

Bioinformatics for structural biologists

Dan Rigden



UNIVERSITY OF
LIVERPOOL

Contents

- Predicting contacts from sequences
- Bioinformatics for MR
 - Predicting domain structure
 - Informs phasing strategy, but also relevant earlier
 - Predicting tertiary structure
 - Homologous relationships identify search models for conventional MR
 - **AMPLE** does unconventional MR with *ab initio* models etc
 - Predicting quaternary structure
- Bioinformatics for structure-based function prediction
 - Using newer/less well-known data
 - Majoring on easily available servers/predictions

PREDICTION OF
STRUCTURE

PREDICTION (AND
EXPLANATION) OF
FUNCTION

```
>target_seq (s)
GYVELYRRSTIGNSLVDALDTLISDG
RIEASLAMRVLETFDKVVAETLKDNT
QSKLTVKGNLDTYGFCDVWTFIVKN
CQVTVEDQSVISVDKLRIVACNSKKS
```

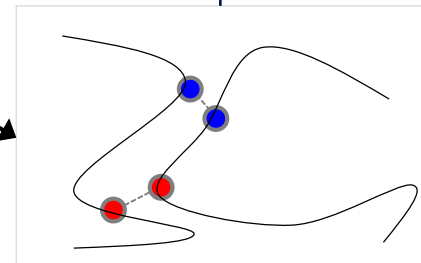
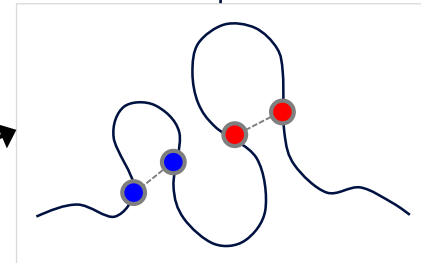
Multiple Sequence Alignment of homologous sequences

...KDNTQSKLTVKGNLDTYGC...
...KDNTQSKLTVKGNLDTYGC...
...KNDANIAKLKLTFGDLDTY...
...RENTQSKLNVKGLLDYGC...
...KNENIVGKSKLIFGDLDTY...
...KNESNIAKSKLTFKGLDTY...
...KHESNIAKSKLTFKGLDTY...
...KNESGISKSKLTFKGLDHTY...
...KDNTQSKLTVKGLLDYGC...
...KSESGIAKSKLTFKGLDHTY...

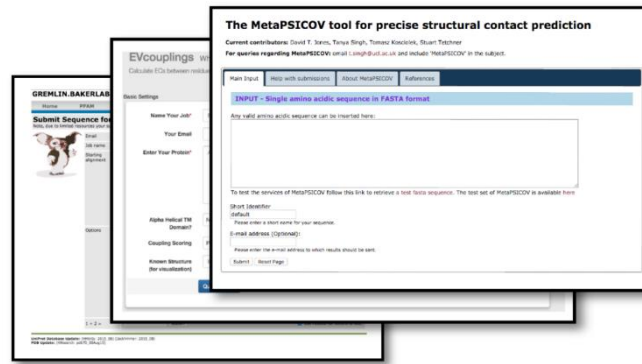
...KDNTQSKLTVK NLDYTC...
...KDNTQSKLTVK NLDYTC...
...KNDANIAKLKLT FKGLDTY...
...RENTQSKLNVK LLDYTC...
...KNENIVGKSKL FKGLDTY...
...KNESNIAKSKL FKGLDTY...
...KHESNIAKSKL FKGLDTY...
...KNESGISKSKL FKGLDHTY...
...KDNTQSKLTVK LLDYTC...
...KSESGIAKSKL FKGLDHTY...

Direct Coupling Analysis of
large MSAs plus Deep
Learning predicts contacts

Applications...



```
>seq
FASDGITF
DRSLFFGH
```

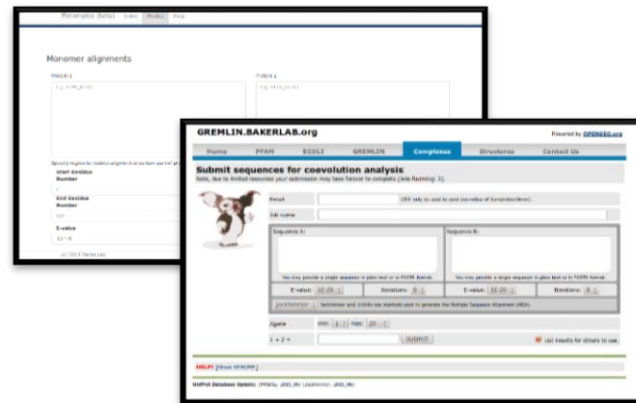


Intramolecular
contacts

Intermolecular
contacts if a
homooligomer

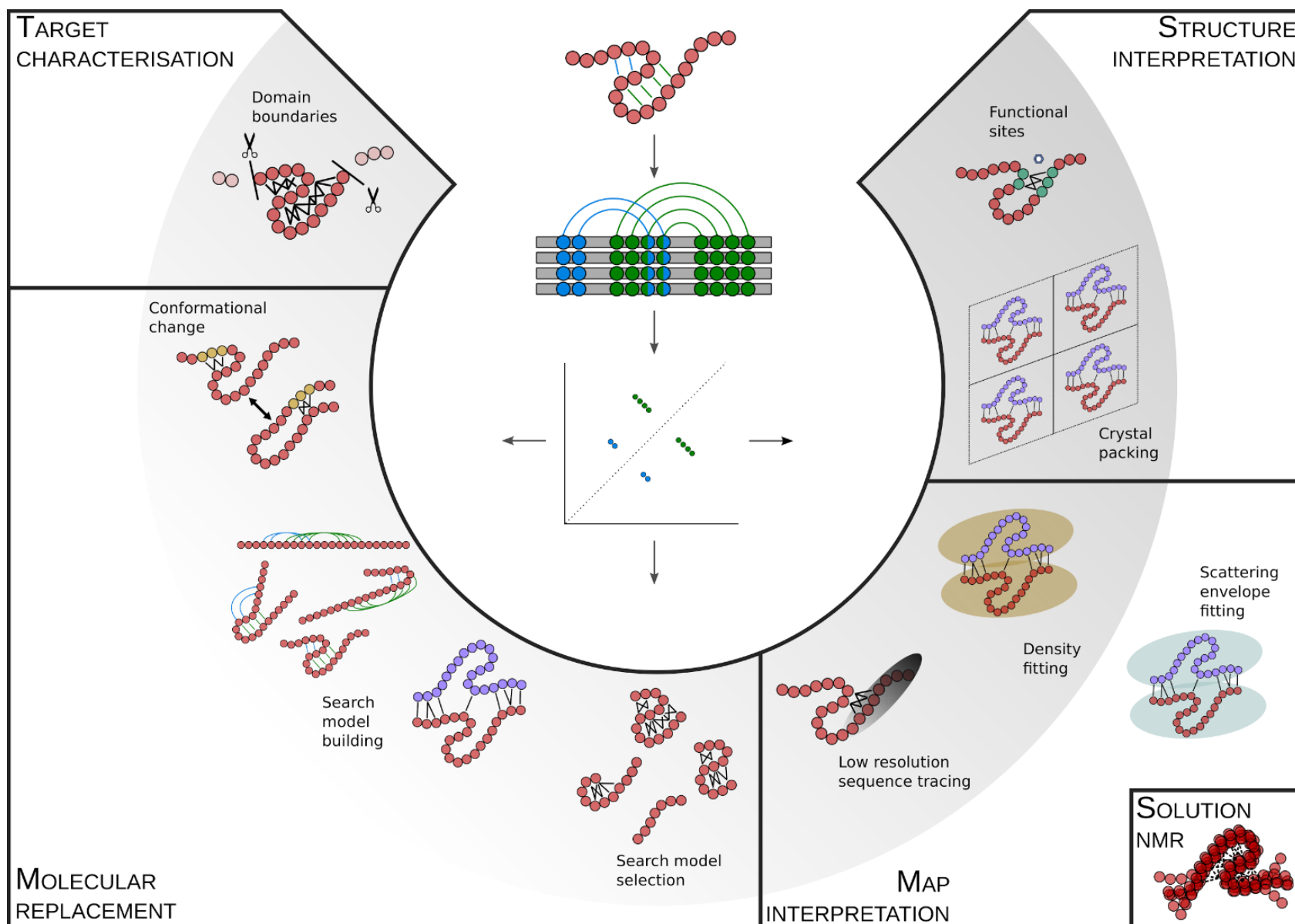
```
>seq1
FASDGITF
DRSLFFGH
```

```
>seq2
NMHKLSDF
PLSERWAQ
```



Intermolecular
contacts
[currently only
reliable for
bacterial proteins in
operons]

ConKit can be used to visualise, analyse and convert contact predictions



Specific value of predicted contacts

- Structures unknown
 - Predicting domain boundaries (ConKit)
 - Making better *ab initio* models for MR (AMPLE)
 - Work with new generation *ab initio* models from databases (AMPLE)
- Structures known
 - Predicting how proteins interact (EVdock, Rosetta)
 - Validating the content of your crystal structure
 - (Predicting functional sites, highlighting conformational states)

Domains and tertiary structure

Which part to express for crystallisation?

Which parts of your crystallised protein might enable phasing by MR, or experimentally?

Recognising folded domains in your sequence

- How novel is your protein target? Recognising distant homology might make it less (or more!) interesting
- You might get extra ideas about its function to guide lab experiments, co-crystallisation, phasing (eg metal-binding sites)
- You might find the whole protein will not express/stay soluble/crystallise etc and want to deal with only part. You might well design construct to exclude disordered regions anyway.
- You might want to parse your protein into domains to explore different MR strategies

Recognising domains by homology with PDB, SCOP, CATH

- (PSI-)BLAST against the PDB might do it

sp|q5i8r5| (257 letters)

RID FRP1NV47014 (Expires on 03-31 21:44 pm)

Query ID gii74681646|sp|q5i8r5|
Description Trypsin-like serine
Molecule type amino acid
Query Length 257

Other reports: [Search Summary](#)

[Graphic Summary](#)

[Show Conserved Domains](#)

Query seq. 

