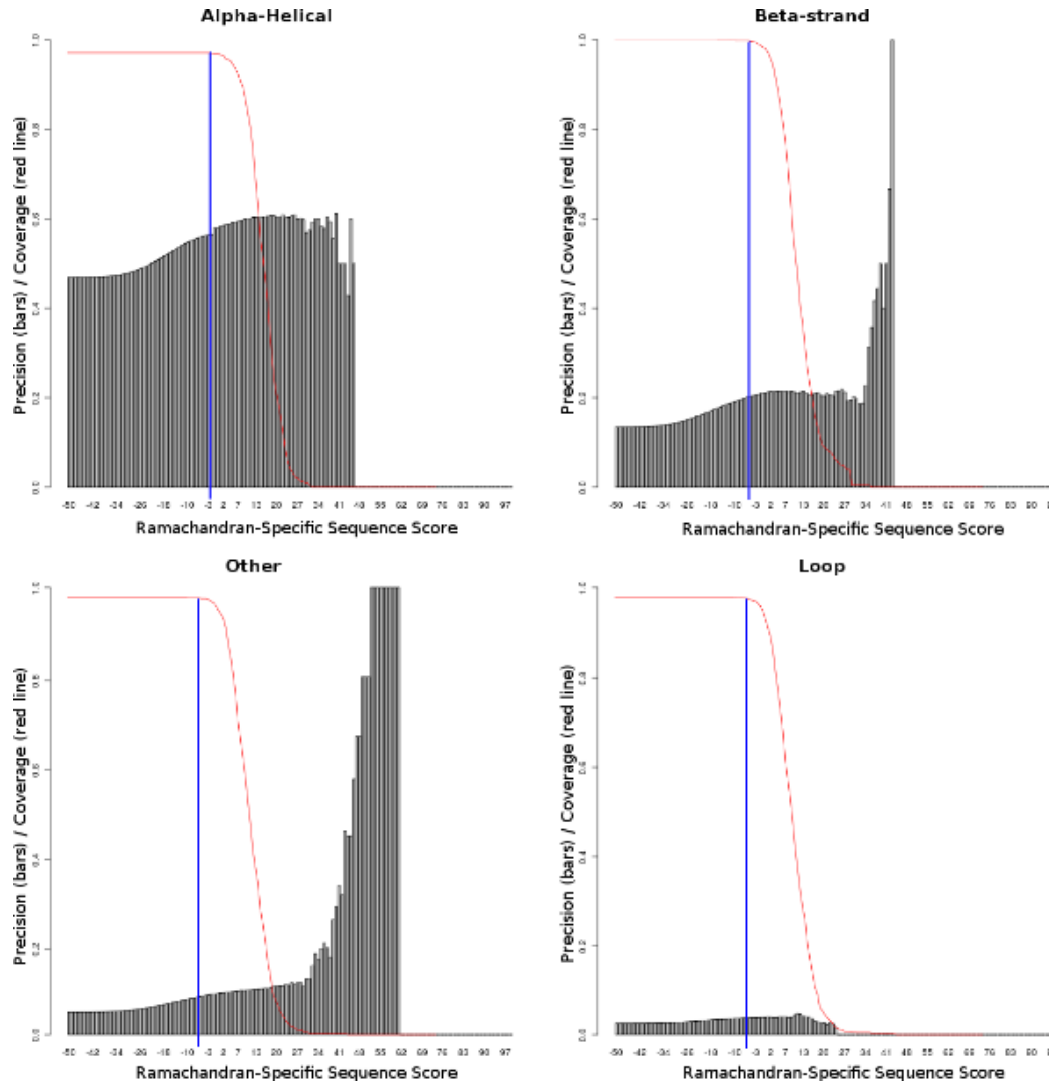
Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library



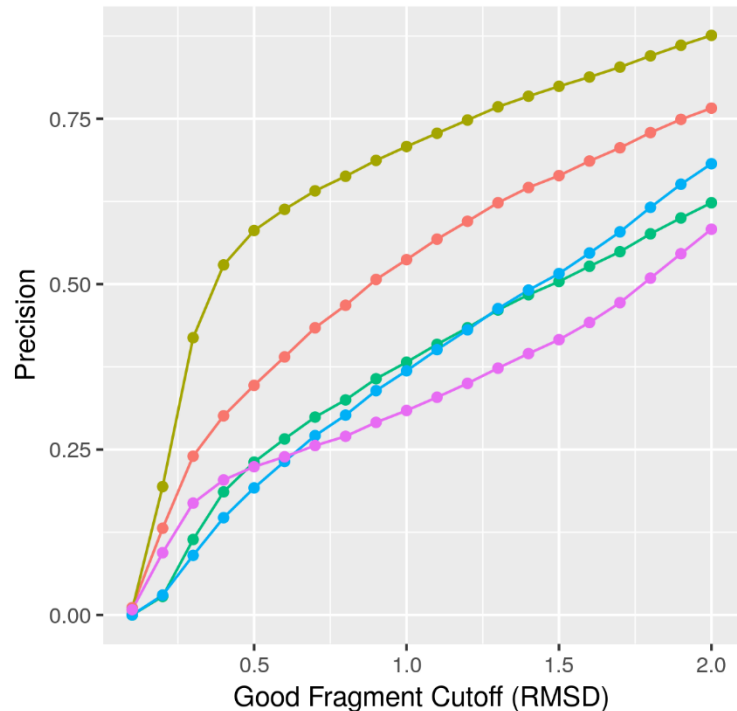
Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library



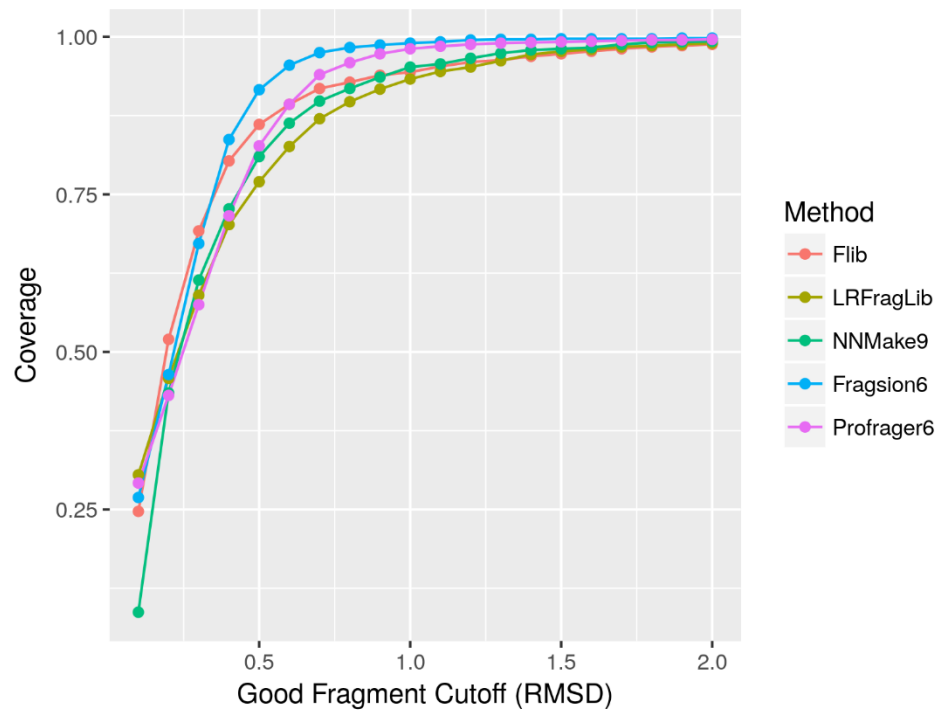
Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library



Method

- Flib
- LRFRagLib
- NNMake9
- Fragsion6
- Profrager6



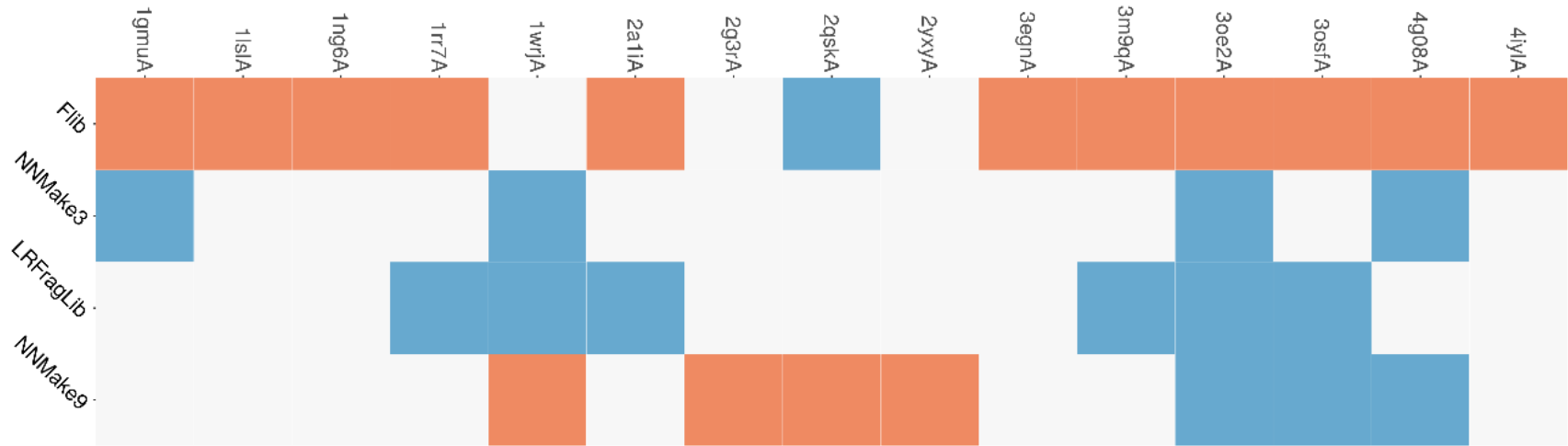
Method

- Flib
- LRFRagLib
- NNMake9
- Fragsion6
- Profrager6

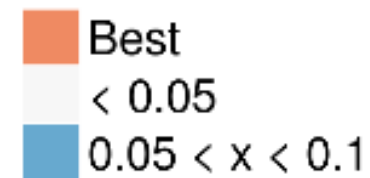
NNMAKE – Gront et al (2011)
FLIB – Oliveira et al (2015)
LRFRagLib – Wang et al (2016)
Fragsion – Bhattacharya et al (2016)
Profrager – Santos et al (2015)

Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library

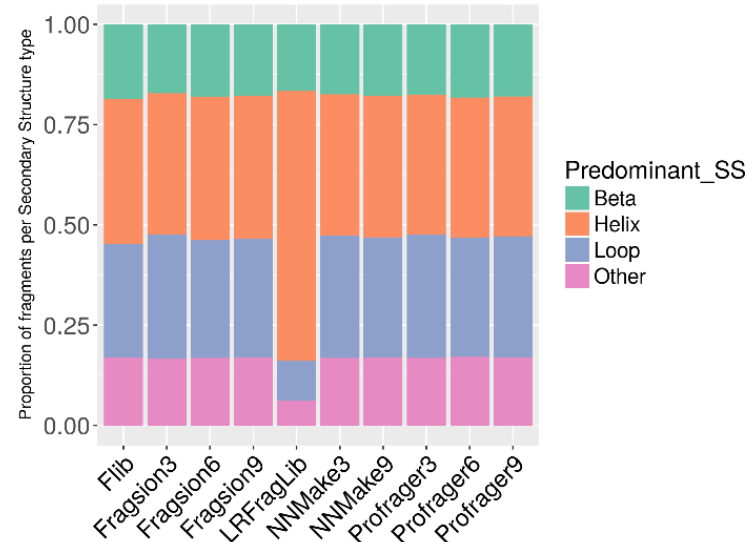
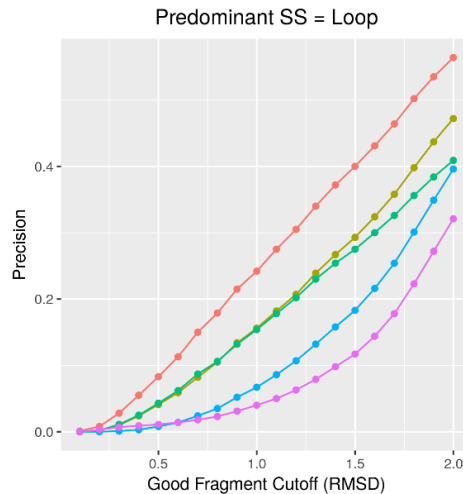
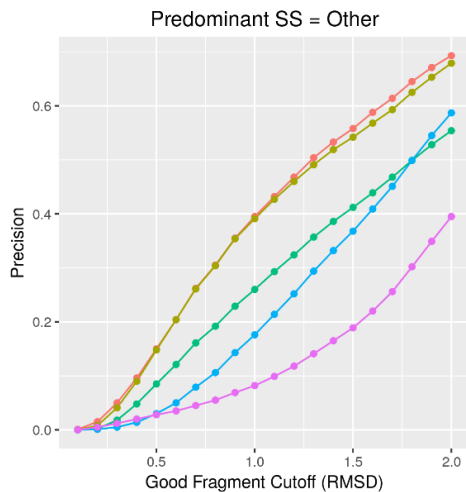
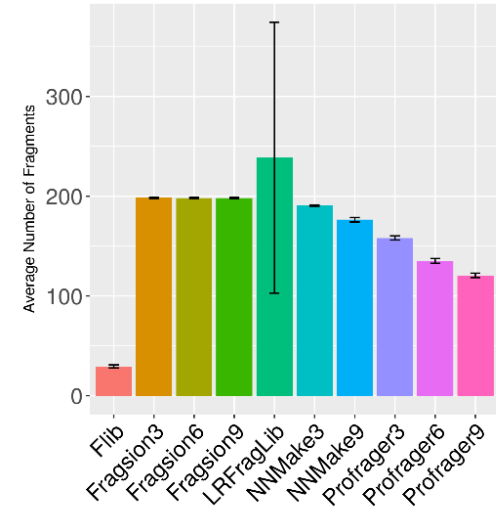
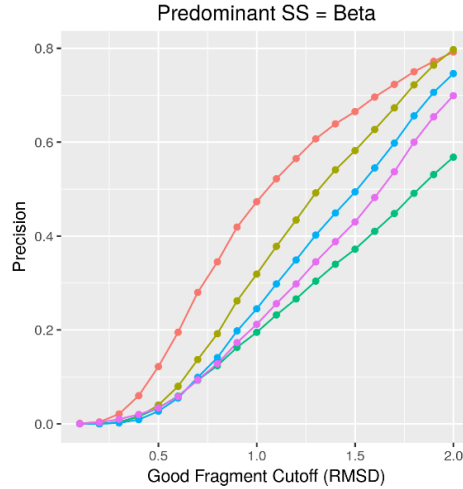
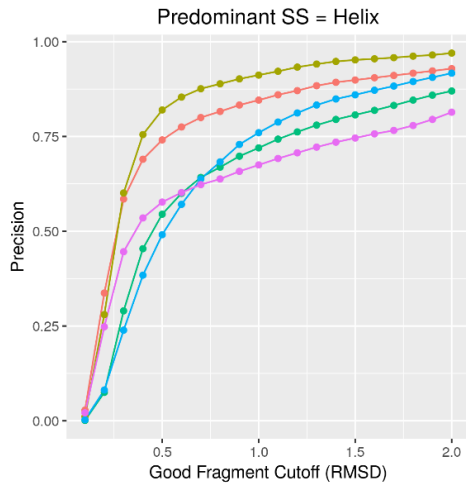


TM-Score
Difference
to Best
Possible
Model



Ways to improve Fragment assembly

- Consider secondary structure when assessing your fragment library



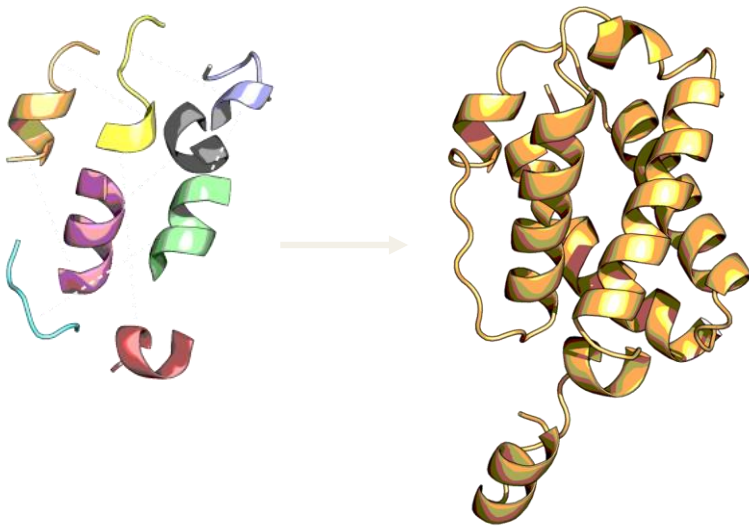
Looking for inspiration in biology: co-evolution

Protein contacts can be predicted using this co-evolutionary signal.

