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)
GYYELYRRSTIGNSLVDALDTLISDG
RIEASLAMRVLETFDKVVAETLKDNT
QSKLTVKGNLDTYGFCDDVWTFIVKN
CQVTVEDQSVISVDKLRIVACNSKKS
```

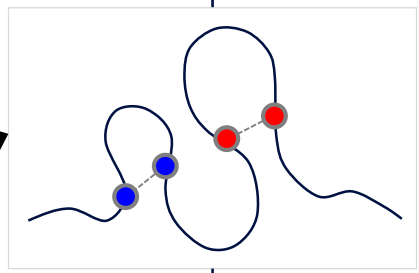
Multiple Sequence Alignment of homologous sequences

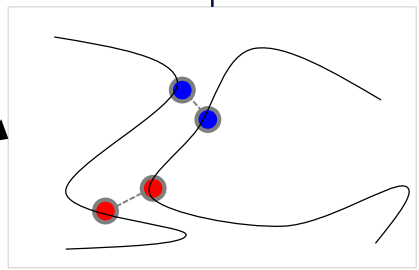
```
...KDNTQSKLTVKGNLDTYGFC...
...KDNTQSKLTVKGNLDTYGFC...
...KNDANIAKLKLTFKGDLDTY...
...RENTQSKLNVKGLLDTYGFC...
...KNENIVGKSKLIFKGDLDTY...
...KNESNIAKSKLTFKGDLDTY...
...KHESNIAKSKLTFKGDLDTY...
...KNESGISKSKLTFKGDLHTY...
...KDNTQSKLTVKGLLDTYGFC...
...KSESGIAKSKLTFKGDLHTY...
```

```
...KDNTQSKLTVK NLDTYGFC...
...KDNTQSKLTVK NLDTYGFC...
...KNDANIAKLKL FKGDLDTY...
...RENTQSKLNVK LLDTYGFC...
...KNENIVGKSKL FKGDLDTY...
...KNESNIAKSKL FKGDLDTY...
...KHESNIAKSKL FKGDLDTY...
...KNESGISKSKL FKGDLHTY...
...KDNTQSKLTVK LLDTYGFC...
...KSESGIAKSKL FKGDLHTY...
```

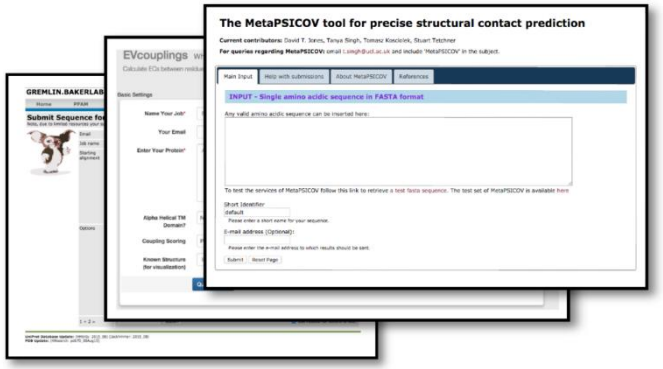
Direct Coupling Analysis of large MSAs plus Deep Learning predicts contacts

Applications…





```
>seq
FASDGITF
DRSLFFGH
```

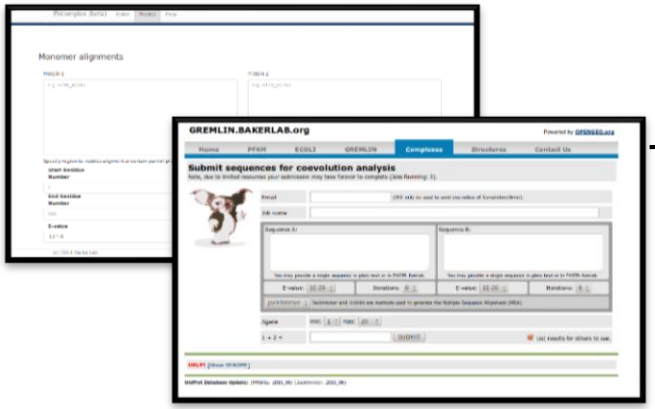


Intramolecular contacts

Intermolecular contacts if a **homo**oligomer

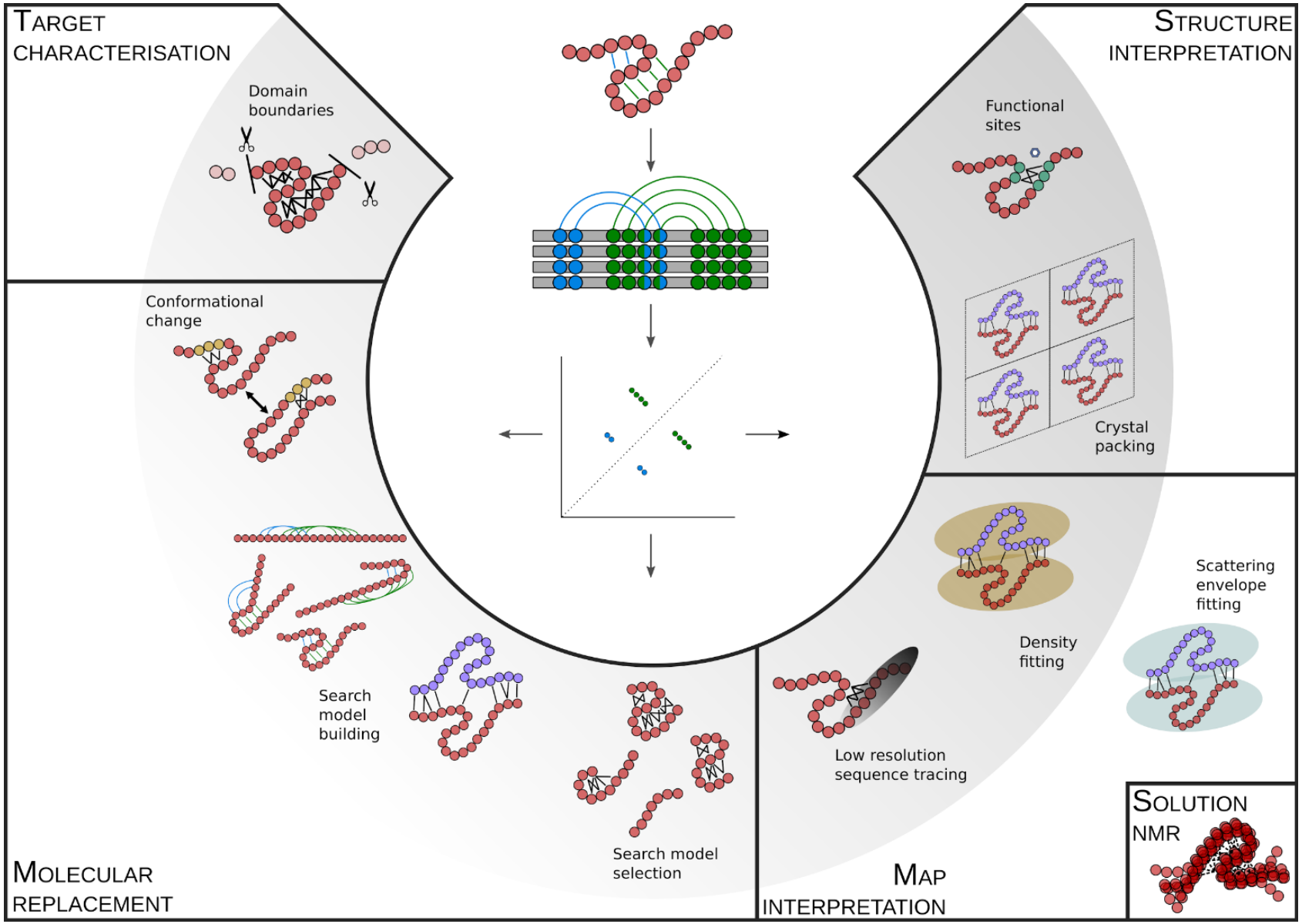
```
>seq1
FASDGITF
DRSLFFGH

>seq2
NMHKLSDF
PLSERWAQ
```



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

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



Simkovic *et al.* (2017) Applications of contact predictions to structural biology IUCrJ **4**, 291-300

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?