Specific hits
Superfamilies
Multi-domains

install-jalview.exe

[Download](#) [GenPept](#) [Graphics](#)

Chain A, Human Hepsin Protease In Complex With The Fab Fragment Of An Inhibitory Antibody
Sequence ID: [pdb|3T2N|A](#) Length: 372 Number of Matches: 1
[See 1 more title\(s\)](#)

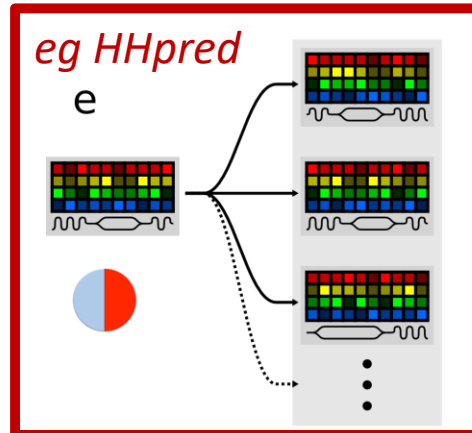
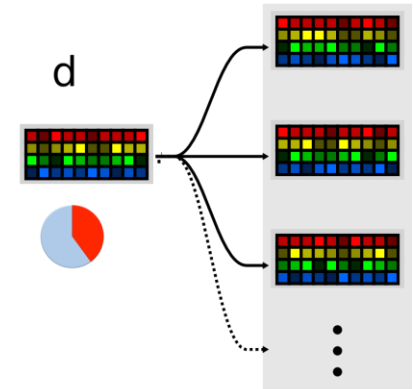
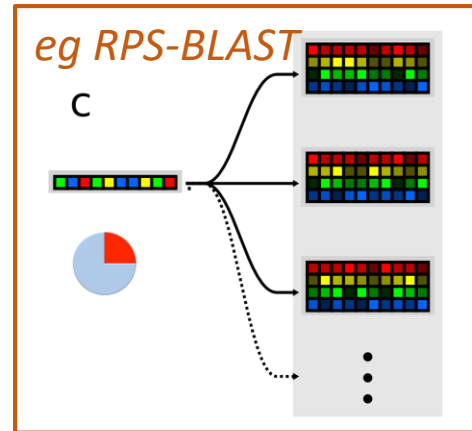
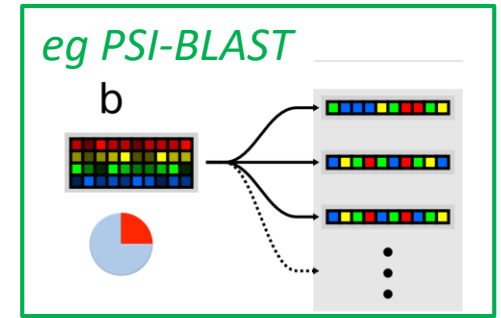
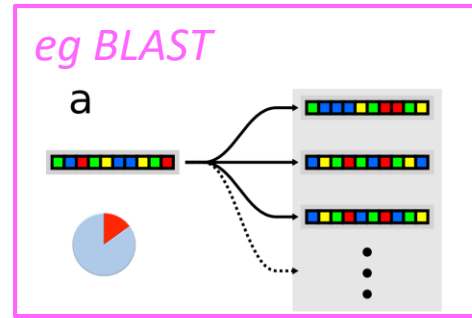
Range 1: 84 to 355 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
142 bits(357)	6e-39	Compositional matrix adjust.	99/281(35%)	140/281(49%)	38/281(13%)
Query 1	M L L S T L S V -----	LGLVTANYKLEKLPSGRIVGGYEVTFKFOYPLIASLEYGSHT	52		
	+LL ++SV	L + + KLP RIVGG + T ++PW SL Y G+H			
Sbjct 84	QRLLEVISVCDPCRGRFLAAICQDCGRRLPVDRIVGGRD-TSLGRWPQVSLRYDGAHL	142			
Query 53	CGGTLYNEKTIISAHC-----	NIGSTSAWSASVHRHDLNEKAEEKESGSNHKII	101		
	CGG+L + +++AAHC	+ + + AS H L +A G			
Sbjct 143	CGGSLLSGDWLTAACHFPERNRVLSRWRFAGAVAQASPHGLQLGVQAVVYHGGYL---	199			
Query 102	ERISHPOYDLN-DDSSNDVSVWKIAAPGNKT---	SGIVLDSGKVSSDGTLLKLVIGWGT	157		
	P D N +++SND+++ ++P T + L + + DG + V GWG T				
Sbjct 200	-----PFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNT	254			
Query 158	TSGGDVSKVLLLEVKVPVFNIDKCKKA--YSTLDTASQFCAGYPEGGKDSCQGDSSGGPIFI	215			
	G + VL E +VP+ + D C A Y FCAGYPEGG D+CQGDSSGGP				
Sbjct 255	QYYGQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGGIDACQGDSSGGPFVC	314			
Query 216	EEKGVAT----LVGVVSWGRGCALKGYPGVYTRVSKVLDFI	252			
	E+ T L G+VSWG GCAL PGVYT+VS ++I				
Sbjct 315	EDSISRTPRWRLCGIVSWGTCALAQKPGVYTKVSDFREWI	355			

Recognising domains by homology with PDB, SCOP, CATH

- Harder cases require a more sensitive tool. I recommend HHpred. Used by **MrBUMP** to find homologues to use as search models



HHpred tips and warnings

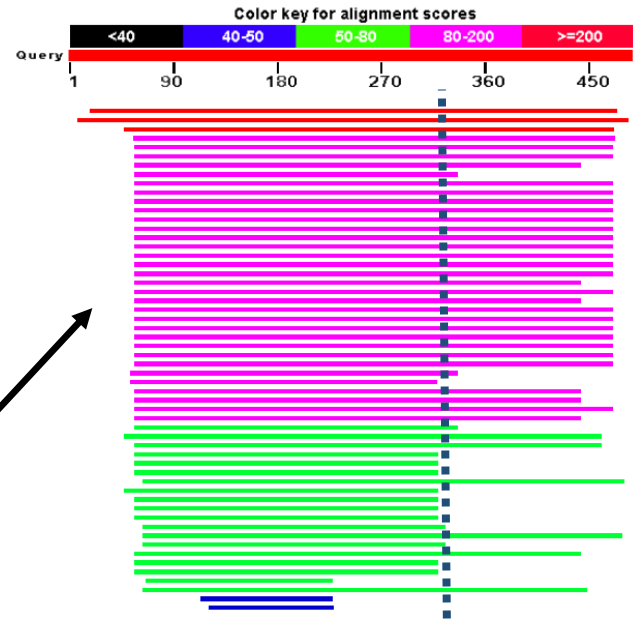
- **Probability** (0-100) is generally a good guide....
- ...but statistics can mislead for **unusual** protein sequences eg coiled-coil, low-complexity, Cys-rich
- Consider if the match makes biological **sense**!
- Look at the matched region
 - Is it a complete domain/structure?
 - If partial, could it reasonably fold?
- Can make reasonable homology models too
- If your query contains multiple domains, '**zooming in**' on particular regions can improve scores and show results not previously seen since
 - Score contains an element favouring similar lengths
 - Only 100 results are shown: can easily get this number for a common domain, making other results invisible!
- **MrBUMP** can search for MR search models with HHpred, most easily at CCP4 online or CCP4Cloud

Recognising matches to domains in **sequence** databases, Pfam, Smart etc

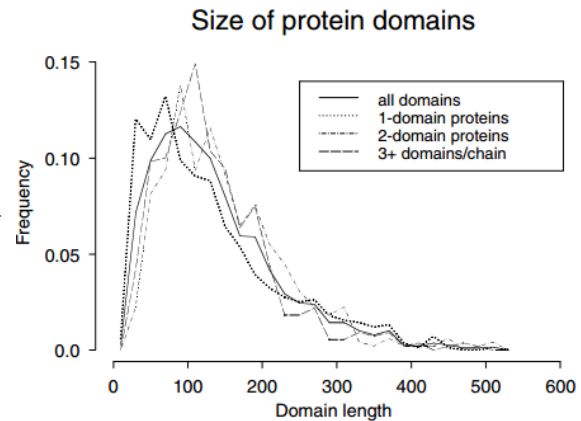
- HHpred is also an excellent, sensitive way to search against these. The same rules for interpretation apply. Matches, even distant ones, can obviously shed light on function
- However, for **structural** purposes need to remember that
 - Many Pfam families don't have structures so that domain limits are much less precise.
 - Many examples where a structure redefines previous Pfam sequence-only domain boundaries
 - Large Pfam entries especially DUFs often turn out to have multiple structural domains
 - Pfam entries for repeats sometimes contain multiple copies

And if the domains
can't be matched?

- *ab initio* approaches to domain boundary identification
 - BLAST matches in sequence databases. Domains are often found in different combinations
 - Secondary structure pattern. PSI-PRED or Jpred4 (faster server)
 - Domain guess by size (server defunct)
 - Contact predictions (ConKit)

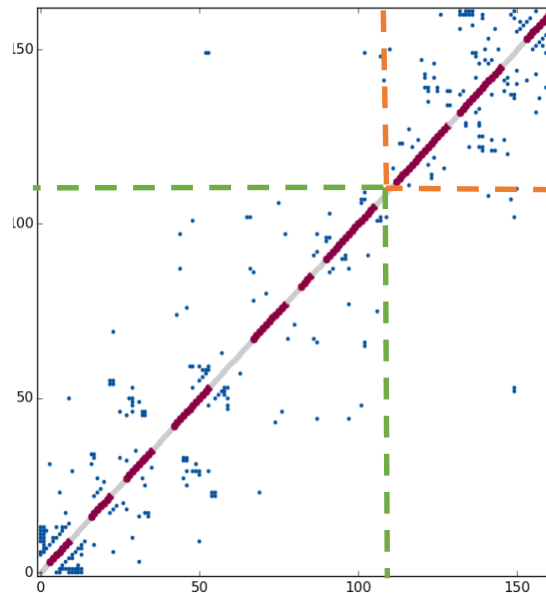
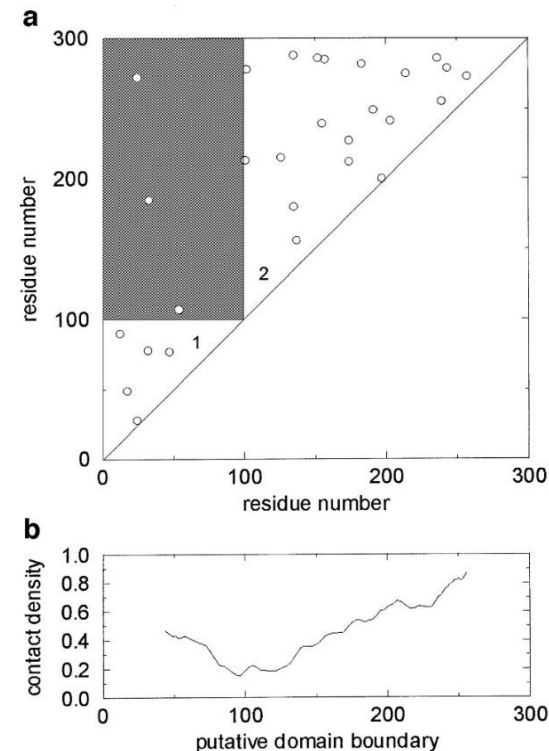


:IDWTHIPKDGLEYNKTIHQIIQENPVEERDRFVMAQLKFLGIEKGFNPNTTEEKKILL EASKVGRAMAQSN DYTKRFTQPYWKGT
 -----HHHHHHHHHHHHHH-----HHHHHHHHHH-----HHHHHHHHHHHHHHHHHHHHHHHH-----
 NWKDAISVSLDQRSENYDELDERAAWFYEAITVS RGMKSLIPGFGQRYLVTVQSDGNWLSGEHTYKLVHPANVPASNFWS TTVYDEI
 -----HHHHHHHHHHHH-----EEEEEEEE-----EEEE-----EEEEEEEE-----
 JRLMIINDAGSPDISSRKNLKVNSDGSIDVYYPGPKPVKGYENNWVC
 -----EEEE-----EE