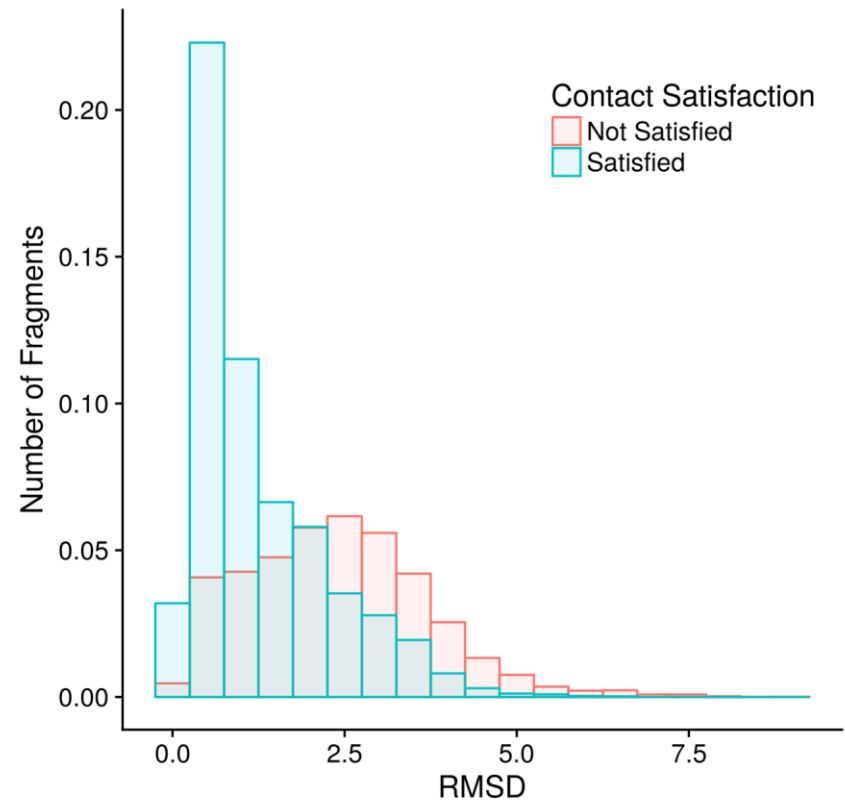
Q5E940 BOVIN	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	M	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_HUMAN	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	M	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_MOUSE	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	M	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_RAT	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	M	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_CHICK	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	M	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_RAMSY	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	M	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
Q7ZUG3_BRARE	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	L	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_ICTTPU	-----M	P	R	E	D	R	A	T	W	S	H	Y	F	L	K	I	T	L	L	D	D	P	K	C	F	V	G	A	D	N	V	G	S	K	M	Q	I	R	L	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76									
RLA0_DROME	-----M	V	R	E	N	K	A	A	W	A	Q	A	Y	E	I	K	V	F	L	F	D	E	F	P	K	C	F	V	G	A	D	N	V	G	S	K	M	N	I	R	T	S	L	R	G	-	A	V	L	M	G	K	N	M	M	R	K	A	I	R	G	H	L	E	N	--	P	A	L	E	76							
RLA0_DICDI	-----M	S	G	A	G	-	S	K	R	K	L	F	T	E	K	A	T	K	L	F	T	T	D	K	H	I	V	A	E	A	D	F	V	G	S	S	O	L	K	I	R	K	S	I	R	G	-	G	A	V	L	M	G	K	N	M	I	R	K	V	I	R	D	L	A	D	S	K	--	P	E	L	D	75				
Q54LPO_DICDI	-----M	S	G	A	G	-	S	K	R	K	L	F	T	E	K	A	T	K	L	F	T	T	D	K	H	I	V	A	E	A	D	F	V	G	S	S	O	L	K	I	R	K	S	I	R	G	-	G	A	V	L	M	G	K	N	M	I	R	K	V	I	R	D	L	A	D	S	K	--	P	E	L	D	75				
RLA0_PLAF8	-----M	A	K	L	S	Q	O	Q	K	M	Y	T	E	K	L	S	L	I	O	S	K	L	I	V	H	V	D	N	V	G	S	H	M	A	S	V	K	S	L	R	G	-	A	L	L	M	G	K	N	M	R	I	T	A	L	K	N	L	A	V	--	P	O	T	E	76												
RLA0_SULAC	-----M	T	C	L	A	V	T	T	K	K	T	A	K	H	V	D	E	A	E	L	E	L	K	T	K	T	I	L	I	A	N	E	G	F	P	A	D	K	H	E	I	K	K	L	R	G	-	A	D	I	K	V	K	N	L	E	A	L	K	N	A	G	--	P	O	T	E	79										
RLA0_SULTO	-----M	R	I	M	A	V	I	T	Q	E	R	K	A	K	K	I	E	E	K	L	E	K	R	E	H	T	I	L	I	A	N	E	G	F	P	A	D	K	H	D	I	K	K	M	R	G	-	A	E	I	K	V	K	N	L	E	G	I	A	A	K	N	A	G	--	P	E	L	D	80								
RLA0_SULSO	-----M	K	R	L	A	L	A	L	K	Q	R	K	V	A	S	W	K	L	E	E	V	K	L	E	L	I	K	N	S	N	T	I	L	I	G	N	L	E	G	F	P	A	D	K	H	E	I	K	K	L	R	G	-	A	T	I	K	V	K	N	L	E	G	I	A	A	K	N	A	G	--	P	E	L	D	80		
RLA0_AERPE	-----M	S	V	S	I	V	G	M	Y	K	R	E	K	P	T	E	P	E	T	L	M	L	R	E	L	E	L	F	S	K	R	V	L	F	A	D	L	T	C	T	P	F	V	P	R	V	K	K	L	W	K	-	D	M	V	A	K	K	I	L	R	A	M	K	A	A	G	L	E	--	P	E	L	D	86			
RLA0_PYRAE	-----M	H	L	A	T	Q	K	R	V	T	R	Q	P	A	R	K	V	S	R	T	E	L	O	K	P	Y	F	L	D	I	H	G	S	E	R	L	H	E	E	K	R	Y	L	R	Y	-	G	V	K	I	G	E	L	E	K	I	A	T	T	Y	G	G	--	P	E	L	D	85										
RLA0_METAC	-----M	A	E	E	R	H	T	E	H	P	Q	W	K	D	E	T	E	N	K	E	L	Q	S	K	V	F	G	M	V	G	E	G	I	L	A	T	K	I	R	D	L	K	D	V	-	A	V	L	K	V	S	R	N	L	E	R	A	L	N	G	L	--	P	E	L	D	78											
RLA0_METMA	-----M	A	E	E	R	H	T	E	H	P	Q	W	K	D	E	T	E	N	K	E	L	Q	S	K	V	F	G	M	V	R	E	G	I	L	A	T	K	I	R	D	L	K	D	V	-	A	V	L	K	V	S	R	N	L	E	R	A	L	N	G	L	--	P	E	L	D	78											
RLA0_ARCFU	-----M	A	A	V	E	S	--	P	E	P	E	V	R	A	V	E	E	K	R	M	I	S	S	K	V	V	A	I	V	E	R	N	V	E	A	G	O	M	O	K	I	R	E	F	R	G	-	A	E	I	K	V	K	N	L	E	R	A	L	D	A	L	G	--	P	E	L	D	75									
RLA0_METKA	-----M	A	V	K	A	G	O	P	S	G	Y	E	K	V	A	E	W	R	R	E	V	K	E	L	M	D	E	V	N	V	G	L	V	L	E	G	I	P	A	P	O	L	E	I	R	A	K	L	R	E	R	D	-	I	I	R	S	R	N	L	M	R	I	A	L	E	E	K	L	D	E	R	--	P	E	L	D	88
RLA0_METTH	-----M	A	H	V	A	E	W	K	K	E	Q	E	L	H	L	I	K	E	V	V	G	I	A	N	L	A	D	I	P	A	R	L	O	K	M	Q	T	L	R	D	S	-	A	L	T	R	S	K	L	I	S	L	A	L	E	K	A	G	R	E	L	--	P	E	L	D	74											
RLA0_METTL	-----M	I	T	A	S	E	H	K	A	P	K	I	E	W	K	I	E	L	K	M	Q	I	V	A	L	D	M	V	A	N	L	E	I	D	I	K	I	R	-	E	T	H	L	K	M	N	L	L	E	R	A	K	E	V	A	E	T	W	N	E	P	A	--	P	E	L	D	82										
RLA0_METVA	-----M	I	D	A	S	E	H	K	A	P	K	I	E	W	K	I	E	L	K	S	N	V	I	A	L	D	M	V	A	N	L	E	I	D	I	K	I	R	-	D	Q	M	L	K	M	N	L	L	K	A	V	E	V	A	E	T	W	N	E	P	A	--	P	E	L	D	82											
RLA0_METJA	-----M	E	T	K	V	K	A	H	V	A	D	K	I	E	W	K	I	E	L	K	S	N	V	I	A	L	D	M	V	A	N	L	E	I	D	I	K	I	R	-	D	K	V	L	M	S	R	N	L	L	I	R	A	L	E	K	A	E	L	N	N	K	I	A	--	P	E	L	D	81								
RLA0_PYRAB	-----M	A	H	V	A	E	W	K	K	E	V	E	L	A	N	I	K	S	V	V	I	A	L	D	V	S	S	M	P	A	P	L	S	M	R	L	I	R	E	N	G	L	L	R	V	S	R	N	L	L	E	L	A	I	K	A	A	E	L	G	K	P	E	L	--	P	E	L	D	77								
RLA0_PYRFU	-----M	A	H	V	A	E	W	K	K	E	V	E	L	A	N	I	K	S	V	V	I	A	L	D	V	S	S	M	P	A	P	L	S	M	R	L	I	R	E	N	G	L	L	R	V	S	R	N	L	L	E	L	A	I	K	A	A	E	L	G	K	P	E	L	--	P	E	L	D	77								
RLA0_PYRKO	-----M	A	H	V	A	E	W	K	K	E	V	E	L	A	N	I	K	S	V	V	I	A	L	D	V	S	S	M	P	A	P	L	S	M	R	L	I	R	E	N	G	L	L	R	V	S	R	N	L	L	E	L	A	I	K	A	A	E	L	G	K	P	E	L	--	P	E	L	D	76								
RLA0_HALMA	-----M	S	E	S	E	R	Q	T	E	V	P	Q	W	R	E	V	D	A	I	N	H	I	S	E	S	V	G	V	N	A	G	I	D	R	L	D	M	R	D	L	H	E	T	-	A	E	L	R	V	S	R	N	L	E	R	A	L	D	D	V	--	P	E	L	D	79												
RLA0_HALVO	-----M	S	E	S	E	R	Q	T	E	V	P	Q	W	R	E	V	D	L	V	D	I	E	S	E	S	V	G	V	N	A	G	I	D	R	L	D	M	R	D	L	H	E	T	-	A	A	V	R	M	S	R	N	L	V	N	R	A	L	D	E	V	N	--	P	E	L	D	79										
RLA0_HALSA	-----M	S	E	S	E	R	Q	T	E	V	P	Q	W	R	E	V	D	L	V	D	I	E	S	E	S	V	G	V	N	A	G	I	D	R	L	D	M	R	D	L	H	E	T	-	A	A	V	R	M	S	R	N	L	V	N	R	A	L	D	E	V	N	--	P	E	L	D	79										
RLA0_THEAC	-----M	K	E	S	O	O	K	E	L	N	E	I	D	R	K	A	S	R	S	V	A	I	V	D	A	G	I	R	I	R	O	D	I	R	G	K	N	K	G	-	I	N	L	K	I	K	L	L	I	F	K	A	L	E	N	L	D	--	P	E	L	D	72															
RLA0_THEVO	-----M	K	E	S	O	O	K	E	L	N	E	I	D	R	K	A	S	R	S	V	A	I	V	D	A	G	I	R	I	R	O	D	I	R	G	K	N	K	G	-	I	N	L	K	I	K	L	L	I	F	K	A	L	E	N	L	D	--	P	E	L	D	72															
RLA0_PICTO	-----M	T	E	P	O	M	I	D	F	K	M	L	E	N	I	S	H	K</																																																												