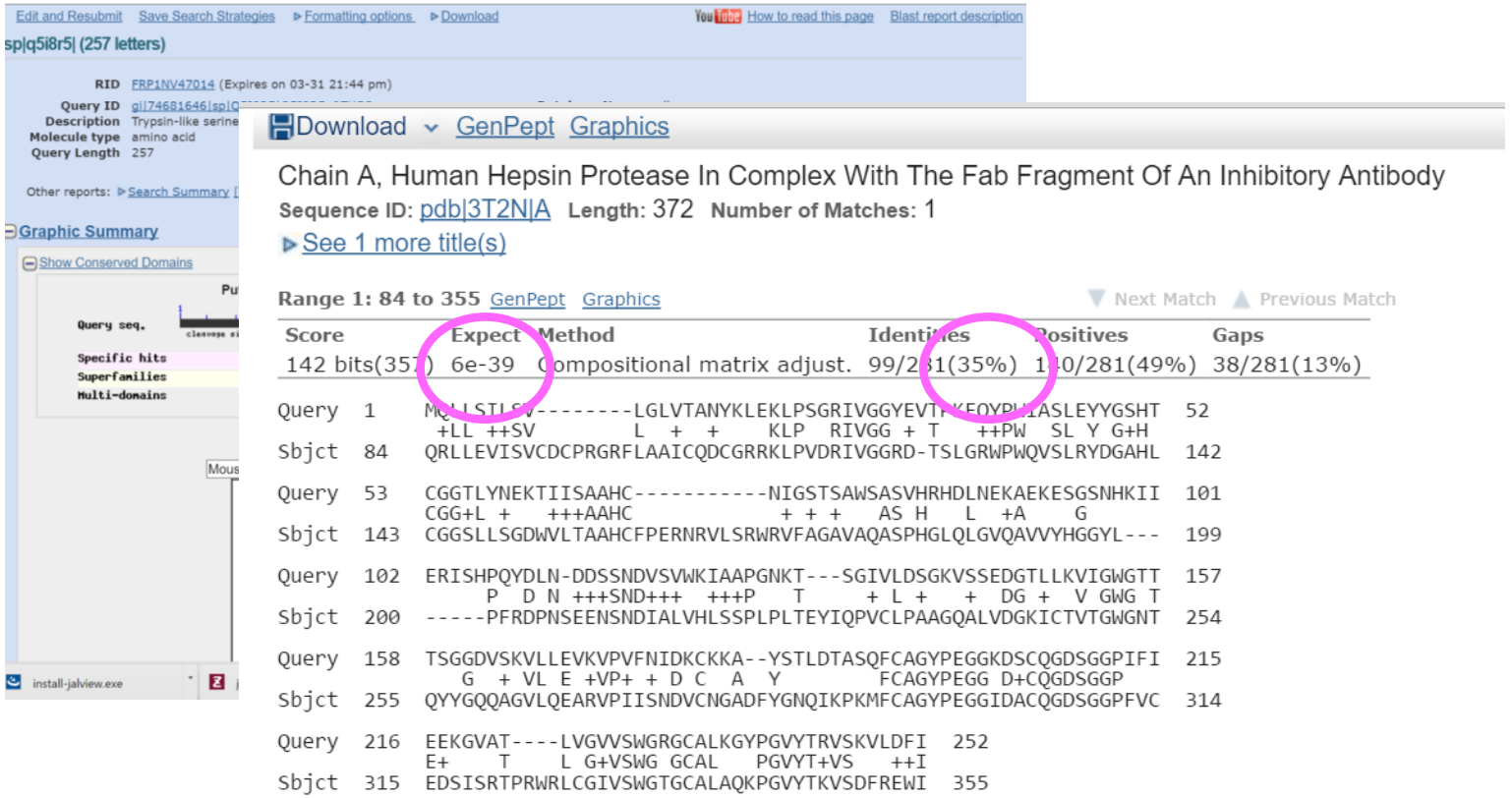
Which parts of your crystallised protein might enable phasing by MR, or experimentally?

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

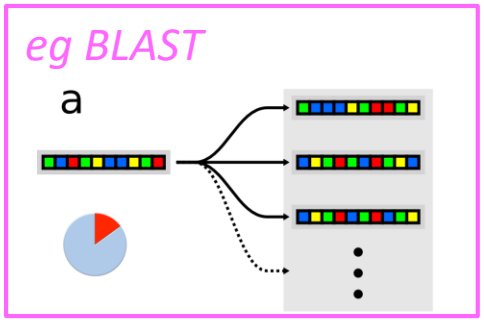
Recognising domains by homology with PDB, SCOP, CATH

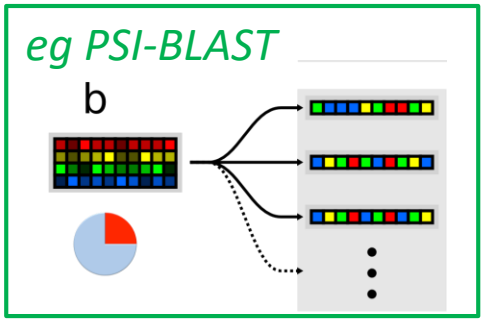
- (PSI-)BLAST against the PDB might do it

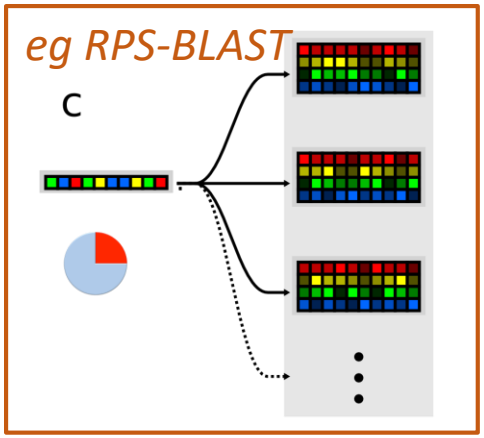


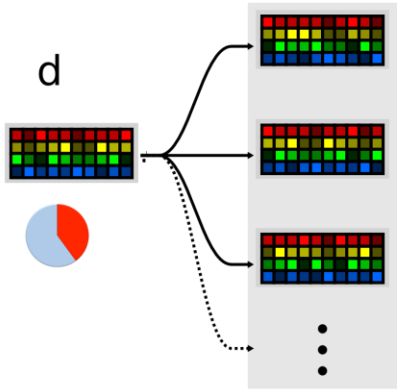
Recognising domains by homology with PDB, SCOP, CATH

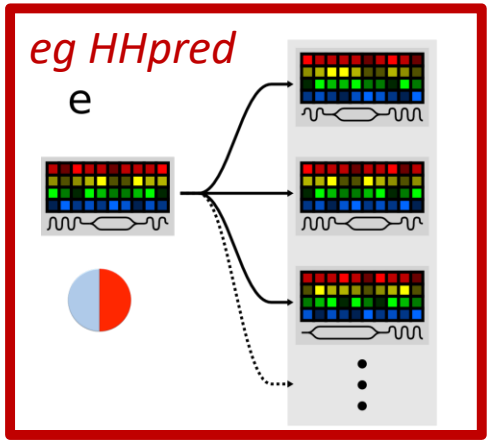
- Harder cases require a more sensitive tool. I recommend HHpred. Used by MrBUMP to find homologues to use as search models







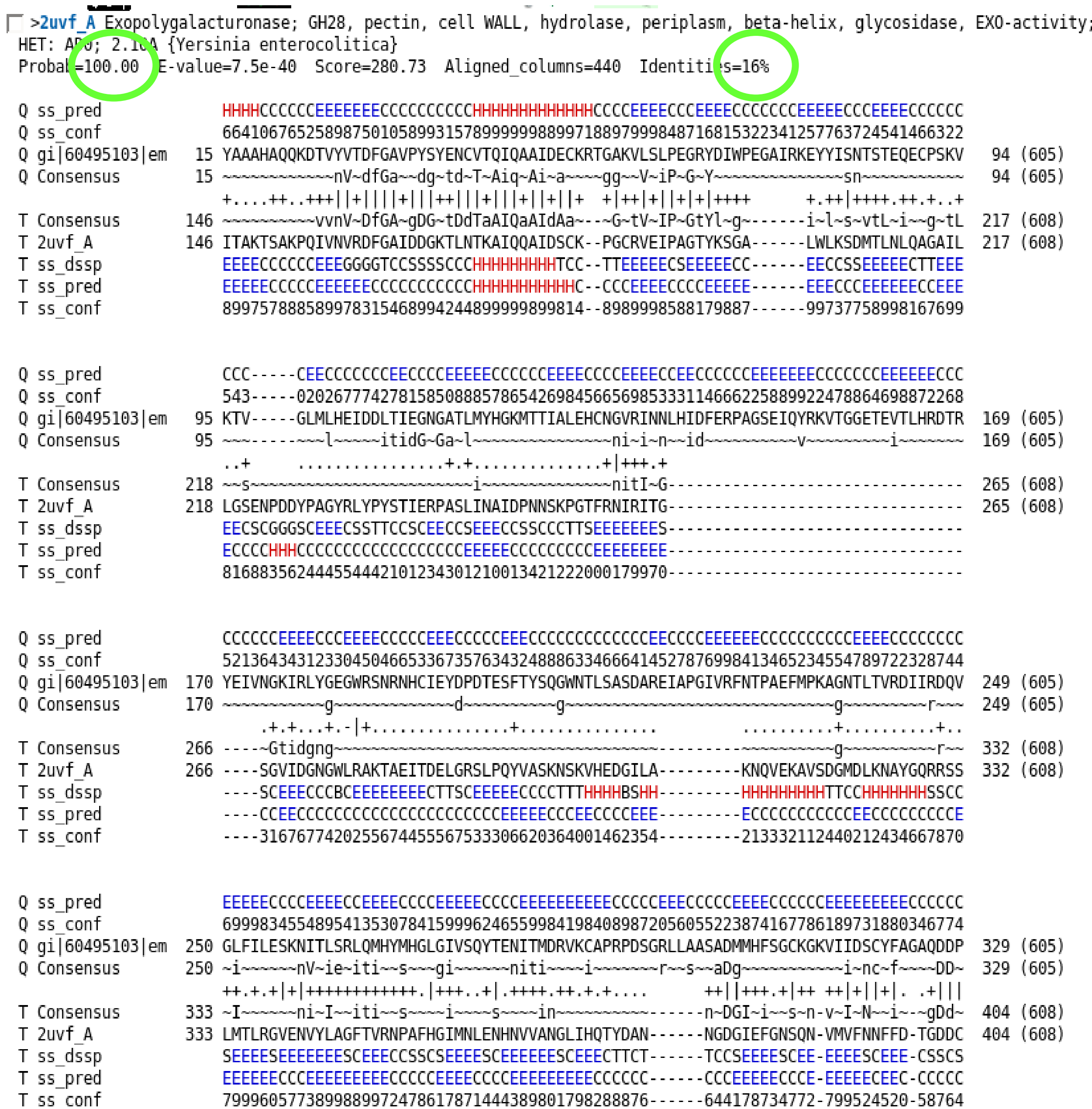




Lawrence Kelley

Recognising domains by homology with PDB, SCOP, CATH

- HHPRED works by comparing HMMs of alignments, not just single sequences
- Matching of secondary structure also scored



HHpred tips and warnings

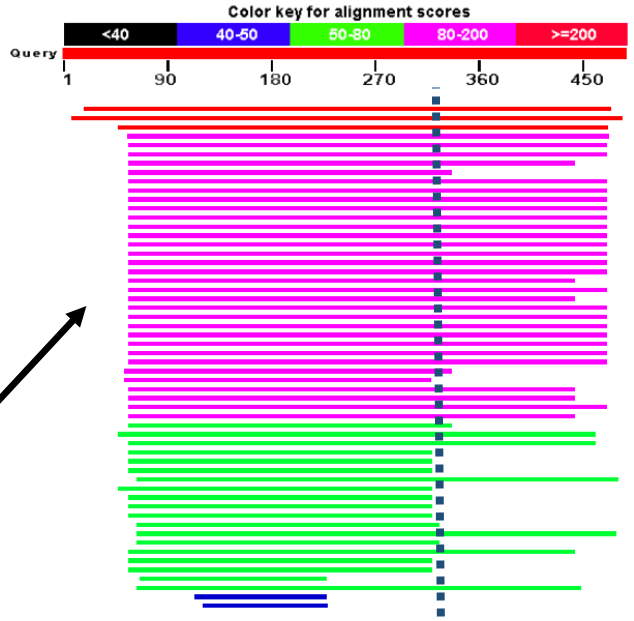
- Probability (0-100) is generally a good guide….
- …but statistics can mislead for unusual protein sequences eg coiled-coil, low-complexity, Cys-rich