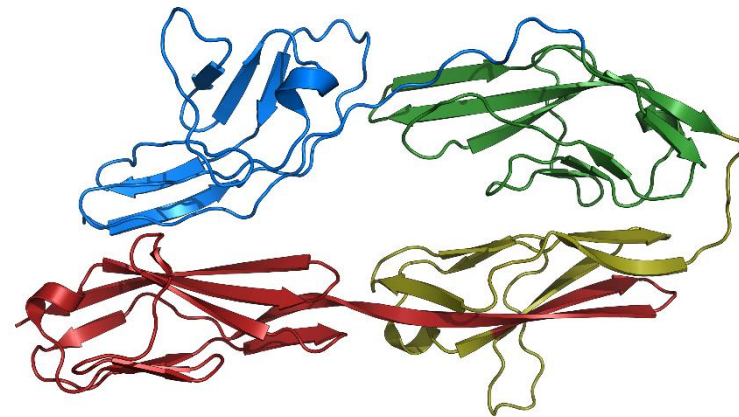
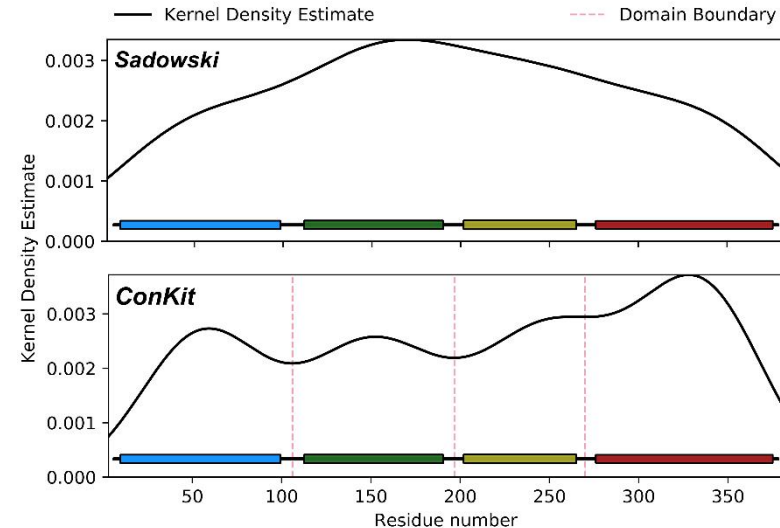


Predicted contacts for defining domains

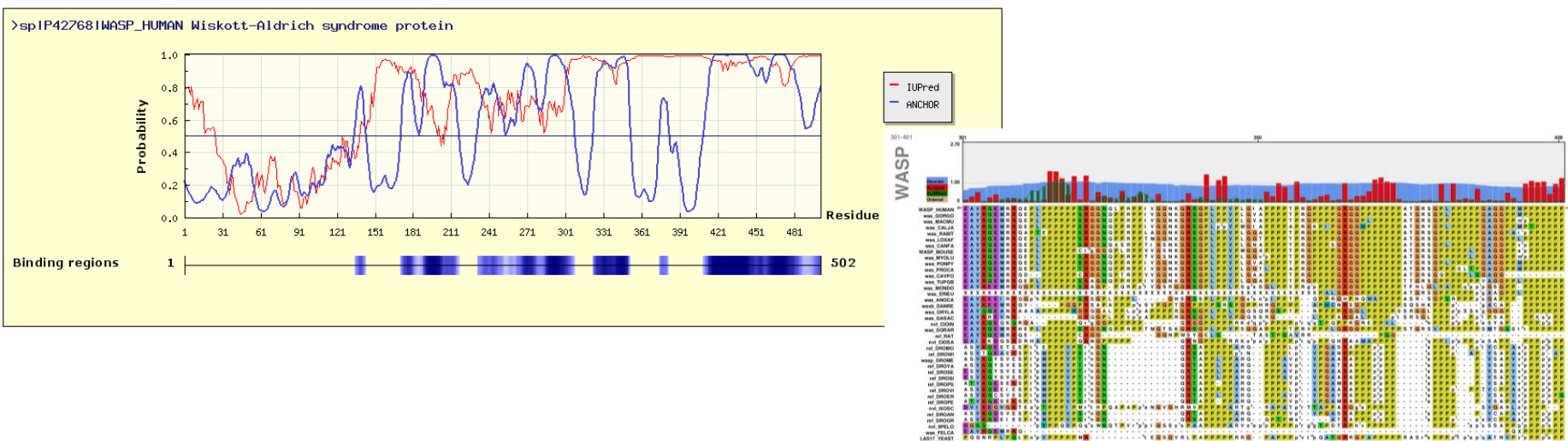
- Domain boundaries required for individual expression *in vitro* and helpful for fold recognition and modelling *in silico*



DUF4158 – two helical domains?



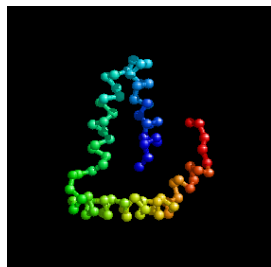
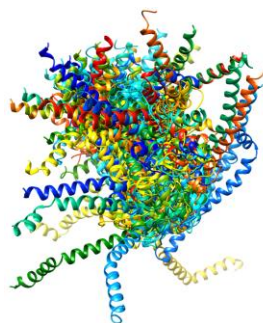
- Not all proteins and protein regions fold into stable structured domains. ID proteins and regions will not crystallise (alone)
- There are many predictors, all performing roughly equally well
- I recommend IUPred (fast) and MetaDisorder (slow but good)
- Can also look for short interaction motifs in ID regions (ANCHOR, SlimPred)



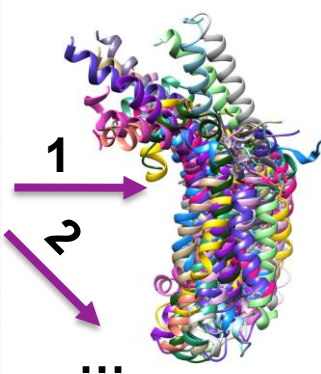
Building **ensemble** search models with



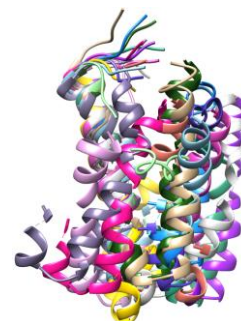
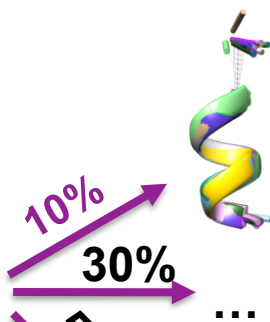
ab initio
models
(ROSETTA/
QUARK)



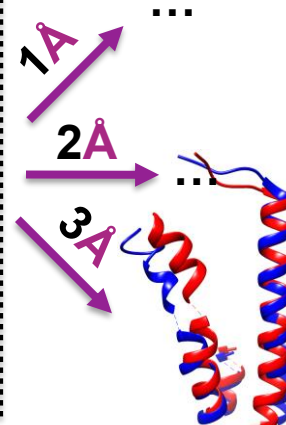
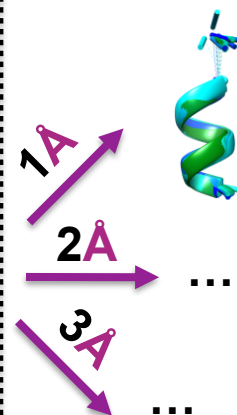
cluster
(SPICKER)



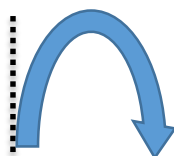
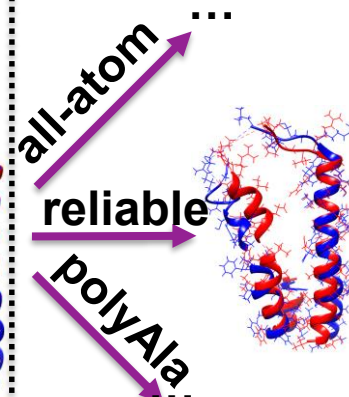
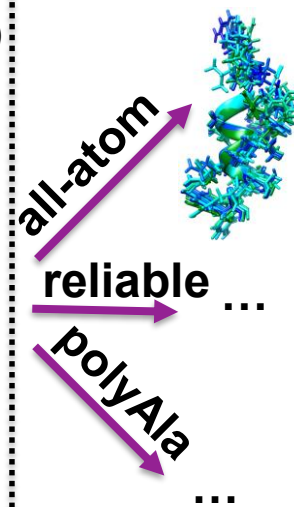
truncate
(THESEUS)



cluster
(MAXCLUSTER)