Coevolution can be used to produce better fragment libraries

Fragments from known protein structures can be used to guide the conformational search.



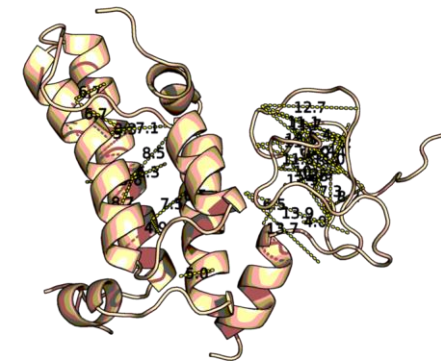
Fragments that satisfy predicted contacts tend to be of better quality.



Coevolution can be used in the energy function

Q5E940	BOVIN	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	HUMAN	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	MOUSE	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	RAT	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	CHICK	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	RAMSY	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
Q7ZUG3	BRARE	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	ICTPU	-----	MPREDRATWKS	NYLKIITLDD	PKCFIVGADNVGSK	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	DROME	-----	MVRENKAQAQY	EIKVVFDFE	PKCFIVGADNVGSK	MQNIETSLRGL	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	DICDI	-----	MSGAG	SKRRKILFTEKATKIFTT	DKMIVAEAD	FVGG	SLOKIRKISIRGI	GAVLMGKMMRKAIRGHLENN	-----	PALE	75									
Q54LPO	DICDI	-----	MSGAG	SKRRKILFTEKATKIFTT	DKMIVAEAD	FVGG	SLOKIRKISIRGI	GAVLMGKMMRKAIRGHLENN	-----	PALE	75									
RLA0	PLAF8	-----	MAKLSQOKQMY	TEKLSLLOQ	SKILIVHYD	VGSHMASVYKSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	SULAC	-----	MLGLAVTTTKK	IAKVVDEVAELTE	KLTKTILIAN	TEGFPADK	HEIKKLRGK	ADIKVKNLEN	IALKNAG	-----	MDIK	79								
RLA0	SULTO	-----	MRIMAVITQEK	IAKVVDEVAELTE	KLTKTILIAN	TEGFPADK	HEIKKLRGK	ADIKVKNLEN	IALKNAG	-----	MDIK	80								
RLA0	SULSO	-----	MKRLALALKQRK	VASWKEEVEKLE	TELKNSNTIL	IGNLEGF	PADKHEIKKLRGK	ADIKVKNLEN	IALKNAG	-----	MDIK	80								
RLA0	AERPE	-----	MSVVSIVG	QMYKREKPTD	EWFTLMLRE	LEELFSKRVVLF	ADLTCPTFVY	RVVKKLWKK	DMVAKKRILL	RAMKAAGLE	-----	LDNN	86							
RLA0	PYRAB	-----	MHLATQKRYVTR	IQDARKKIV	SYRTELOK	QVWFLD	DMGLS	RIIIE	RYKRLRY	GVKIKTE	FLKTAFTY	GG	-----	LDVA	85					
RLA0	METAC	-----	MAEERHTEH	EPQW	KDDEIENIKEL	IQSKVFG	MYRTEG	LLATKMK	IRRDLDKV	AVLKVSRN	LLERAINGL	-----	ETID	78						
RLA0	METMA	-----	MAEERHTEH	EPQW	KDDEIENIKEL	IQSKVFG	MYRTEG	LLATKMK	IRRDLDKV	AVLKVSRN	LLERAINGL	-----	ETID	78						
RLA0	ARCFU	-----	MAAVRGS	-----	PEVY	VRVAEELKRM	ISSKVVVAI	VERNVFAG	QOKIRREF	FRGK	AEIKVKNLEN	IALDALC	-----	GDIL	75					
RLA0	METKA	-----	MAVKAKG	QPSG	YEPKVAE	WRREVKELK	EMDENV	GVLDLEGG	IPAPLOE	IRAKLRERD	ITIRMSRNL	LMRLALEEK	LDER	-----	PELE	88				
RLA0	METTH	-----	MAHVAE	WKKEVQ	ELHDLTK	EVVGL	IANLADIP	ARDLQ	KMQRTLRDS	ALTRMSK	RLISL	ALEKAGREL	-----	FWVD	74					
RLA0	METTL	-----	MITAE	SEHKAPKIE	VNKLKEL	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	METVA	-----	MIDAK	SEHKAPKIE	VNKLKEL	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	METJA	-----	METK	VKAHVADKIE	VNKLKEL	KMQQIRMSLRGK	AVVLMGKMMRKAIRGHLENN	-----	PALE	76										
RLA0	PYRAB	-----	MAHVAE	WKKEVQ	ELHDLTK	EVVGL	IANLADIP	ARDLQ	KMQRTLRDS	ALTRMSK	RLISL	ALEKAGREL	-----	FWVD	74					
RLA0	PYRFU	-----	MAHVAE	WKKEVQ	ELHDLTK	EVVGL	IANLADIP	ARDLQ	KMQRTLRDS	ALTRMSK	RLISL	ALEKAGREL	-----	FWVD	74					
RLA0	PYRKO	-----	MAHVAE	WKKEVQ	ELHDLTK	EVVGL	IANLADIP	ARDLQ	KMQRTLRDS	ALTRMSK	RLISL	ALEKAGREL	-----	FWVD	74					
RLA0	HALMA	-----	MSSEFRKTET	EPQW	KDDEIENIKEL	IQSKVFG	MYRTEG	LLATKMK	IRRDLDKV	AVLKVSRN	LLERAINGL	-----	ETID	78						
RLA0	HALVO	-----	MSSEFRKTET	EPQW	KDDEIENIKEL	IQSKVFG	MYRTEG	LLATKMK	IRRDLDKV	AVLKVSRN	LLERAINGL	-----	ETID	78						
RLA0	HALSA	-----	MSSEFRKTET	EPQW	KDDEIENIKEL	IQSKVFG	MYRTEG	LLATKMK	IRRDLDKV	AVLKVSRN	LLERAINGL	-----	ETID	78						
RLA0	THEAC	-----	MKE	SNQKEL	VNTEIT	IRKASR	VAVDLAG	IRIR	RODID	IRKNGK	INLKV	IKLLIF	KALENLD	-----	EKIS	79				
RLA0	THEVO	-----	MKK	SNQKEL	VNTEIT	IRKASR	VAVDLAG	IRIR	RODID	IRKNGK	INLKV	IKLLIF	KALENLD	-----	EKIS	79				
RLA0	PICTO	-----	MTE	SNQKEL	VNTEIT	IRKASR	VAVDLAG	IRIR	RODID	IRKNGK	INLKV	IKLLIF	KALENLD	-----	EKIS	79				
tuLee	1	10	20	30	40	50	60	70	80	90	

Protein contacts can be predicted using this co-evolutionary signal.



Multiple sequence alignments can be used to identify pairs of residues that co-evolve.

Predicted contacts improve scoring and lead to accurate protein structure prediction.