- Consider if the match makes biological sense!
- Look at the matched region
  - Is it a complete domain/structure?
  - If partial, could it reasonably 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

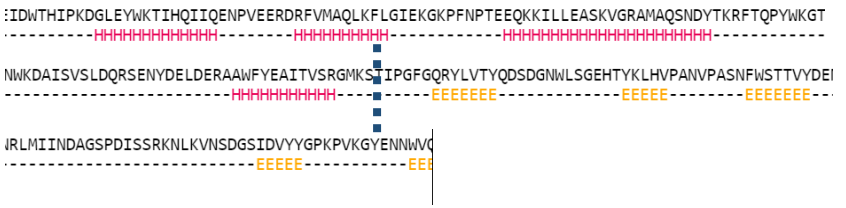
Recognising matches to domains in **sequence** databases, Pfam, Smart etc

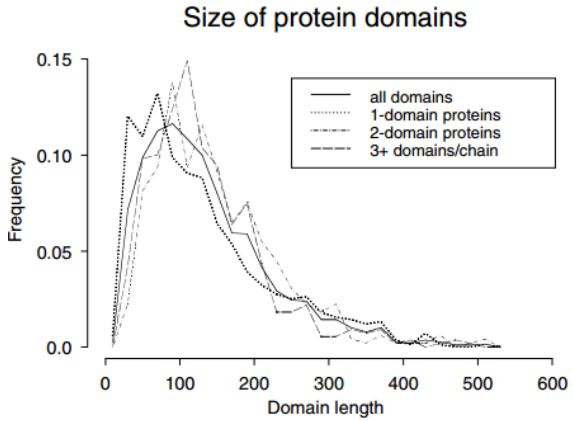
- HHpred is also an excellent, sensitive way to search against these. The same rules for interpretation apply. Matches, even distant ones, can obviously shed light on function
- However, for **structural** purposes need to remember that
  - Many Pfam families don't have structures so that domain limits 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)

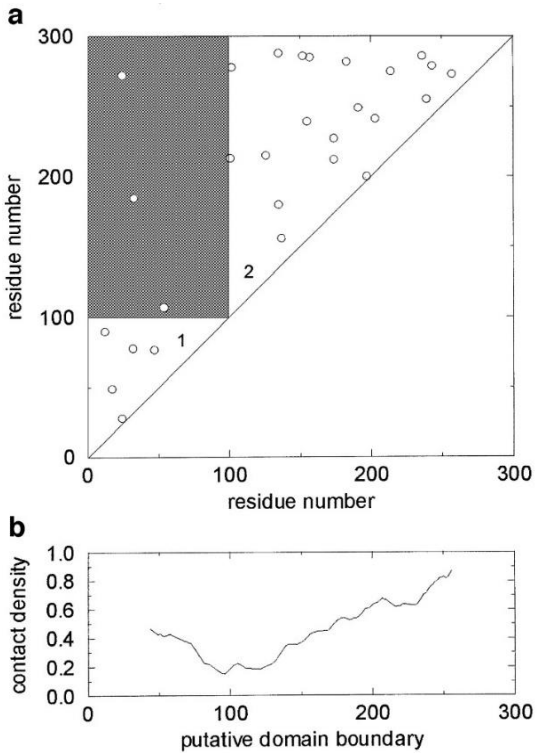


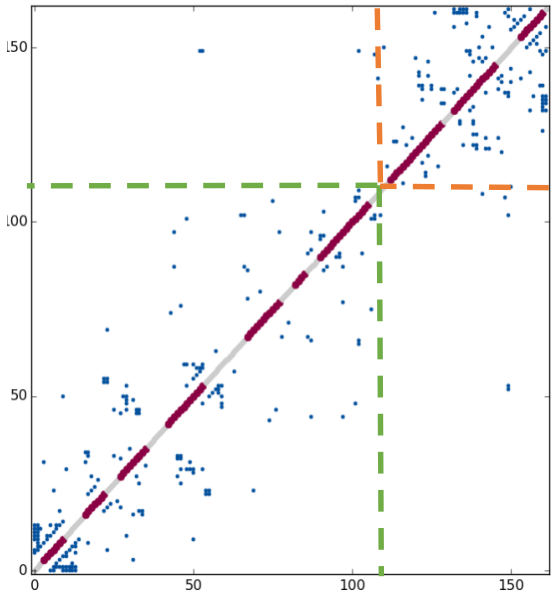




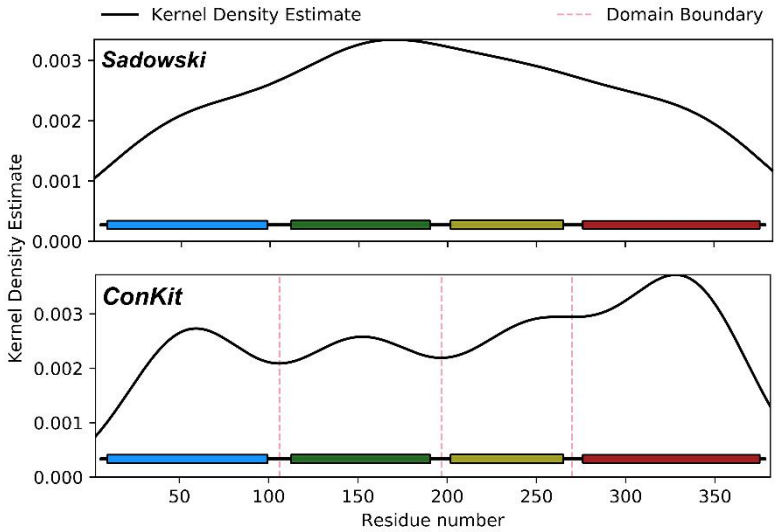
Predicted contacts for defining domains

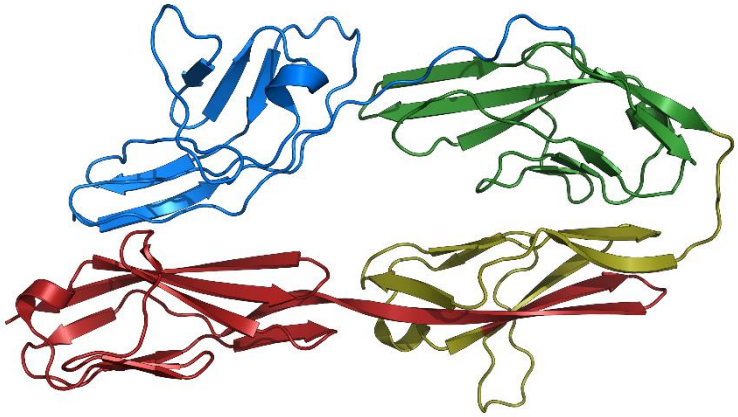
- Domain boundaries required for individual expression *in vitro* and helpful for fold recognition and modelling *in silico*





DUF4158 – two helical domains?