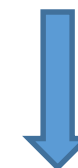
side chains



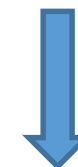
MrBUMP



Phaser

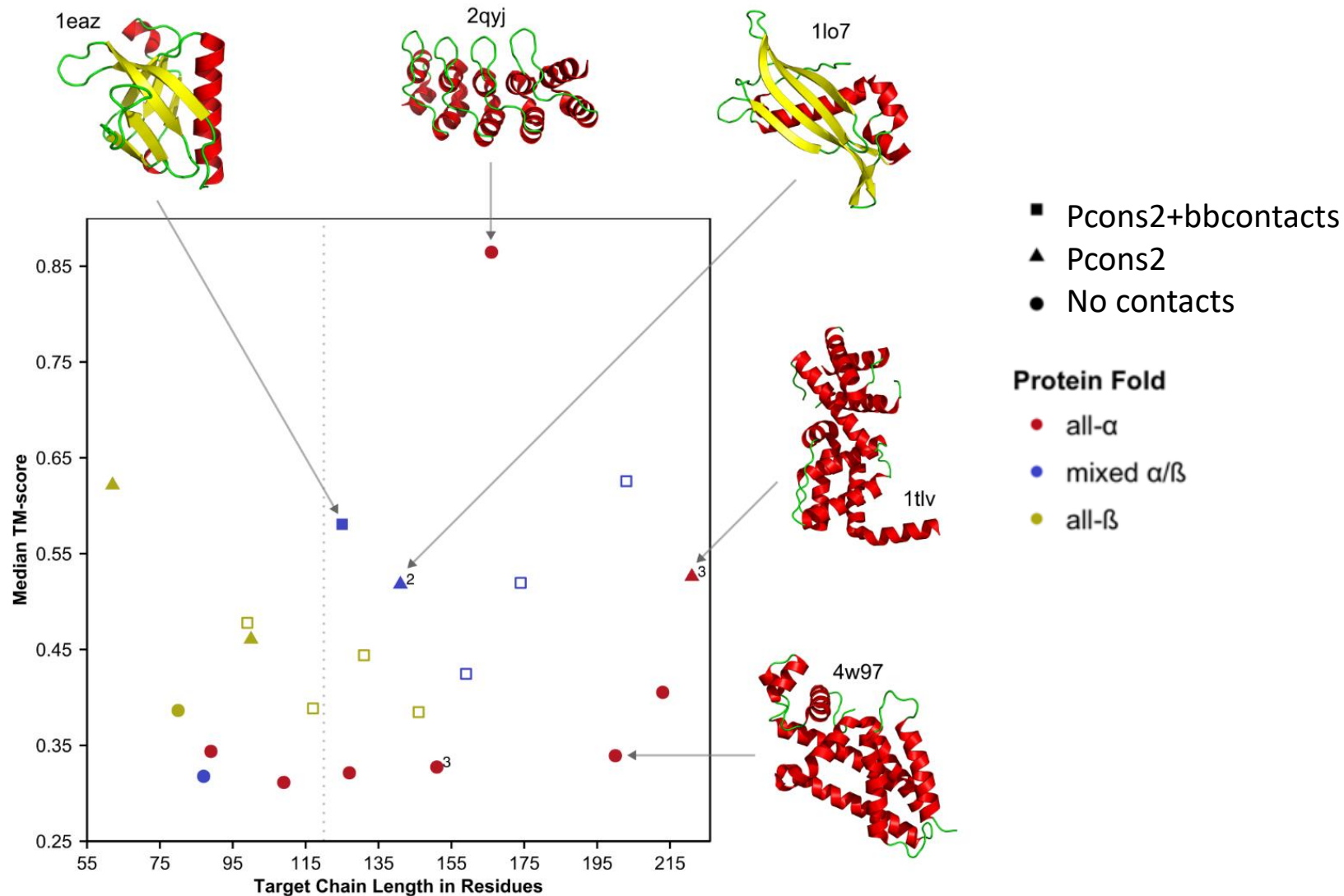


ShelxE



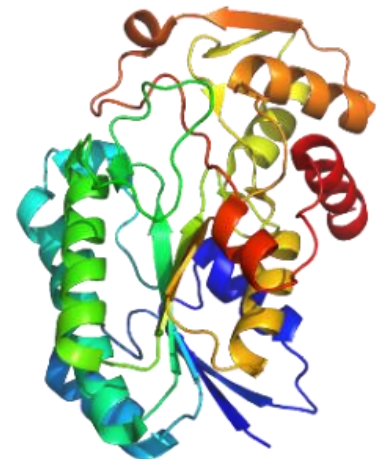
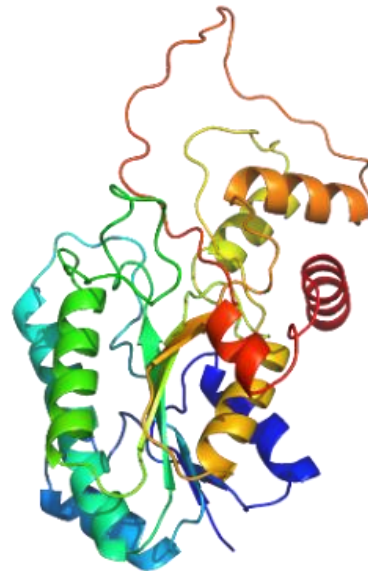
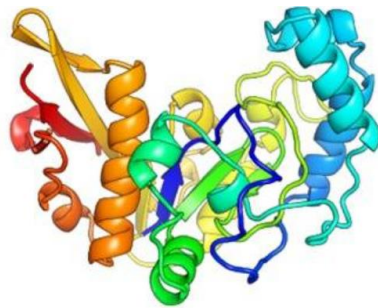
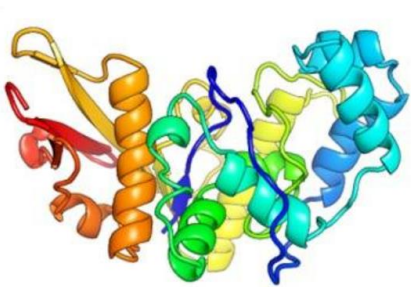
**Buccaneer
and/or
ARP/wARP**

With contact predictions, **AMPLE** and ROSETTA produces models that succeed with larger and more β -rich proteins



Using *ab initio*, contact-driven models from databases

- GREMLIN metagenome repository
- Models for 614 Pfam families (predicted high quality)
- 30 best models per family, or single best
- Scored according to predicted reliability
- gremlin2.bakerlab.org/meta.php
- Pfam and InterPro plan to integrate models made this way
- PConsFam database
- Models for 13,617 Pfam families
- Single best model per family
- Scored according to predicted reliability
- pconsfam.bioinfo.se



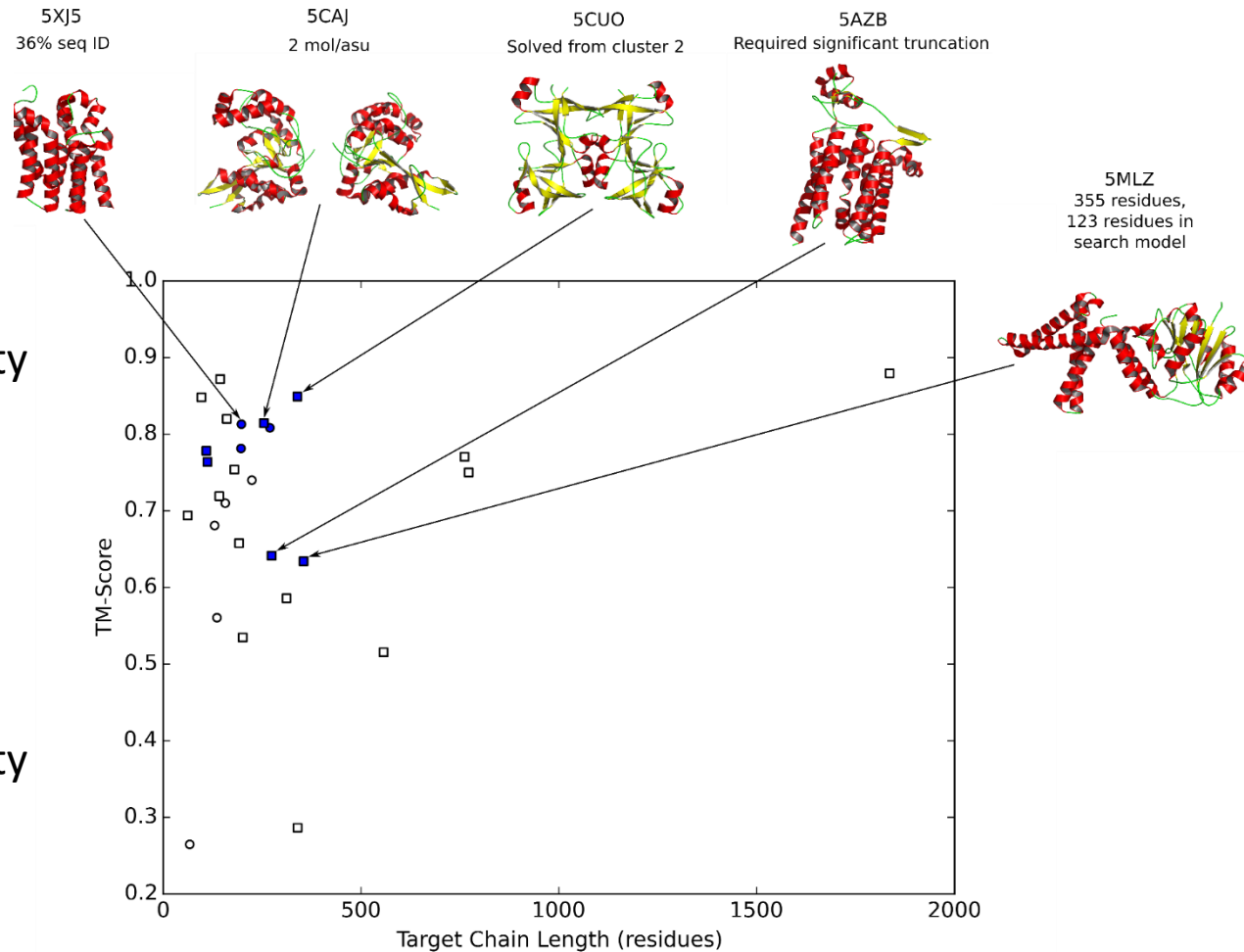
AMPLE results with GREMLIN database deposits

- Targets

- 1-12 mol/asu,
- 1.5-3 Å resolution,
- 62-1836 residues,
- 36-100% sequence identity to modelled protein

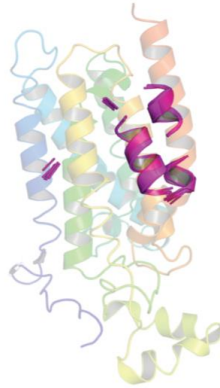
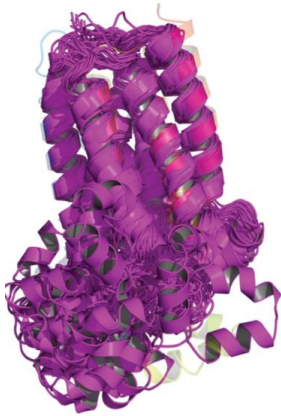
- 9/27 succeed

- 1-3 mol/asu,
- 1.5-2.9 Å resolution,
- 109-355 residues,
- 36-100% sequence identity to modelled protein)