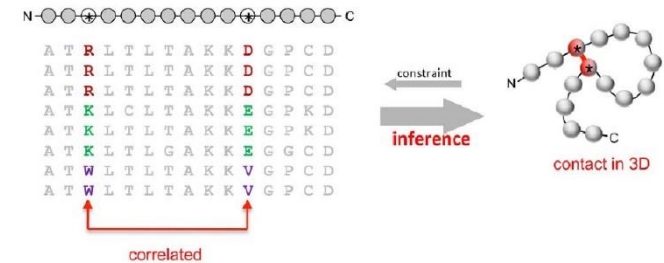
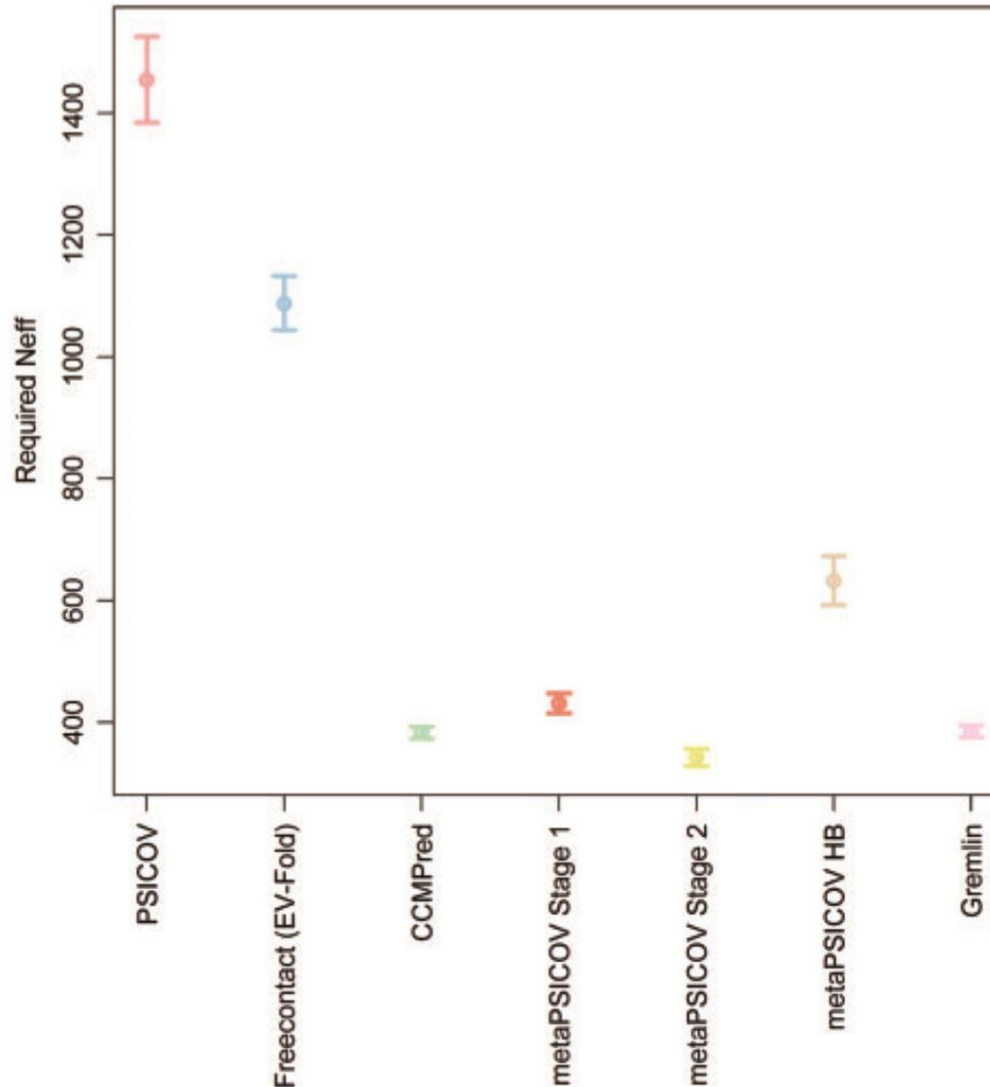
Co-evolution methods

- 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)

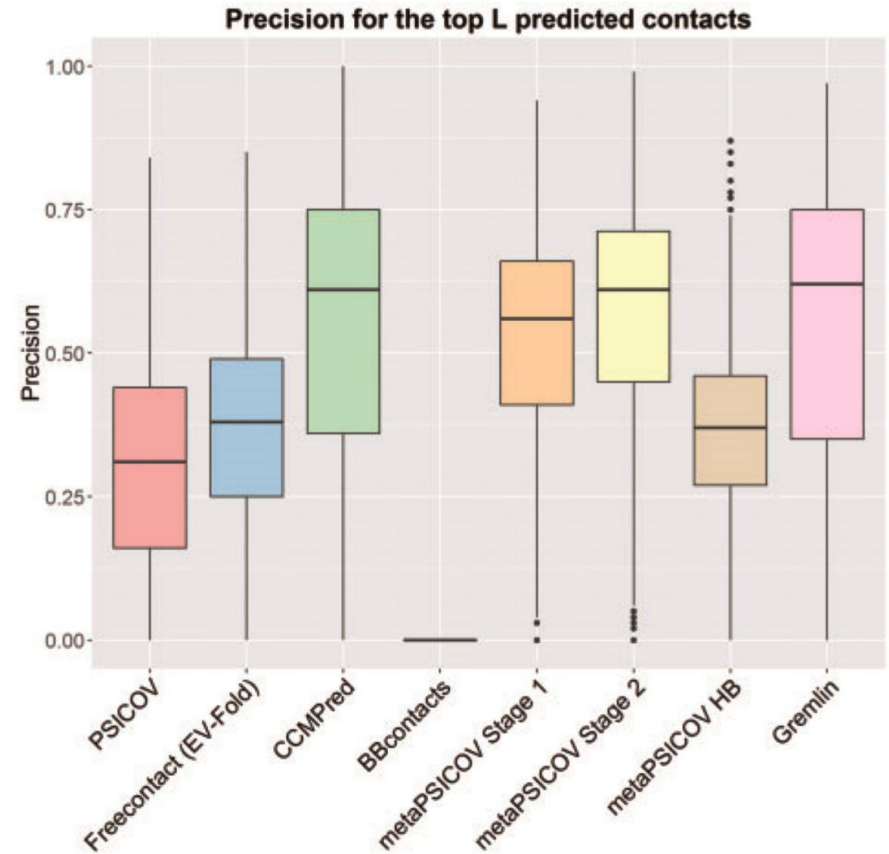
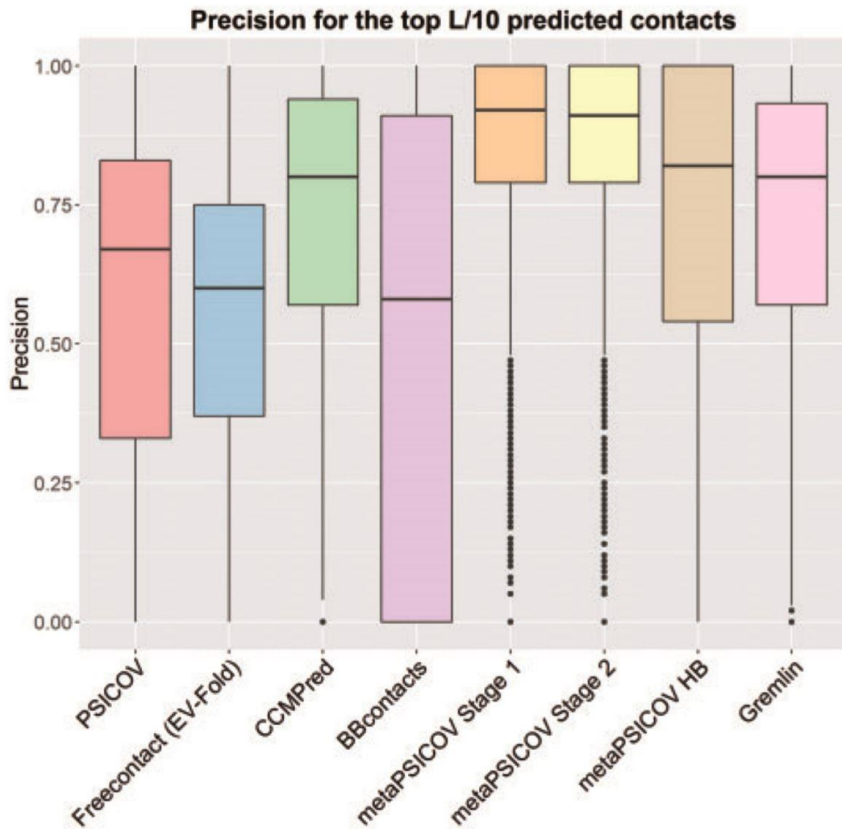
Contact definition

- Two protein residues are defined to be in contact if their C- β s (C- α s for Glycine) are less than 8 Å 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 $5 \leq |i - j| \leq 23$.
- A long range contact is defined when $|i - j| > 23$

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



How accurate are the methods?