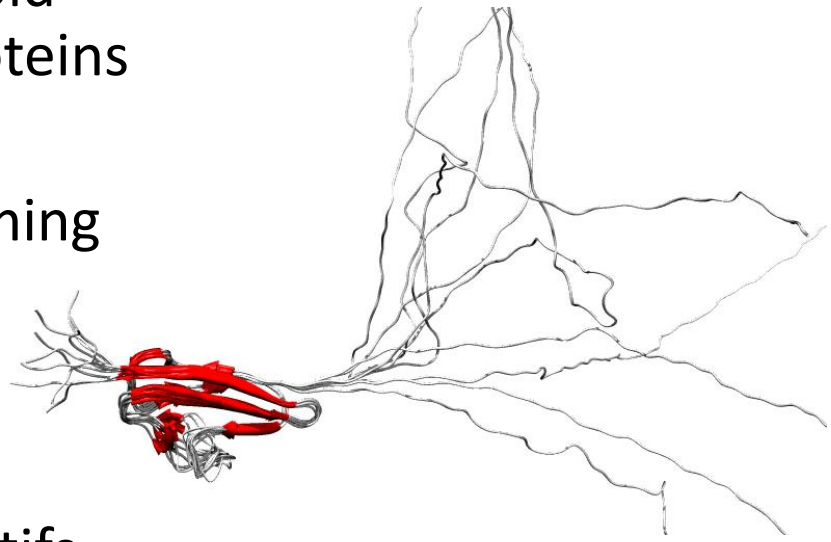


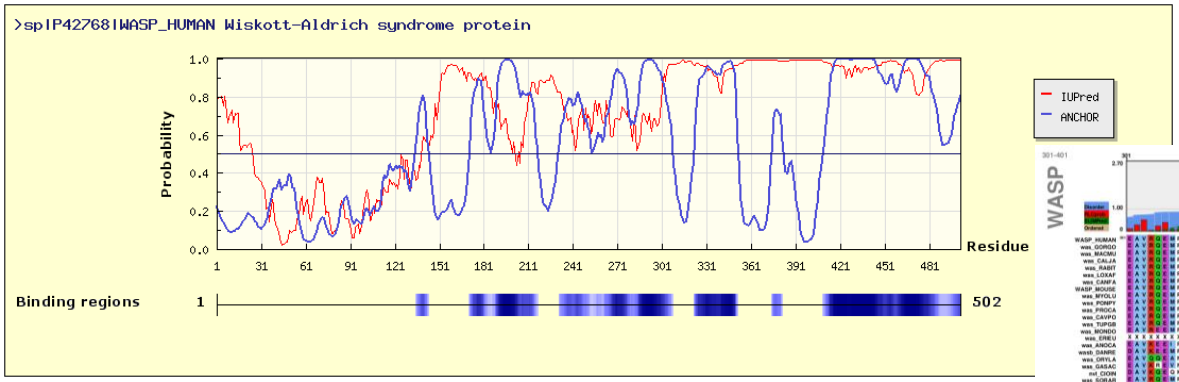


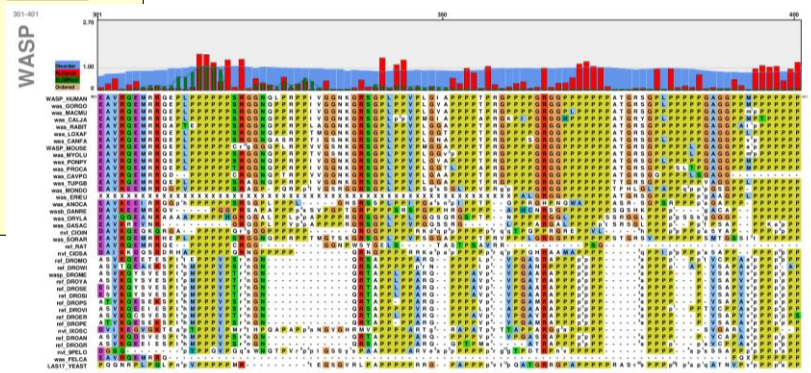
Rigden (2002)
Prot. Eng. **15**, 65

Intrinsic disorder prediction

- 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)

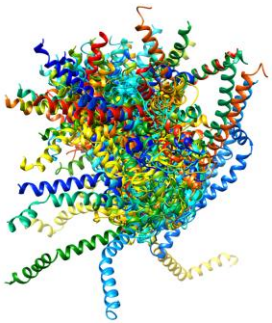


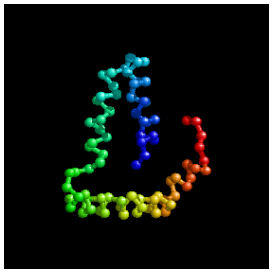




Building ensemble search models with AMPLE

***ab initio* models (ROSETTA/ QUARK)**



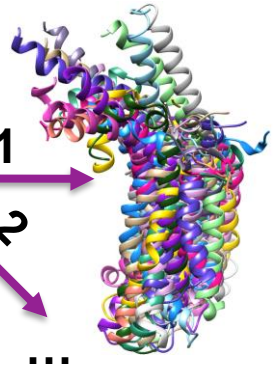


cluster (SPICKER)

1

2

…



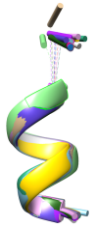
truncate (THESEUS)

10%

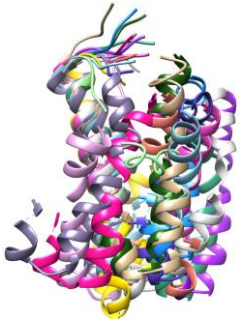
30%

50%

…



…