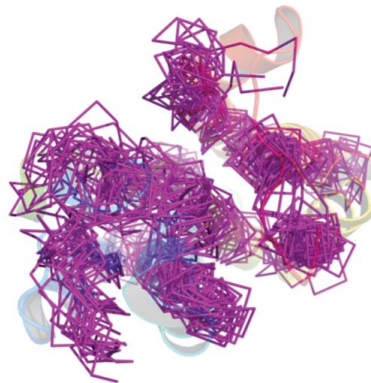


AMPLE working on *ab initio* model database deposits

GREMLIN



PCONSFAM



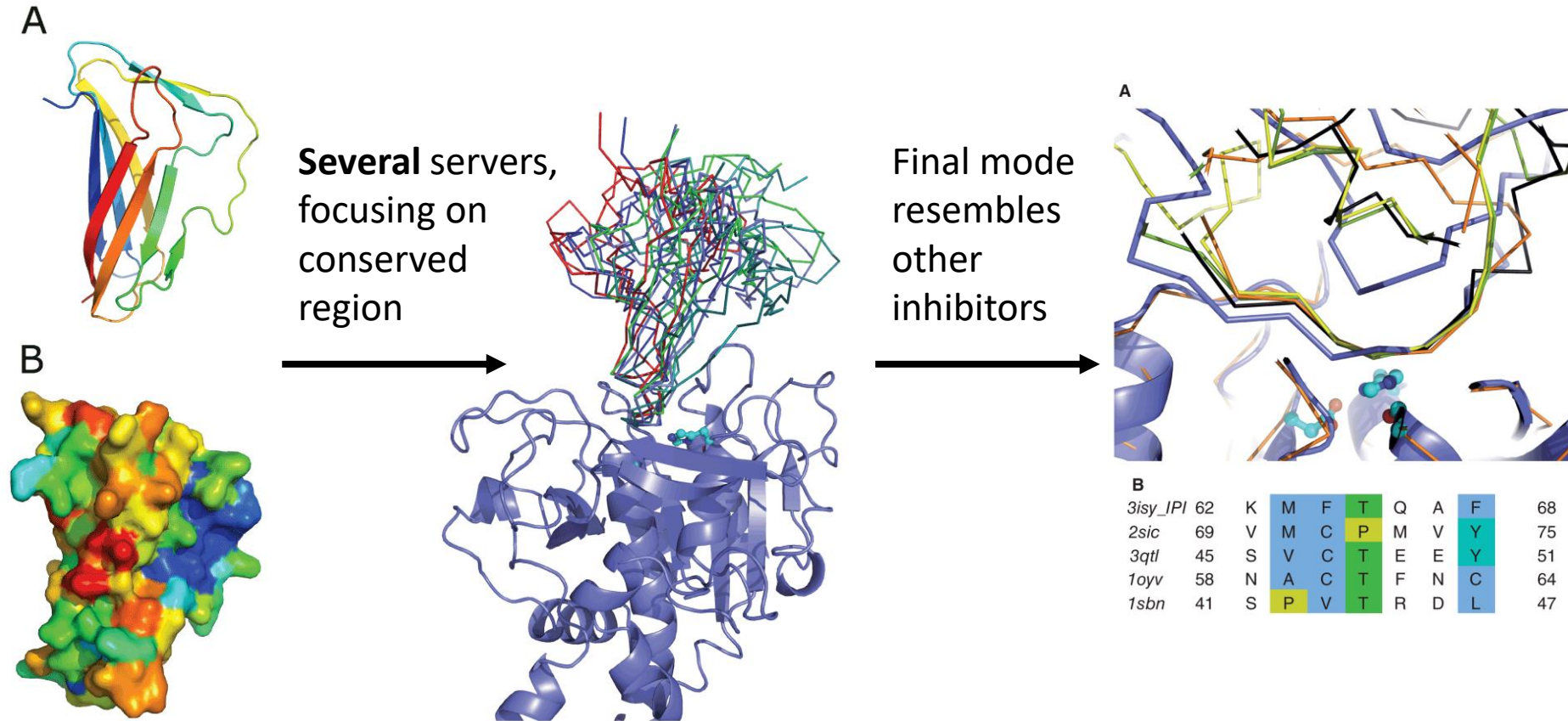
Quaternary structure

Intermolecular interactions

Predicting protein-protein interactions

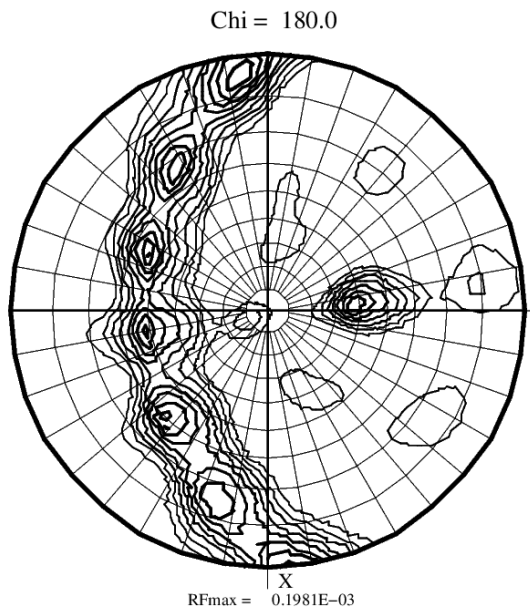
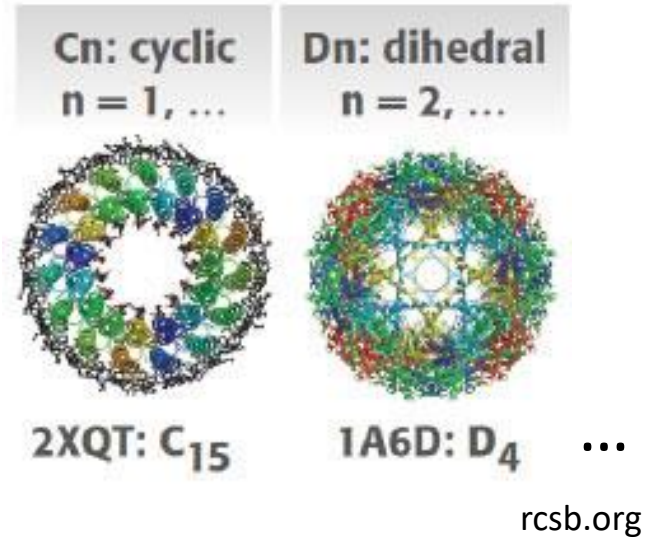
- Relevant to MR eg proteins A and B are cocrystallised but neither alone solves. An accurately predicted complex, being larger, might solve
- Many methods predict complexes based on steric complementarity plus other scoring functions
- Recommendable servers include
 - ClusPro, the best performing docking method
 - Haddock, which has a good server with different modes
 - Each allows inclusion of other information eg predicted interface residues
 - Symmetric docking at ROSIE server

Multiple methods in bioinformatics: *B. subtilis* IPI docking to protease

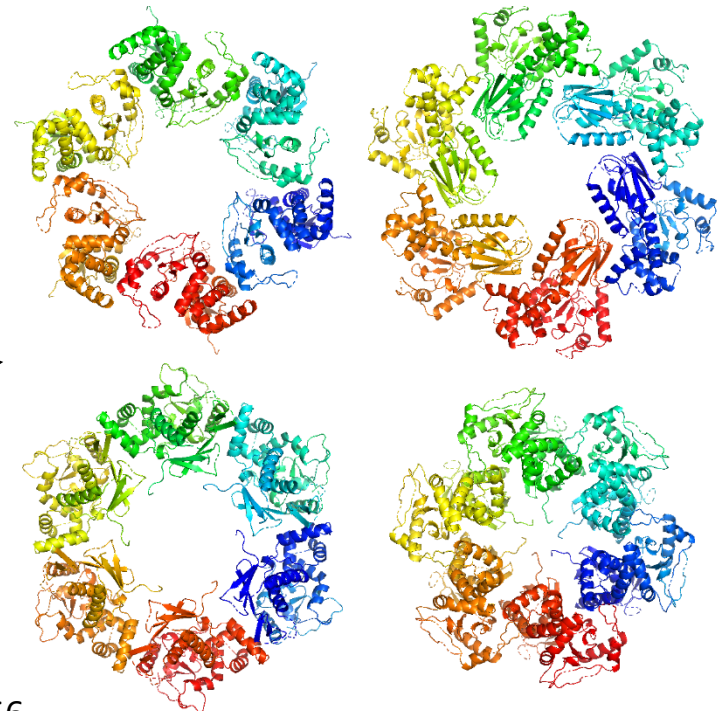


ROSIE symmetric docking for oligomers

- Only cyclic (C_n) or dihedral (D_n) symmetry at server
- Clues from self rotation function may be available