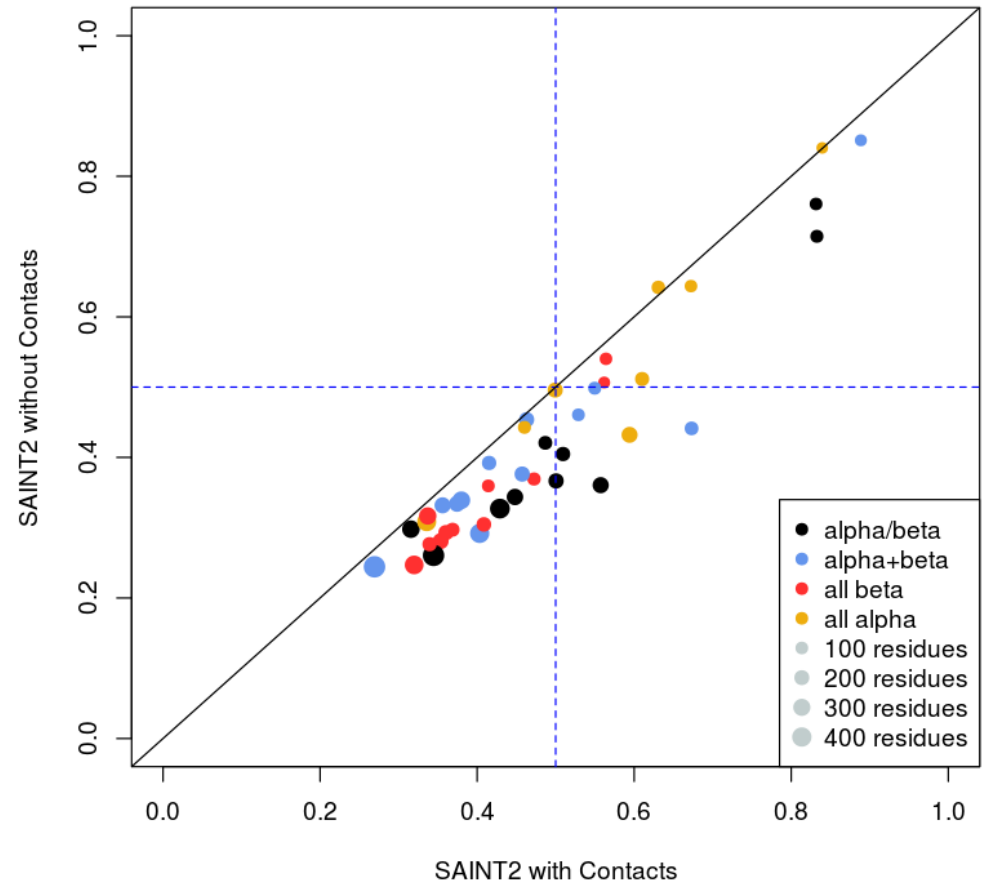
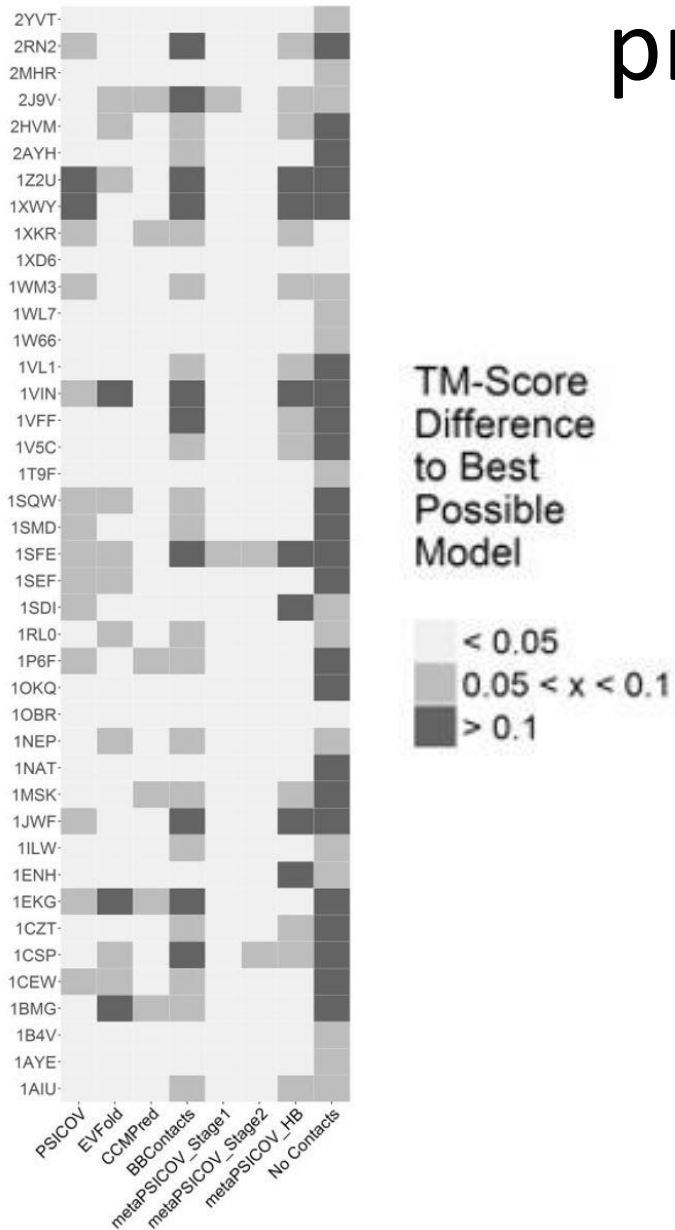
Putting co-evolutionary contacts into protein structure prediction

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

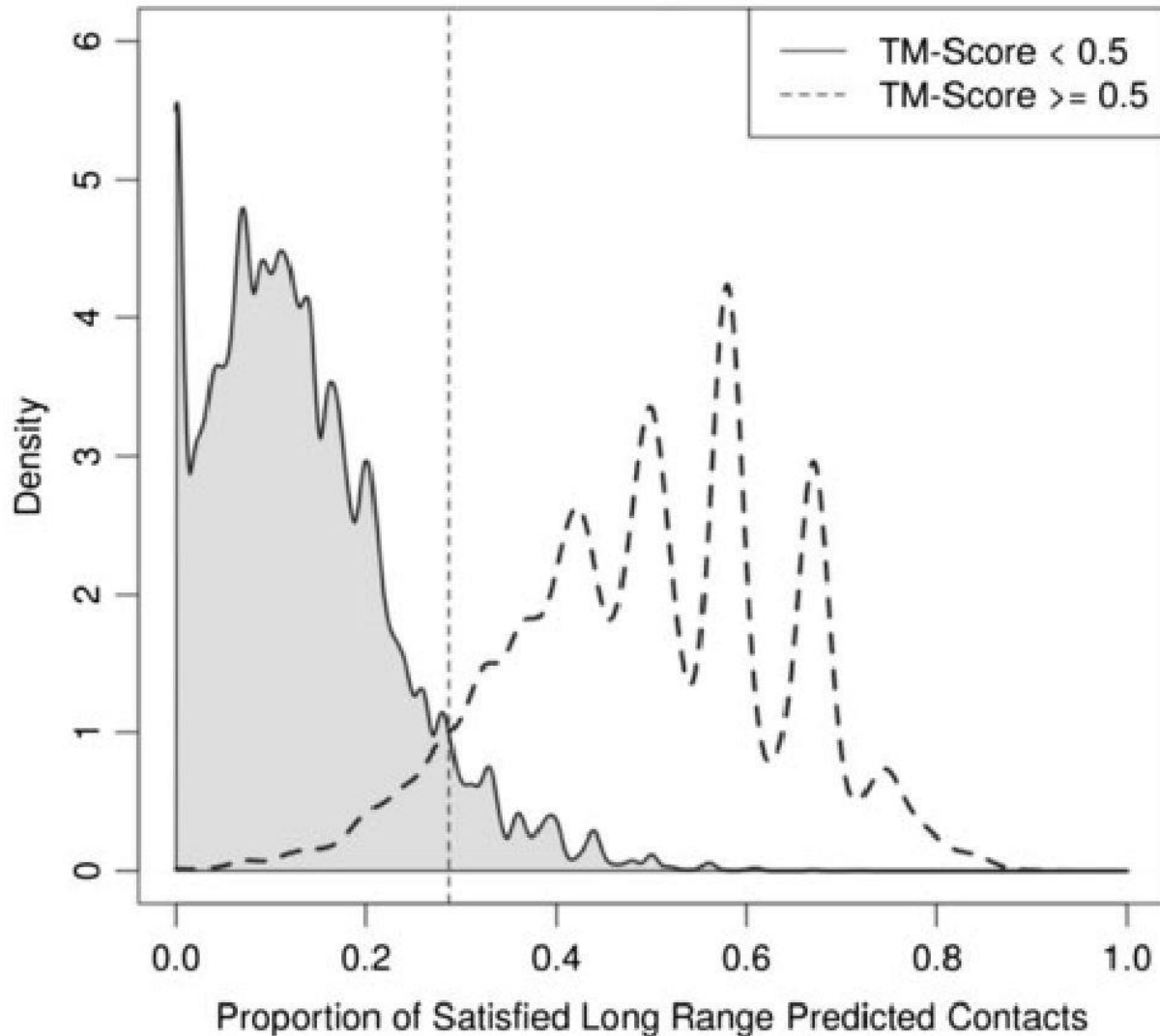
Where $\mathbf{C}_\beta(i)$ and $\mathbf{C}_\beta(j)$ represent the coordinates of the C- β s (C- α s in the case of glycine) of residues i and j and:

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

How do they influence structure prediction?