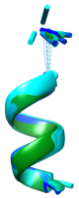
cluster (MAXCLUSTER)

1Å

2Å

3Å

…



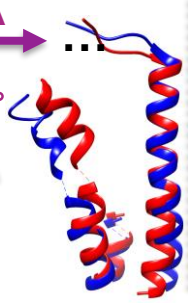
…

1Å

2Å

3Å

…

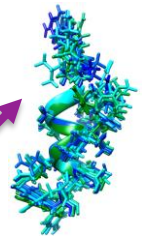


side chains

all-atom

reliable

polyAla



…

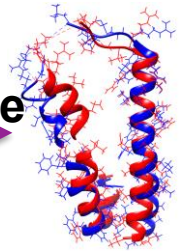
…

all-atom

reliable

polyAla

…



MrBUMP

Phaser

ShelxE

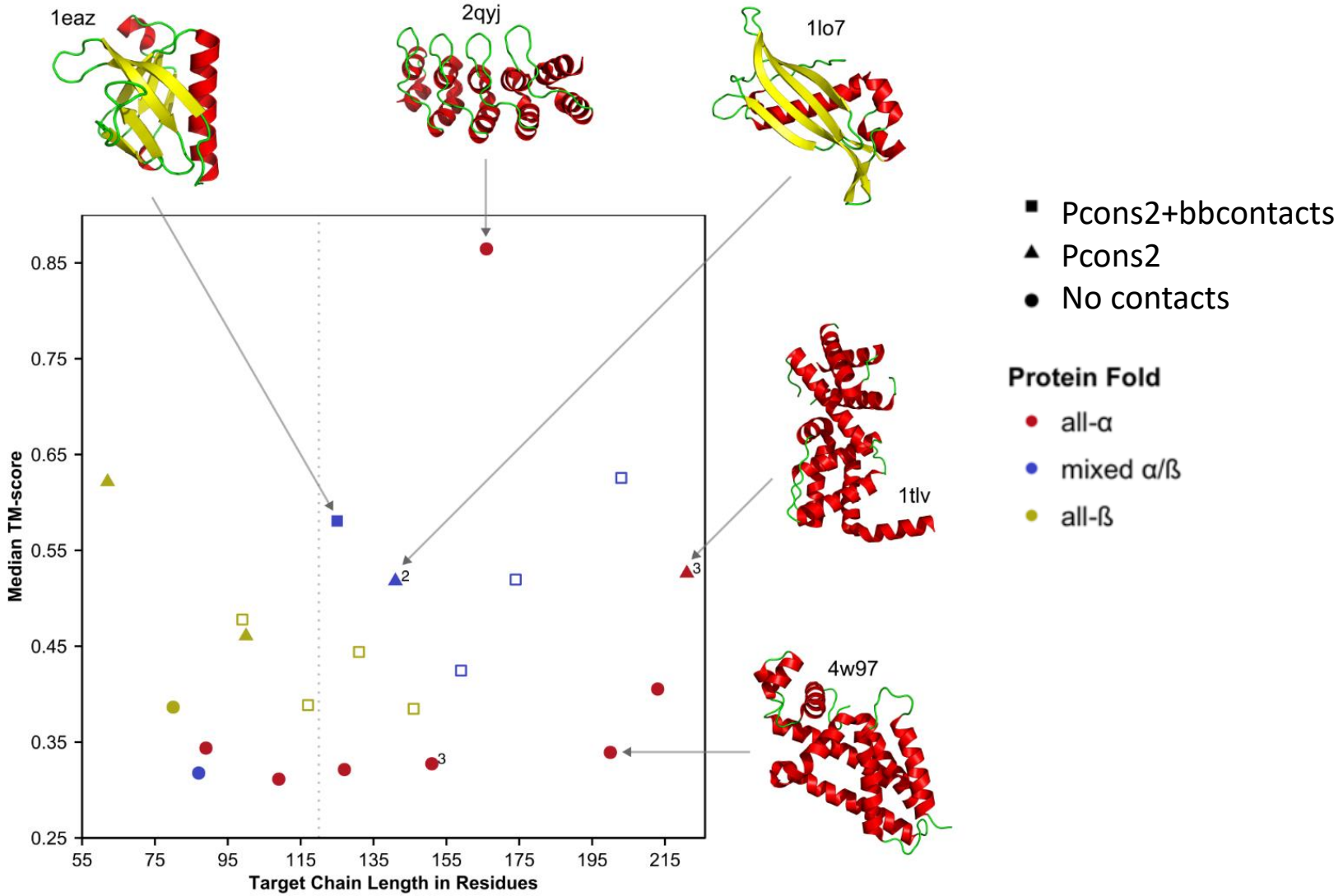
Buccaneer and/or ARP/wARP

http://www.cs.ucl.ac.uk/staff/D.Jones/t42morph.html

Bibby *et al.* (2012) Acta Cryst D**68** 1622

up to 180 ensemble search models

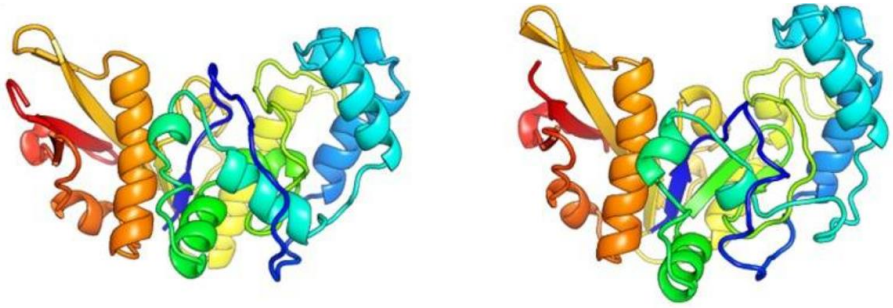
With contact predictions, AMPLE and ROSETTA produces models that succeed with larger and more β-rich proteins