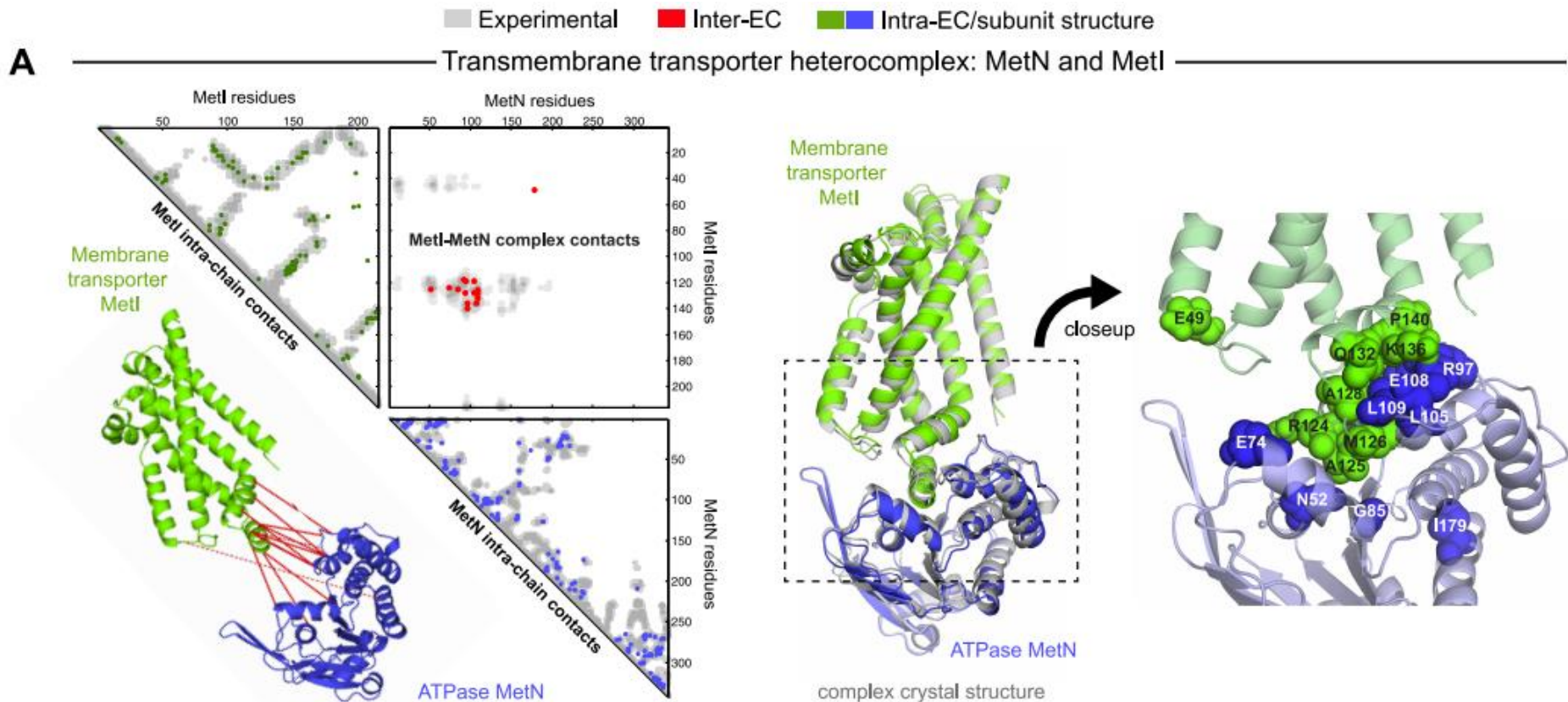


Generate
hexamers

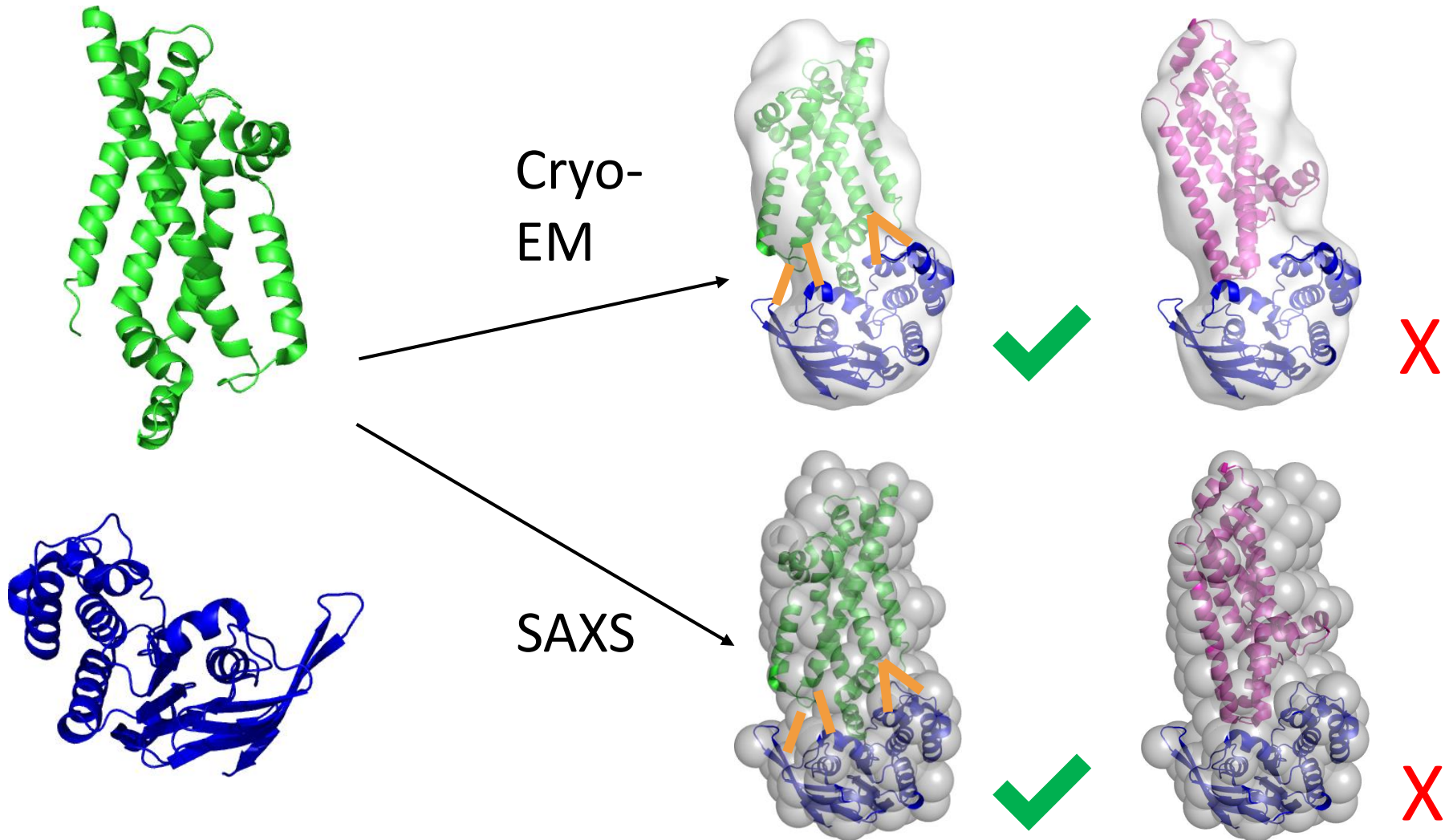


Andre et al. (2009) PNAS 104:17656

Using predicted contacts to help predict complexes



Fitting to low-resolution envelopes



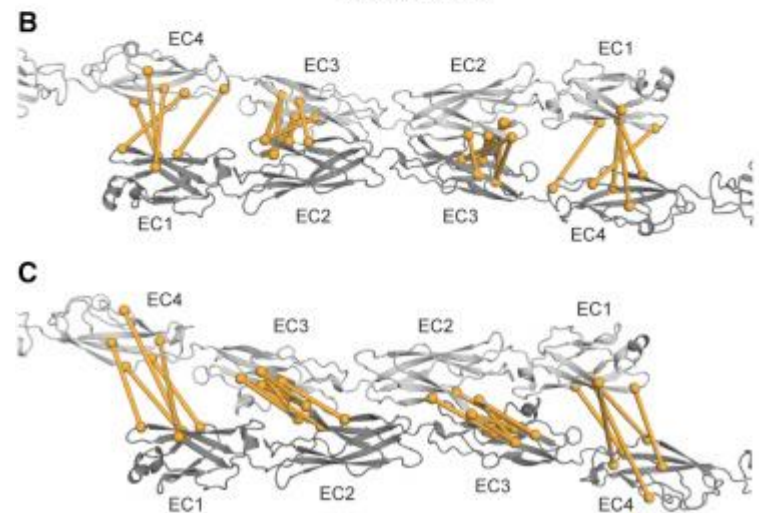
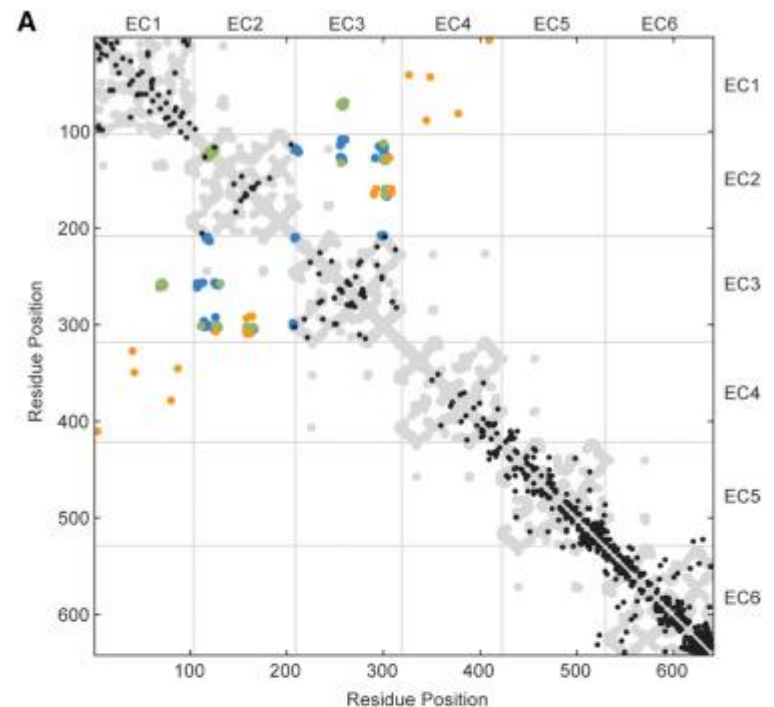
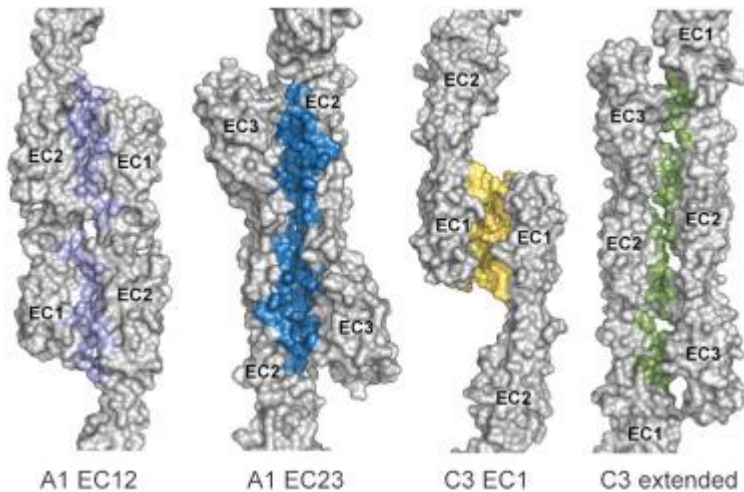
Structure-based function interpretation

What is the biologically relevant quaternary structure?

Where are the functional/catalytic sites?

Validating crystal structure contents

- **PISA** is an excellent general method, but contact predictions help in some cases
- Crystal showed various ways in which protocadherins could interact
- Predicted contacts supported two of the four modes

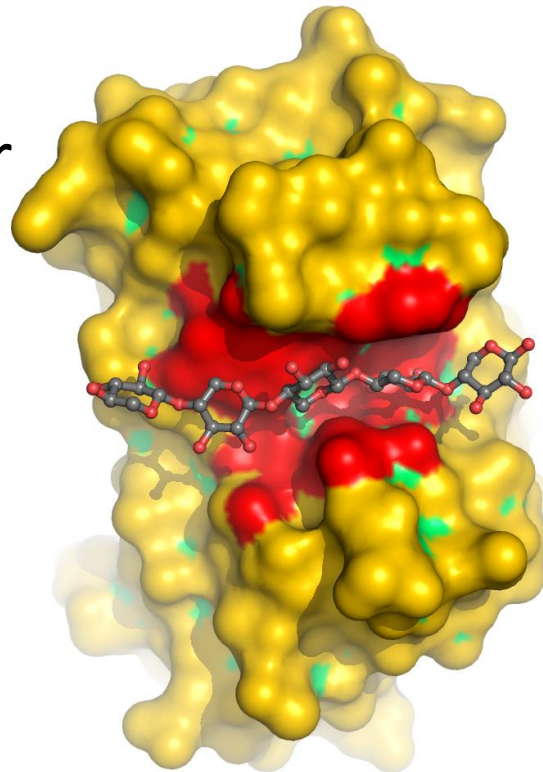


Some lesser-known structure-based function annotation methods

- Finding functional sites is based on their being different somehow to the rest of the protein surface. Important generic methods are based on
 - Shape (castP, ProFunc, PyMOL)
 - Electrostatics (PyMOL, APBS)
 - Evolutionary conservation (Consurf)
- Less well-known but valuable characteristics are
 - Statistics of surface atom 'triangles' (STP)
 - Probe interaction energetics (ISMBLab)
 - Predicted pKa values (THEMATICS/POOL)

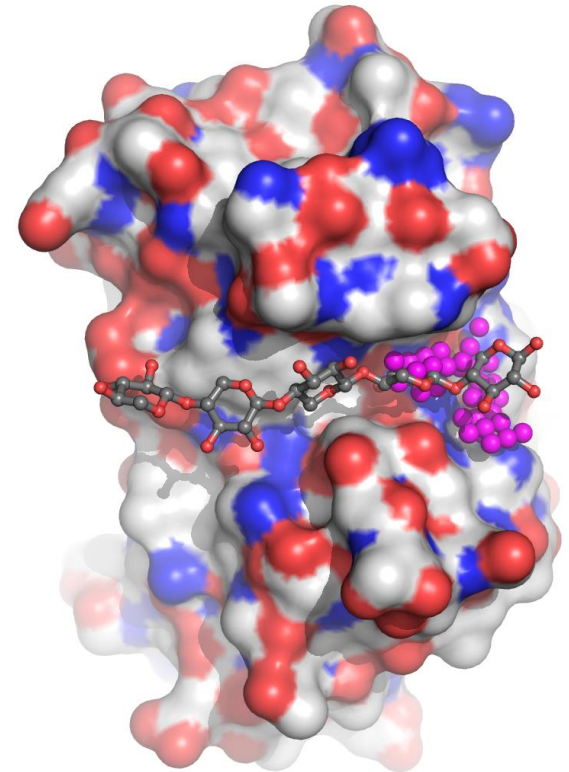
Different probes for different binding sites

- Hydroxyl group can be used to probe for carbohydrate binding sites
- Phosphate oxygen used for binding sites of phosphorylated ligands