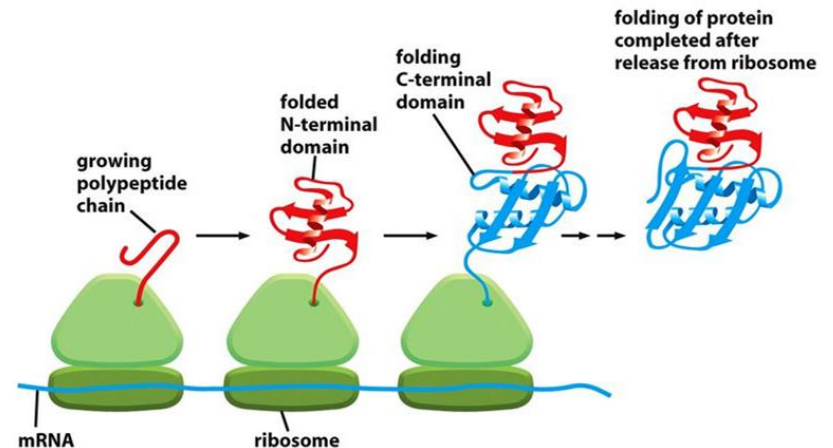


Using co-evolution contacts to identify good models



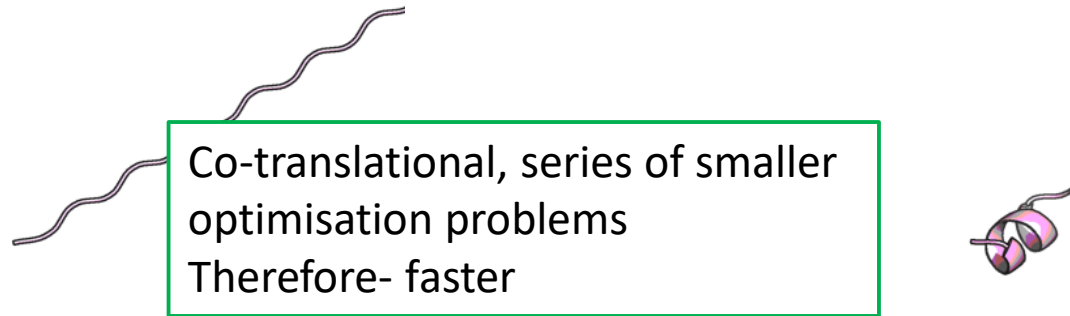
Inspiration from biology: cotranslational protein folding

- Folding is orders of magnitude faster than translation.
- Cotranslational protein folding not necessary for all proteins to reach their native state.
- Cotranslational protein folding is faster/more efficient than *in vitro* re-folding.



Hypothesis: CT folding guides proteins towards their native state by restricting the conformational search space.

Template-free search strategies with SAINT2



Traditional:
Non-sequential approach

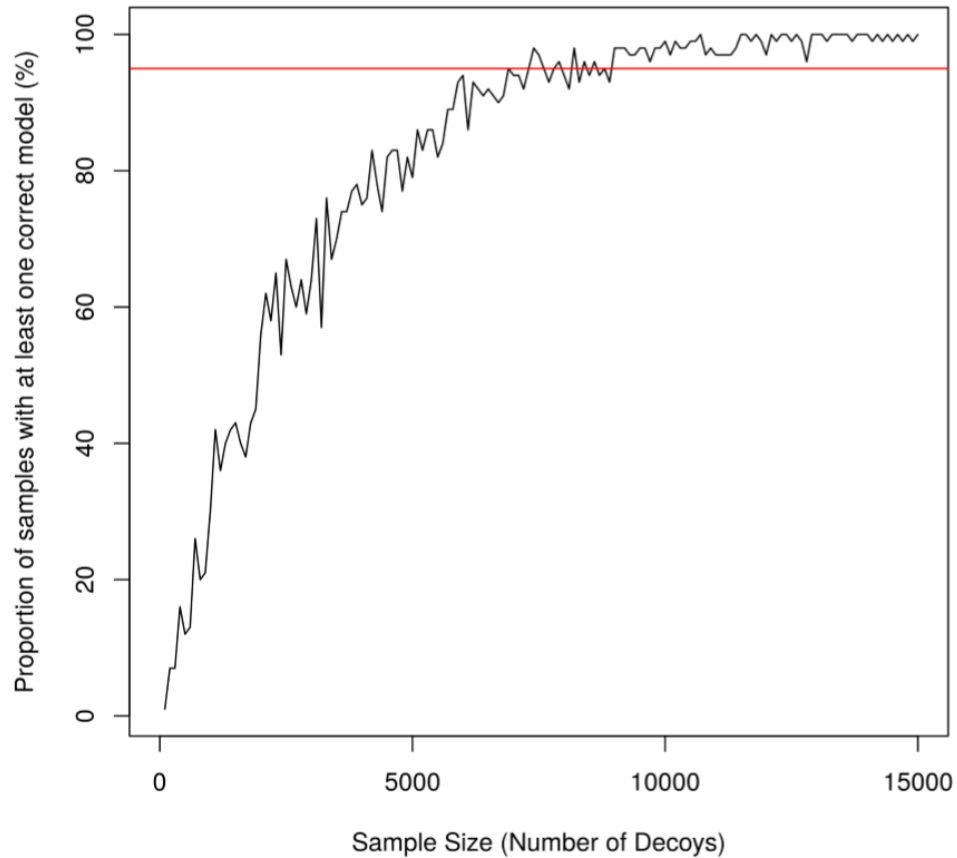
Biologically Inspired:
Cotranslational approach

Number of decoys required

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

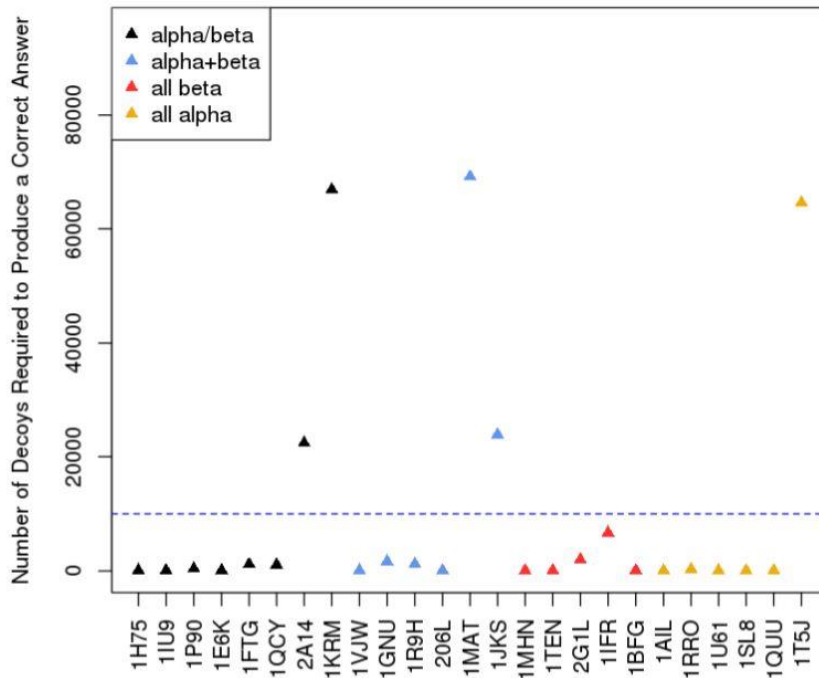
Method:	Number of Decoys:
FRAGFOLD (6)	200
CABS(7)	360
MBS (8)	3,000
RBOaleph (9)	1,000-5,000
QUARK (10)	5,000
Nefilim (11)	150,000
EDAfold (12)	200,000
Rosetta (13)	20,000-900,000

Number of decoys required



Number of decoys required

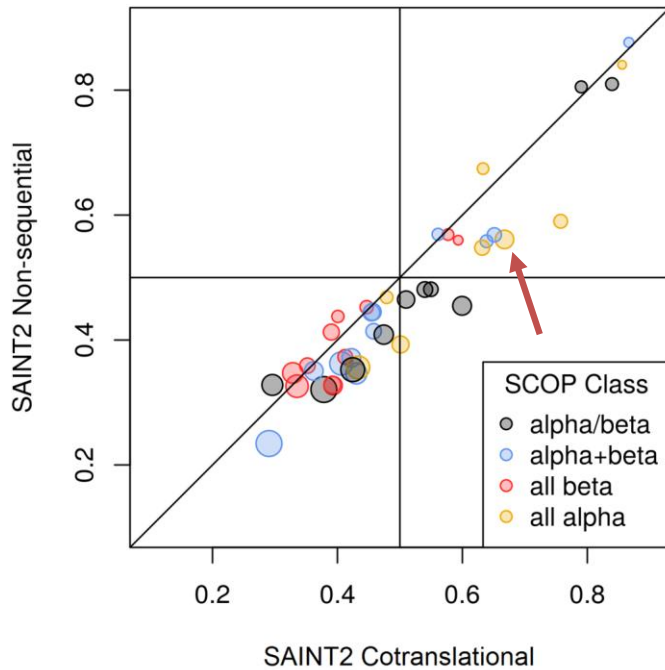
A



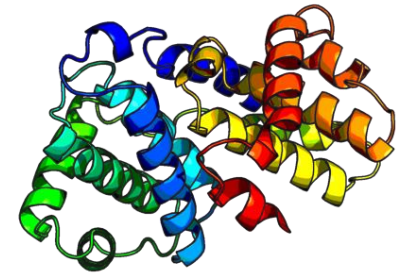
- Number decoys to get a correct answer ~10,000
- Number of decoys to get best answer ~20,000
- Not dependent on protein length (if length <250)