Simkovic *et al.* (2016) IUCr J. **3**, 259-270

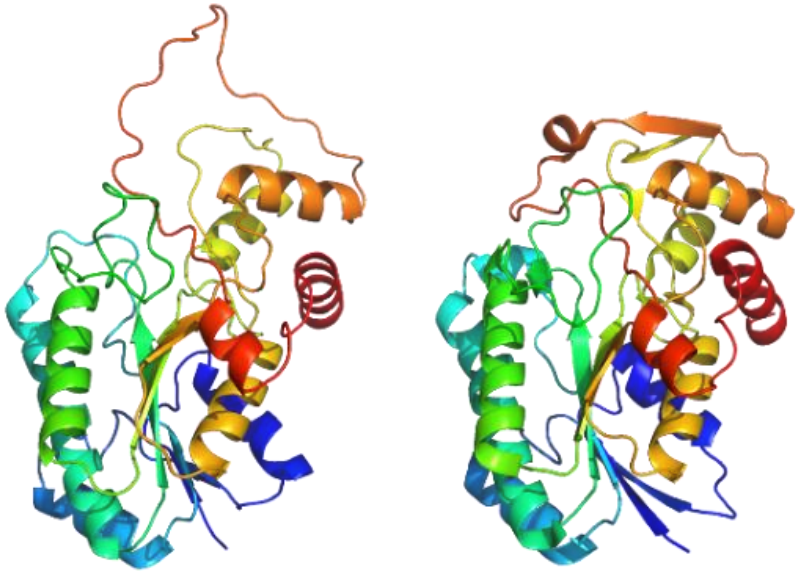
Using *ab initio*, contact-driven models from databases

- 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



Ovchinnikov *et al.* (2017) Science **355** 294

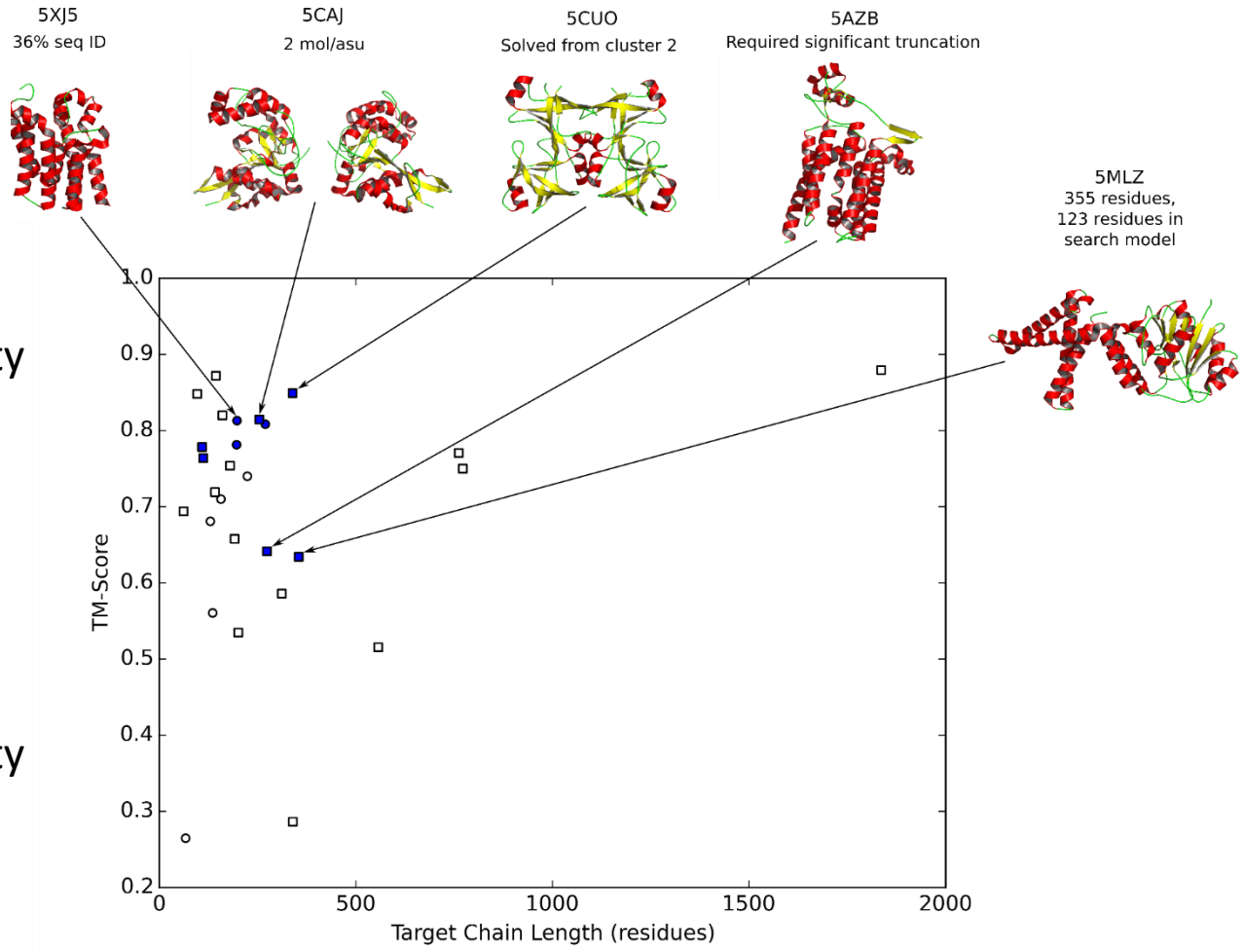
- PConsFam database
- Models for 13,617 Pfam families
- Single best model per family
- Scored according to predicted reliability
- pconsfam.bioinfo.se



Michel *et al.* (2017) Bioinformatics **33** i23

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)

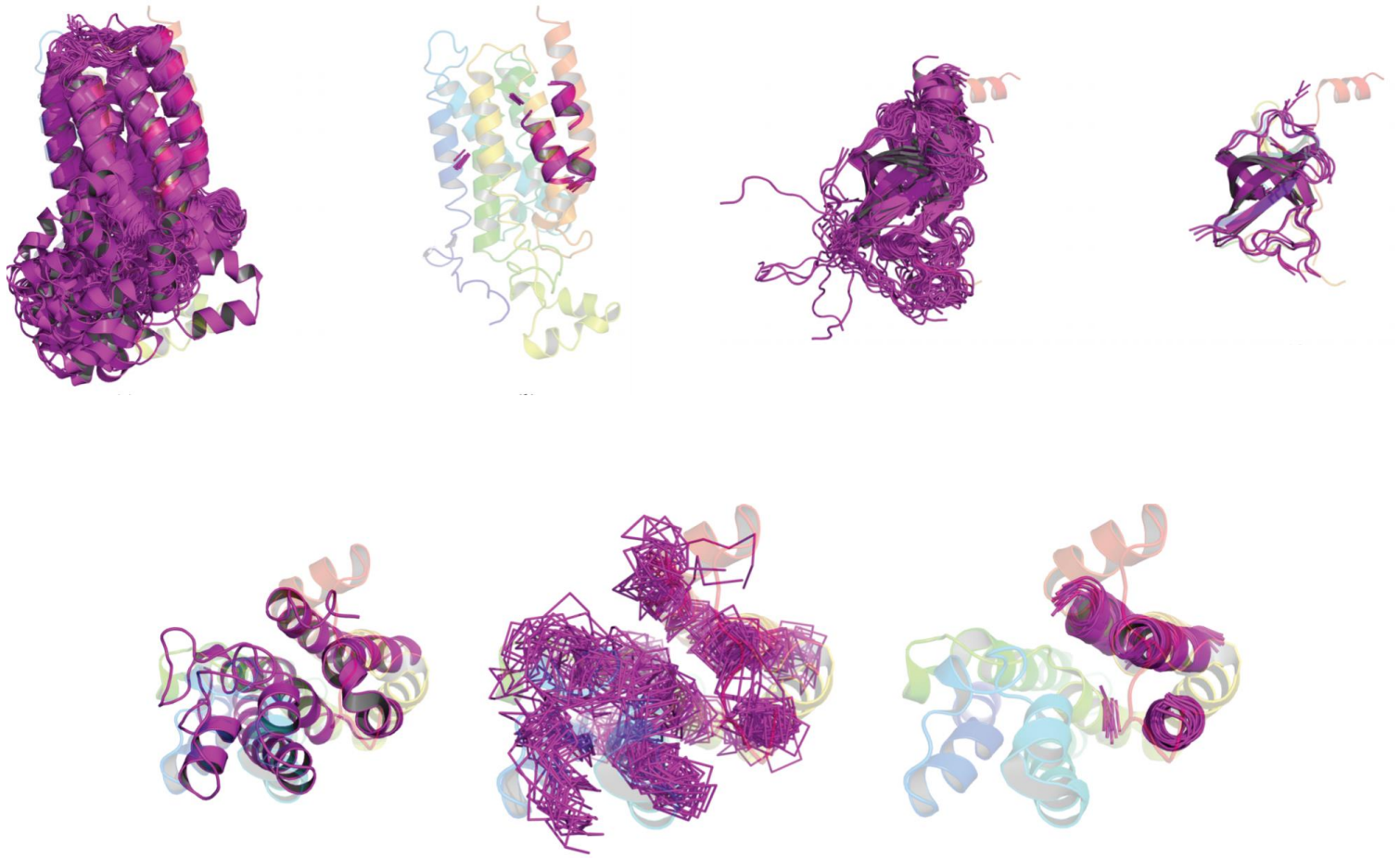