ISMBLab

ismblab.genomics.sinica.edu.tw

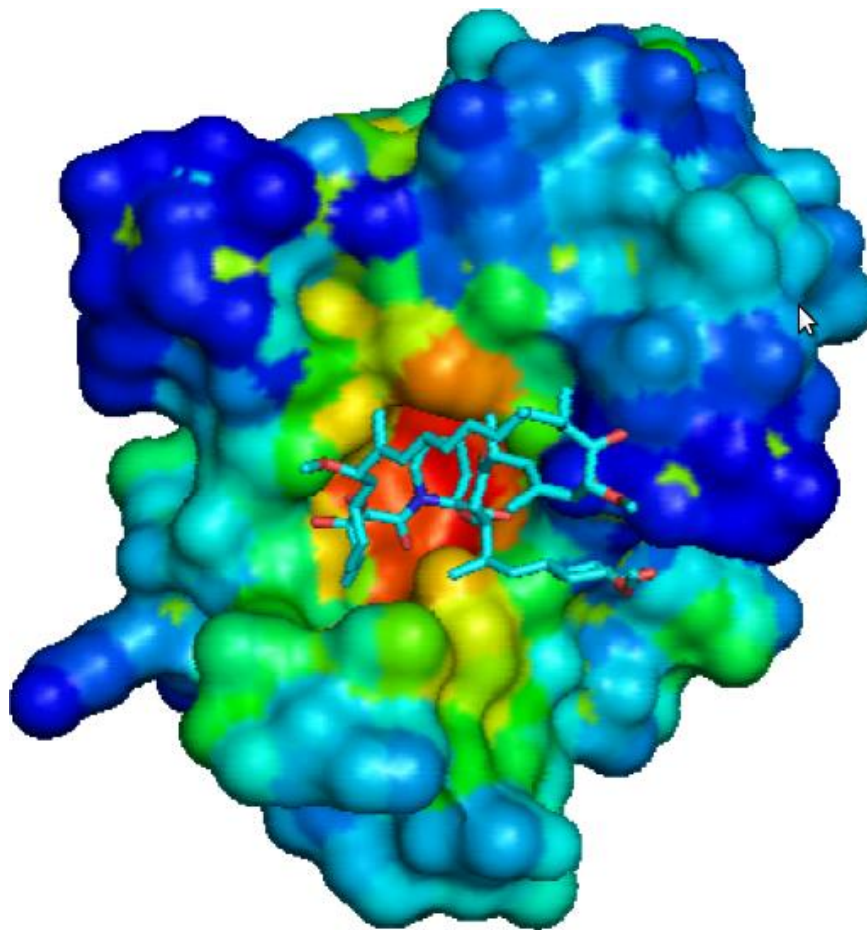


SiteHound

scbx.mssm.edu/sitehound

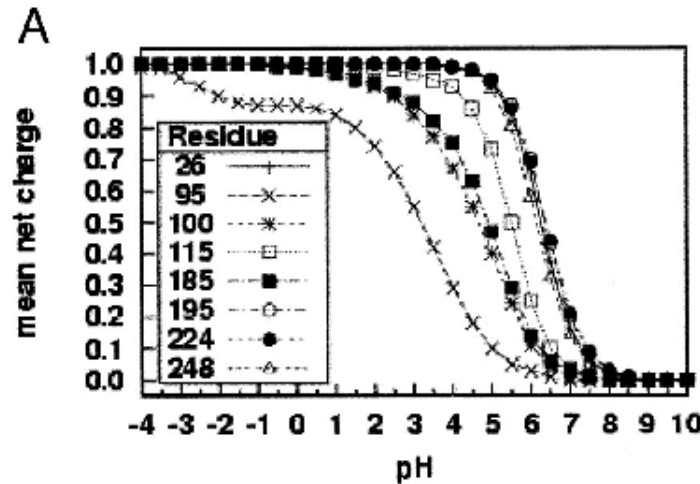
Binding sites from statistics

- STP (surface triplet propensities)
- 13 atom types $\rightarrow$ 455 triplets
- Distribution in binding vs non-binding sites varies
- Designed for small molecules, works on PPIs, including flat surfaces

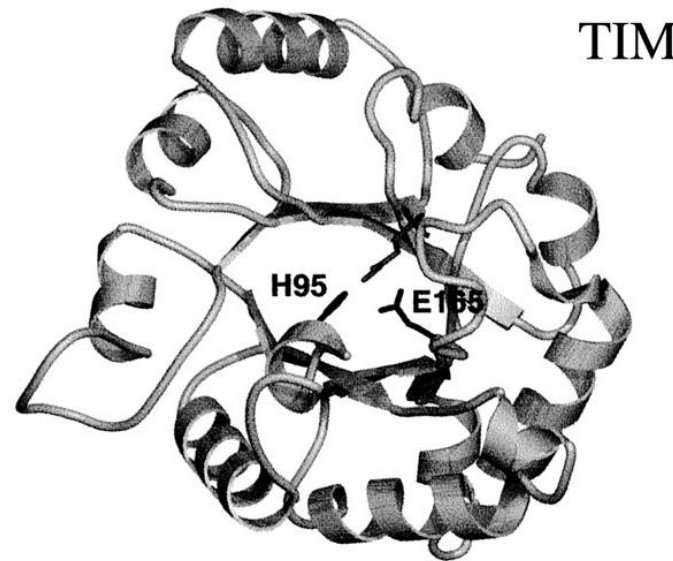


Theoretical microscopic titration

- Computer analysis of a reliable protein structure can predict pKa values for acids and bases. Residues with perturbed pKa values are possible catalytic residues, especially if clustered.



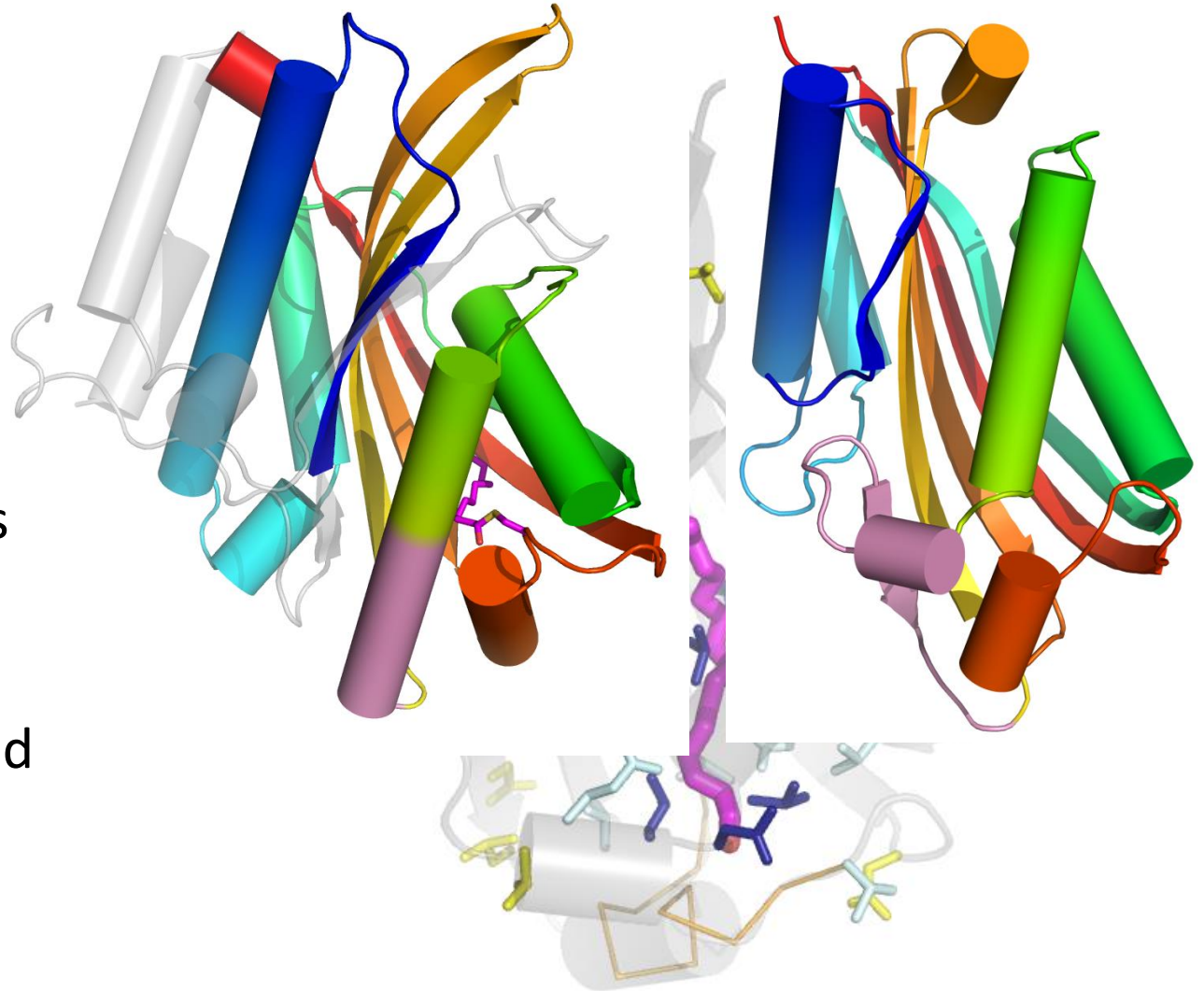
pKa of His95 is atypical compared to other His residues in enzyme



His95 and other residues with atypical pKa cluster at catalytic site

Multiple methods in bioinformatics: Structure comparisons of Evf

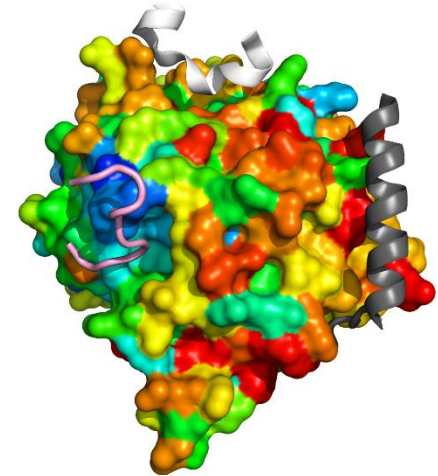
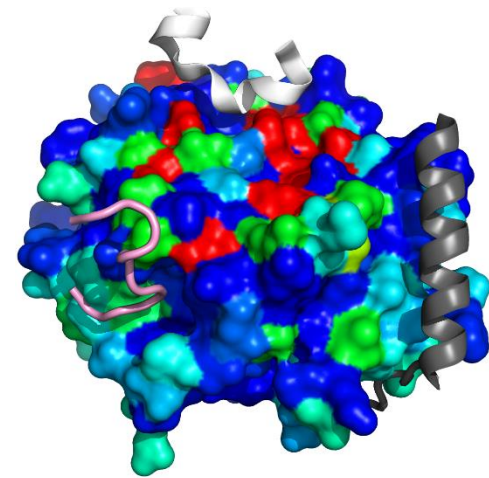
- Reported as novel fold...
- ... but in fact related to *Bacillus* toxin structures
- Both bind to host insect membranes
- Palmitate seen in Evf structure. Matches conserved region of toxins...
- **GESAMT** is an excellent CCP4 option



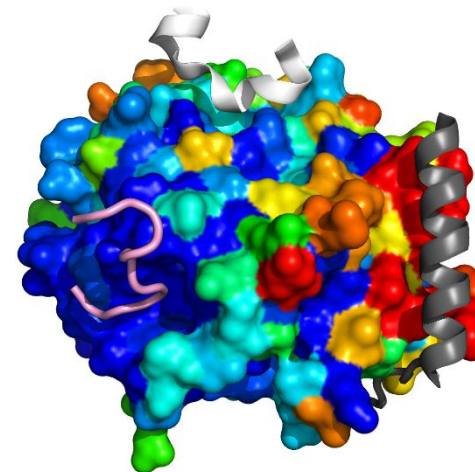
Some servers require thought...