Cotranslational prediction produces better models

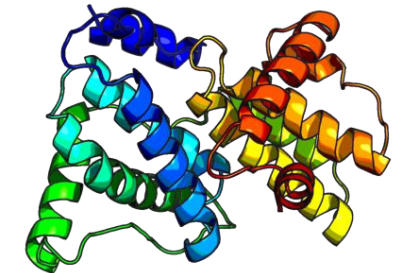
Comparing search strategies:



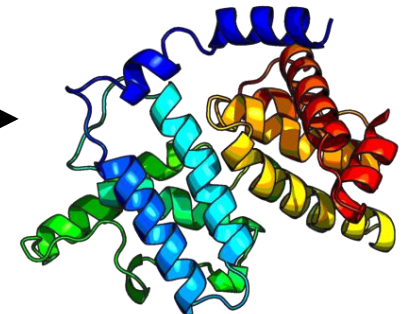
Native: 1VIN →



Cotranslational →



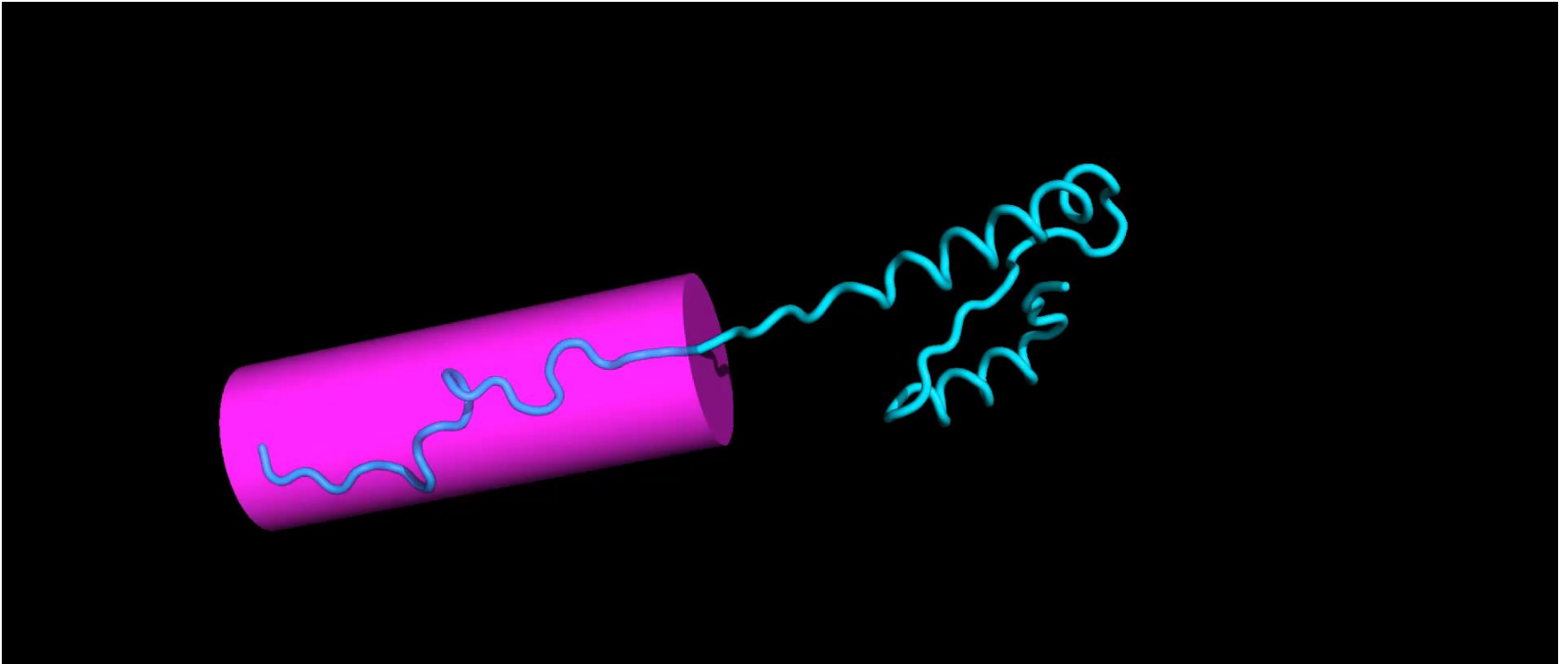
Non-sequential →



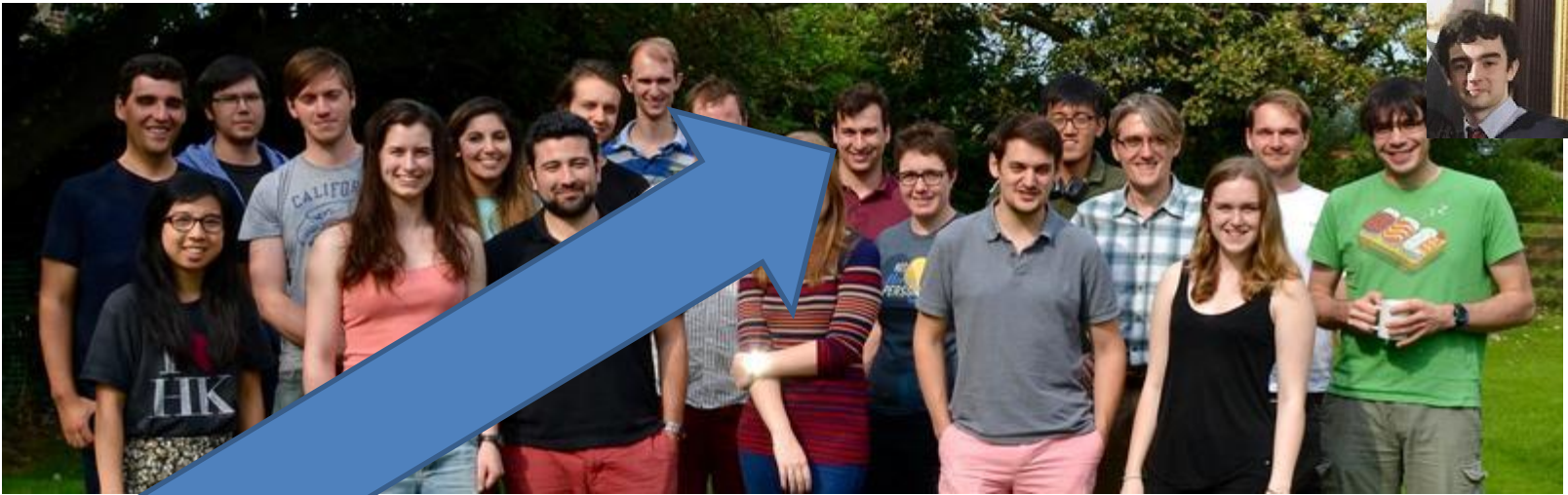
Improving the search: Cotranslational protein structure prediction

- 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
 - Requires fewer decoys to produce a good answer
- 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

What next?



ACKNOWLEDGEMENTS



e-Therapeutics plc



WONKA and OOMMPPAA

Select compound to be shown

OOMMPPAA Targets Download Help Input SMILES or code Search Built With Bootstrap

CDK2

Pharmacophore change

Improving activity

Reducing activity

Centre of mass of each pair

View in 3D

Pharmacophoric changes:

Activity change: 6

Refresh

Pharmacophore changes

Hydrophobic Acceptor Donor Aromatic

INCREASE in activity

DECREASE in activity

All points

Increasing activity

Display compounds Remove

Max

IC50 Activity: 7.17 IC50 Activity: 5.00

IC50 Activity: 7.17 IC50 Inactive: under 5.00

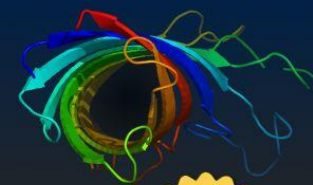
2D activity data in matched pairs

Control data shown

Filter data shown

Memoir

Membrane protein modelling pipeline

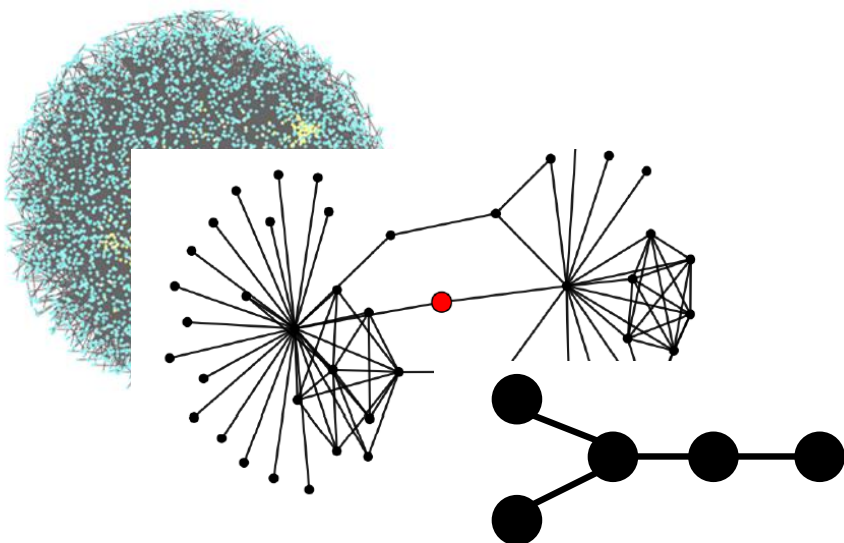


Help

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