Simpkin *et al.* (2019) Acta Cryst. **D75**, in press

AMPLE working on *ab initio* model database deposits

GREMLIN

PCONSFAM



Simpkin *et al.* (2019) Acta Cryst. **D75**, in press

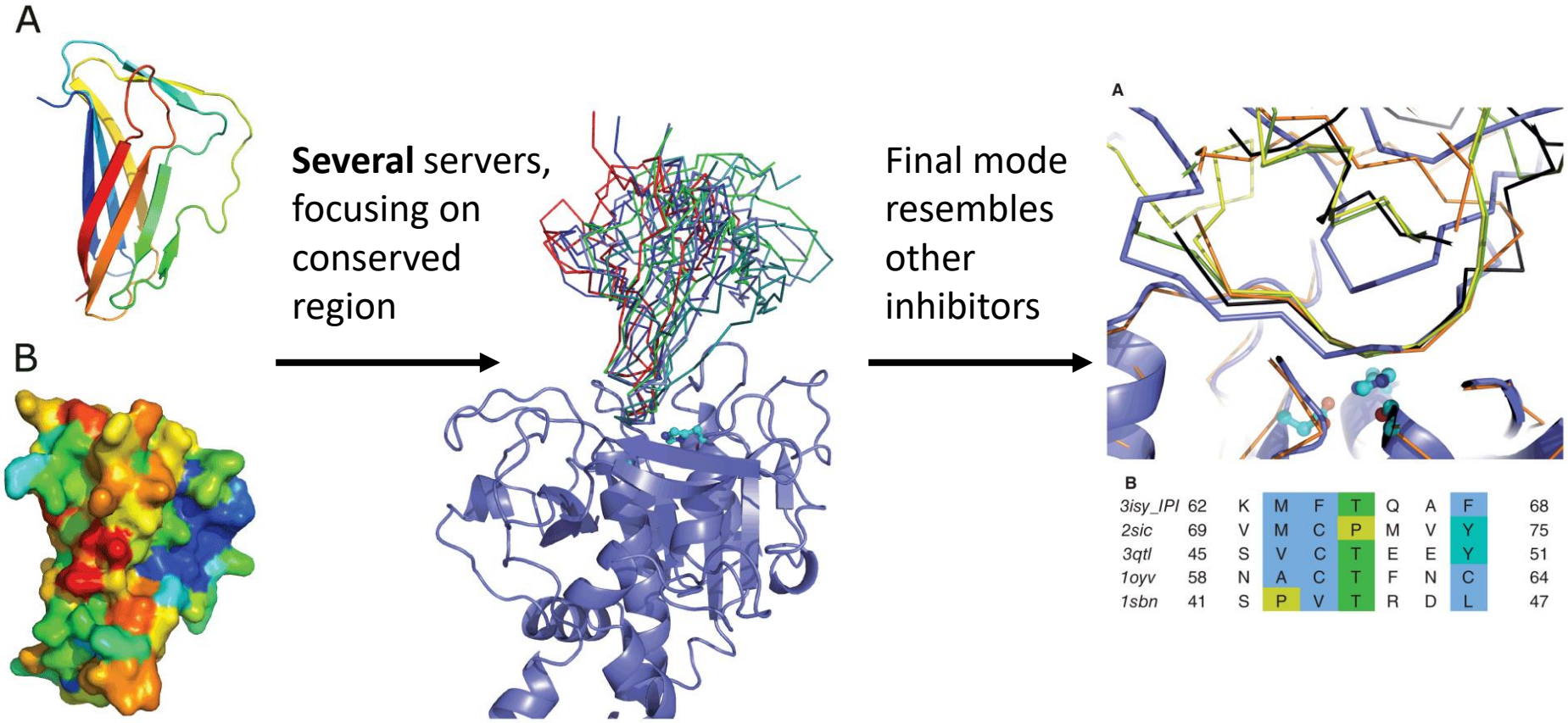
Quaternary structure

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

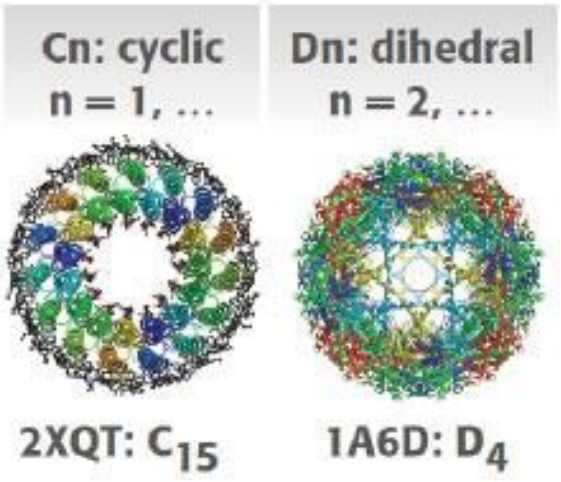
Multiple methods in bioinformatics: *B. subtilis* IPI docking to protease



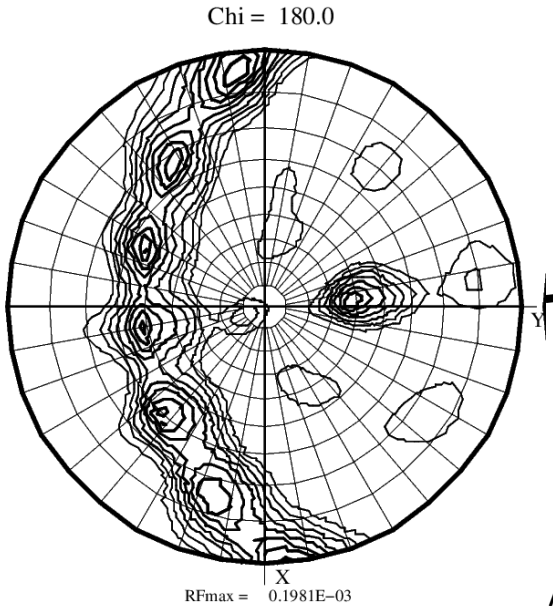
Rigden et al. (2013) F1000Research 2:154

ROSIE symmetric docking for oligomers

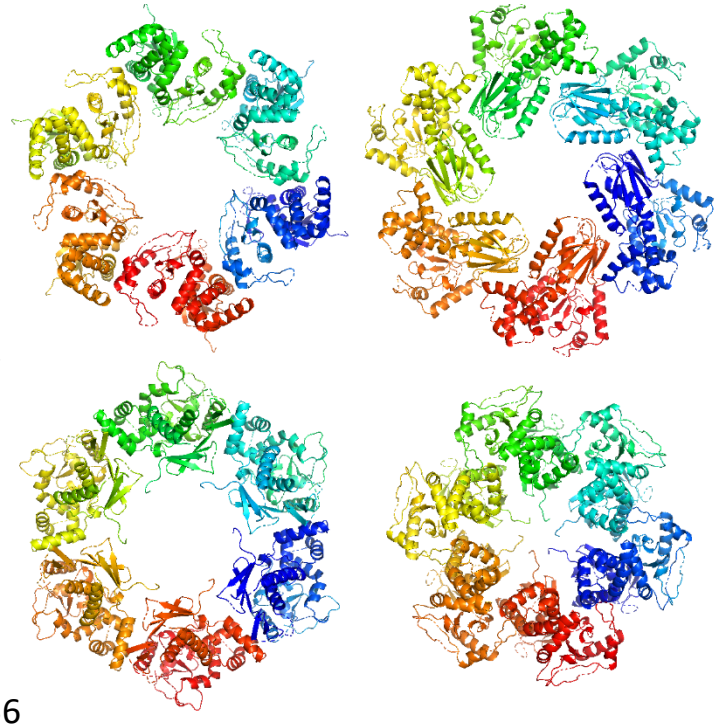
- Only cyclic (C_n) or dihedral (D_n) symmetry at server