- Consurf maps sequence conservation onto a structure revealing functional sites
- Excellent, general method, but results depend on sequence set chosen for mapping: selecting all or only near relatives gives different results. Either might be more appropriate for you

Mapping 300 homologues mixes different activities so no information on binding sites



But restricting to a single protein family shows only 'pink' site is function in both Diptera and Lepidoptera



...and finally, you're putting a
manuscript together

Calculating and presenting sequence alignments

Your sequence alignment

- Don't use ClustalW! It's 22 years old! Modern methods like MUSCLE, Probcons and MAFFT are much better

ClustalW misses relatively obvious RHG motif in some of diverse sequence set...

Sequence alignment from ClustalW. The alignment shows 10 sequences from position 150 to 230. Two black ovals highlight gaps in the alignment at positions 160 and 220, where the 'RHG' motif is present in some sequences but missed in others.

```
150      160      170      180      190      200      210      220      230
ji|548534|sp|P00950|PMG1_YEAST/1-487---MPKLVLYRHGSEWNEK----NLFTGWV-DVKLSAKGQQ---EAARAGELLKEKKVYPDV---LYTSKLSRAITQANIALEKADR
ji|3024420|sp|P96121|GPMA_TREP/1-487---MKLVLI RHGSEWNEK----NLFTGWT-DVPLTPRGES---EAQEGGRVLQEAGDFDL---CYTSY LKRAIRTLNFVLQALDR
ji|12751461|gb|AAK07665.1|/1-487-----TRHGEIKWVE---RRMQGWQ-DSPLTEKGRQ---DAMRLCKRLEA--VELAA---IYTSYSGRALTAIEIVRGCR--
ji|1169587|sp|P32604|F26_YEAST/1-487TSYENSEPHVAADFLEK---IRQYERF-YEPLDQKDK---DMTFIKLVNIEEVVINK---IRTVLESRIVYMMNIRPKPKY
ji|1730554|sp|P52086|COBC_ECOL/1-487---MRLWLI RHGSEWNEK----GLYSCHA-PTPLTARGIE---QAQNLHTLLHG--VSFDL---VLCELEAQTARLVLSDRQ-
ji|3183165|sp|P76502|SIXA_ECOL/1-487---MQVFIMRHGDAALD-----AASDS-VRPLTTNGCD---EIRLMANWLKQKVEIER---VLVSPLRAEITLVEEGDCLN-
ji|2895490|gb|AAC38954.1|/1-487---LELPVDICIRHGTQGNTEPRVFQGVQDYA--NNQLTQQCQQQAAAAATKLEAMAAAKEFI PDL---LLSSPLLRAVHTAQPFVDANK
ji|15229917|ref|NP_187168.1|/1-487---KLLPKRIILVRHGESEGNLDT-AAYTTTPDH-KIQLTDSGLLQAQEGARLHALISSNPSSPEWRVYFVVSVDRTFSTLREIGRSFSR
ji|16130187|ref|NP_416755.1|/1-487---MLAFCRSLKSKYIIILLALAAIAGLCTHAAWSNGLR--INKTLARLAQHPVVVLFV-----FIERDORSTNQCLSDKTK
```

... but MUSCLE gets it

Sequence alignment from MUSCLE. The alignment shows 10 sequences from position 220 to 300. Two red ovals highlight the 'RHG' motif at positions 230 and 290, which is correctly aligned across all sequences.

```
220      230      240      250      260      270      280      290      300
ji|2895490|gb|AAC38954.1|/1-537-----LELPVDICIRHGTQGNTEPRVFQGVQDYANQLTQQCQQQAAAAATKLEAMAAAKEFI PDL---LLSSPLLRAVHTAQPFV
ji|1730554|sp|P52086|COBC_ECOLI/1-537-----MRLWLI RHGSEWNEK----GLYSCH--HAPLPTARGIEQAQNLHTLLHGVSF-----LV--LCSELEAQTARLV
ji|12751461|gb|AAK07665.1|/1-537-----TRHGEIKWVE---RRMQG---WQDSPLTEKGRQDAMRLCKRLEAVELA-----AI---YTSTSGRALETAIE-IV
ji|548534|sp|P00950|PMG1_YEAST/1-537-----MPKLVLYRHGSEWNEK--NLFTG---WVDVKLSAKGQQEAAARAGELLKEKKVY---PD-VL---YTSKLSRAITQAN-IA
ji|3024420|sp|P96121|GPMA_TREPA/1-537-----MKLVLI RHGSEWNEK--NLFTG---WTDVPLTPRGESAEQEGGRVLQEAGFD---FD-LC---YTSFLKRAIRTLN-FV
ji|16130187|ref|NP_416755.1|/1-537-----PVVILFRHAERCDRSTNQCLSD-----KTGITVKGTQDARELQNAFSADI-----PDFDL---YSSNVRTIISA----
ji|3183165|sp|P76502|SIXA_ECOLI/1-537-----MQVFIMRHGDAALDAASDSVR-----PLTTNGCDESRLMANWLKQKVE---IE-RV---LVSPFLRAEITL---
ji|15229917|ref|NP_187168.1|/1-537-----KLLPKRIILVRHGESEGNLDTAAYTT-TDHHKIQLTDSGLLQAQEGARLHALISSNPSSPEWRVYFVVSVDRTFSTLREIG
ji|1169587|sp|P32604|F26_YEAST/1-537YMMNIRPKPKYITWLSRHGESEYVVE-KKICG-----DSSLSEKGFYAKKLEQLVKESAGE-----INLTV---WTSTLRTQTAN-YL
```

Jalview.org, recommended for sequence alignments

- All these alignment methods and more are available through Jalview on Dundee servers

Jalview 2.9.0b2

File Tools Vamsas Help Window

MAFFT Multiple Sequence Alignment of Retrieved from Uniprot

File Edit Select View Annotations Format Colour Calculate Web Service

Original Phosphorylation Site Predictions MAFFT Alignment Order

FER_CAPAA/1-97 -ASYYKVLITPDGPIEFDCDDVYILDQA
FER_CAPAA/1-144 MASYYKVLITPDGPIEFDCDDVYILDQA
FER1_SOLLQ/1-144 MASYYKVLITPDGPIEFDCDDVYILDQA
Q93XJ9_SOLLQ/1-144 MASYYKVLITPDGPIEFDCDDVYILDQA
FER1_PEA/1-149 MASYYKVLITPDGPIEFDCDDVYILDQA
Q7XA98_TRIPR/1-152 MATYYKVLITPDGPIEFDCDDVYILDQA
FER1_MESOR/1-148 MAAYKVLITPDGPIEFDCDDVYILDQA
FER1_SPIOL/1-147 MAAYKVLITPDGPIEFDCDDVYILDQA
FER3_RAPSA/1-96 -ATYYKVEITPDGPIEFDCDDVYILDQA
FER1_ARATH/1-148 MATYYKVEITPDGPIEFDCDDVYILDQA
FER_BRANA/1-96 -ATYYKVEITPDGPIEFDCDDVYILDQA
FER2_ARATH/1-148 MATYYKVEITPDGPIEFDCDDVYILDQA
Q93Z60_ARATH/1-118 MATYYKVEITPDGPIEFDCDDVYILDQA
FER1_MAIZE/1-150 QATYYKVLITPDGPIEFDCDDVYILDQA
O80429_MAIZE/1-140 QATYYKVLITPDGPIEFDCDDVYILDQA

Secondary Structure 1 2 3 4 5
Iron Sulphur Contacts Fe Fe Fe
Conservation 0 7 6 6 9 7 3 6 7 7 9 5 6 8 9 3 5 9 9 9 3 7 3 9
Quality
Consensus A Y Y K V L I T P D G P I E F D C D D V Y I L D Q A E E L G D L P Y S C R A G C S S C A G K V V S G S V

Sequence 1 ID: FER_CAPAA Residue: PRO (36)

Alignment

- Secondary Structure Prediction
- Protein Disorder
- Analysis
- Conservation
- Fetch DB References
- Muscle with Defaults
- MAFFT with Defaults
- MSAProbs with Defaults
- GLProbs with Defaults
- Clustal
- ClustalO

Align with default settings

Average distan...

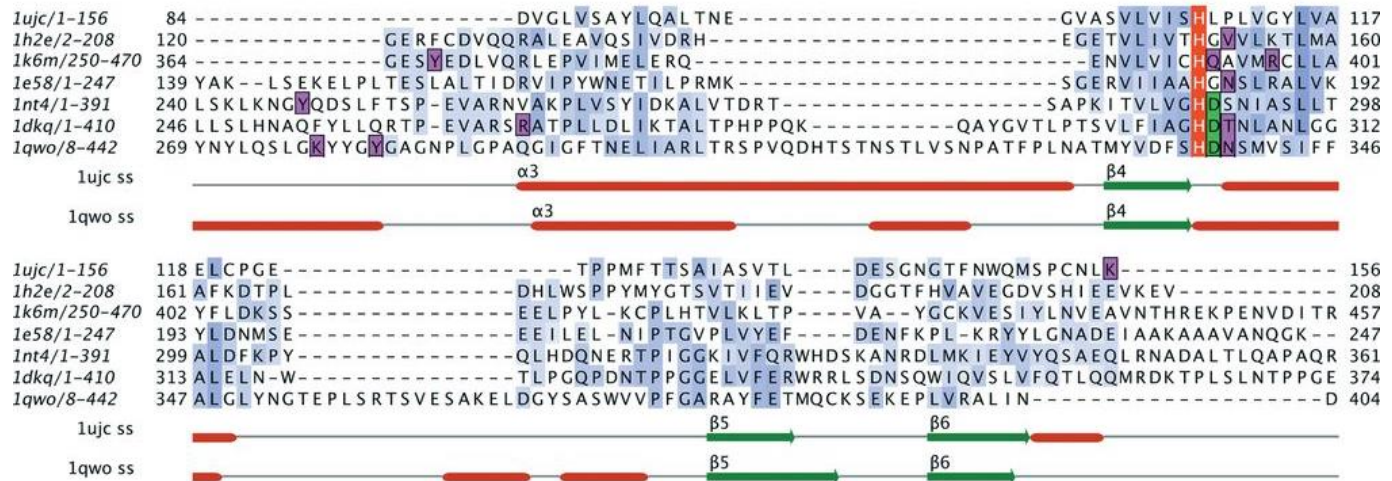
File View

3D view for FER1_S...

THR 89

Jalview

- Also helps you produce figures like this...



- ... rather than like this



Questions? Feedback?