- Clues from self rotation function may be available



rcsb.org

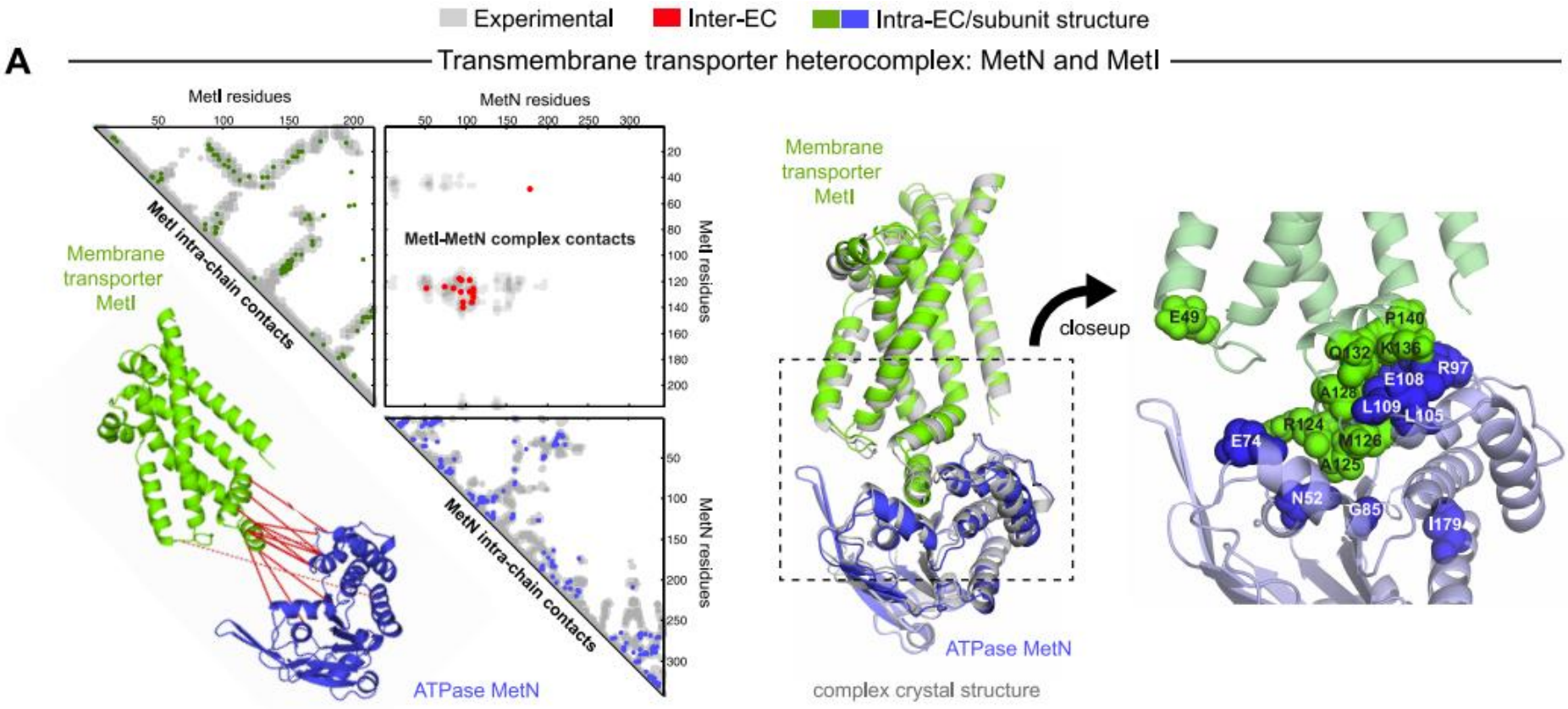


Generate hexamers



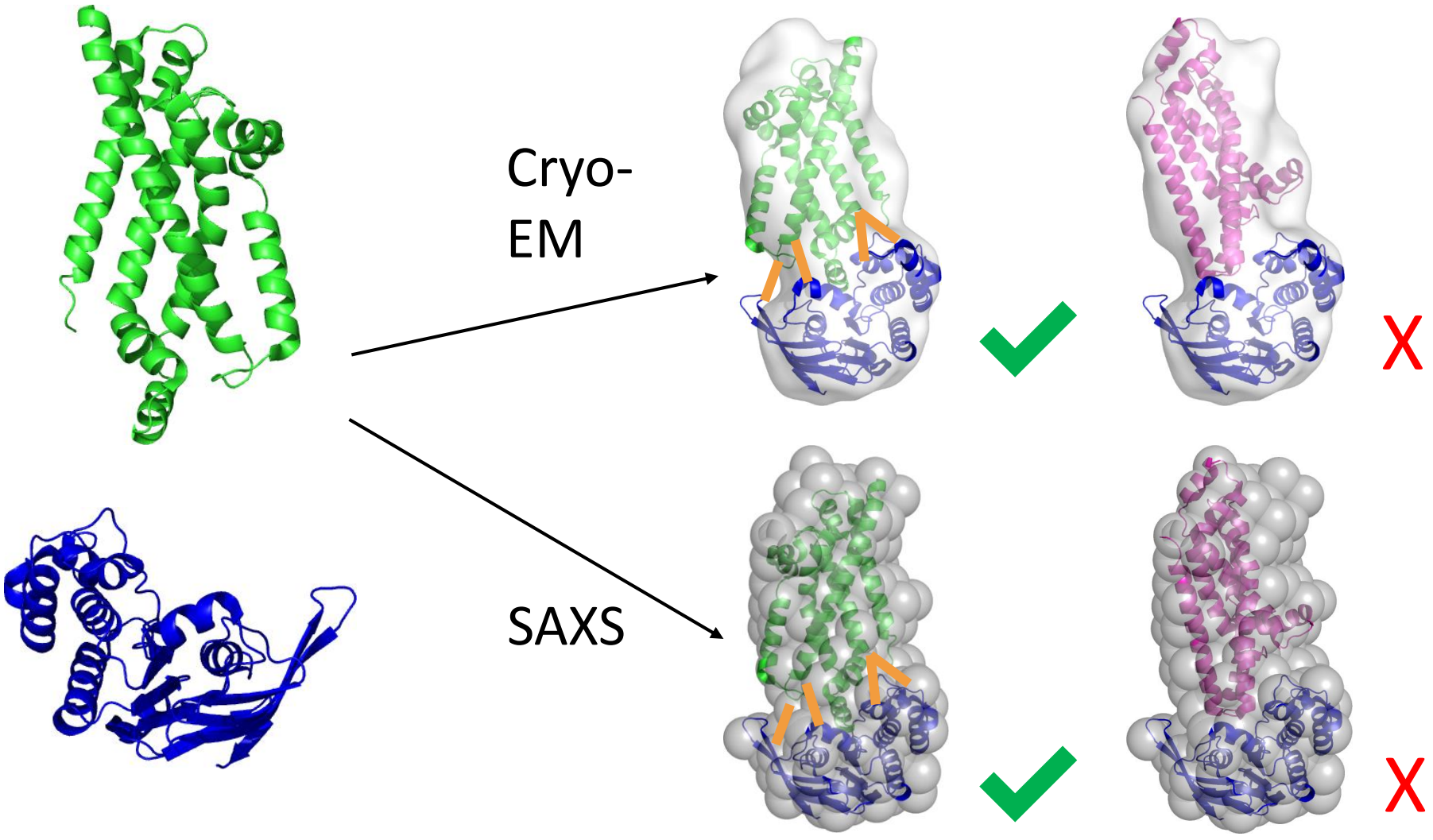
Andre et al. (2009) PNAS 104:17656

Using predicted contacts to help predict complexes



Hopf et al. eLife (2014) 3,e03430.

Fitting to low-resolution envelopes



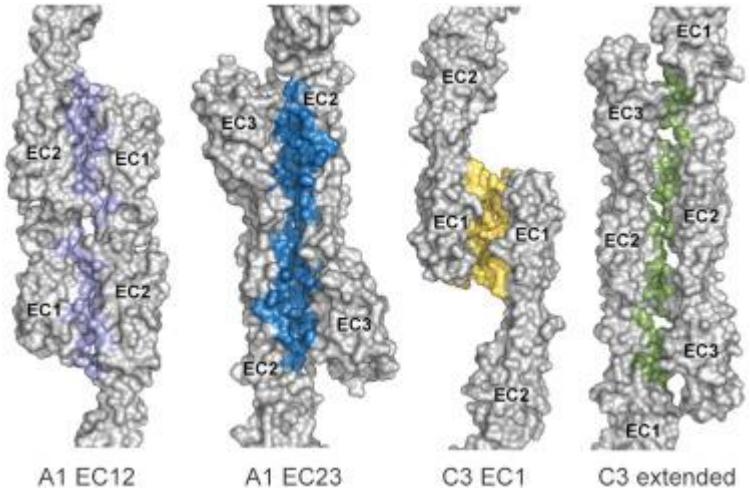
Structure-based function interpretation

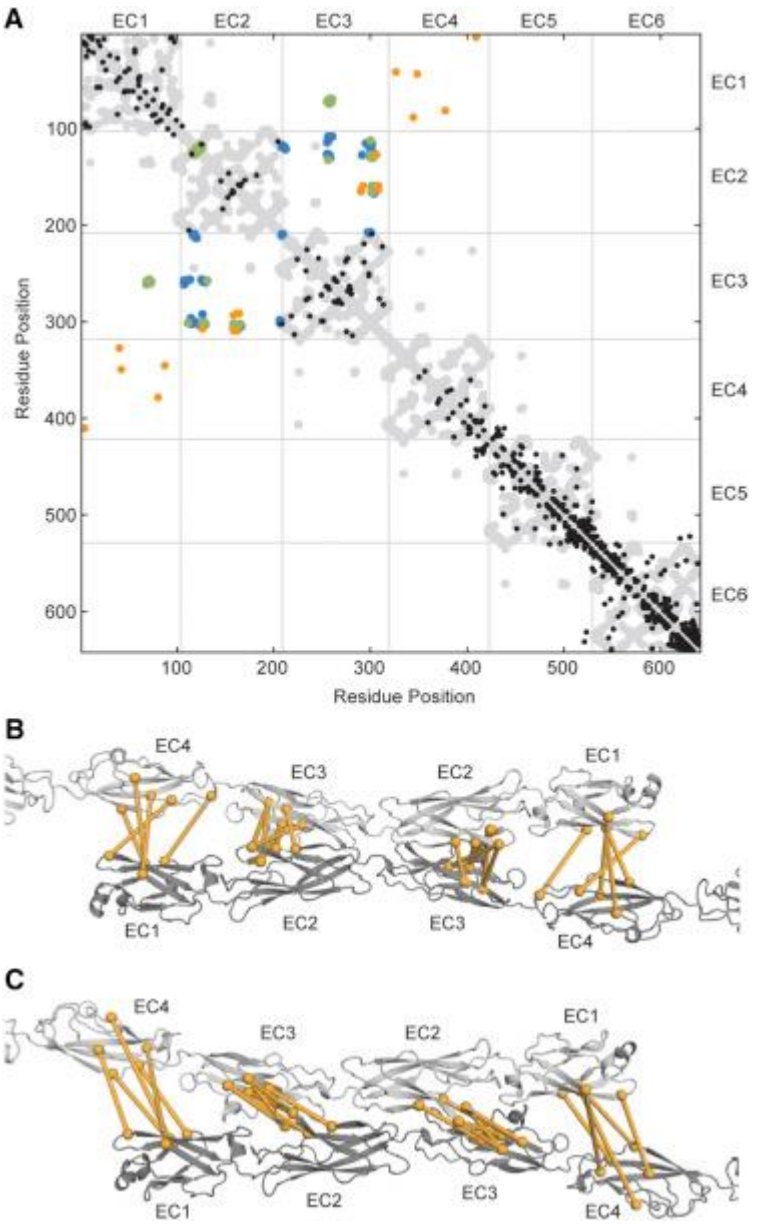
What is the biologically relevant quaternary structure?

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





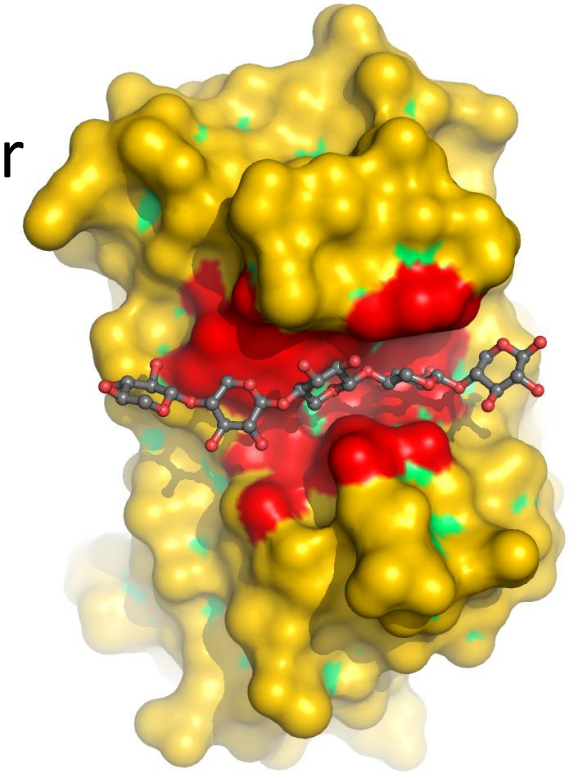
Nicoludis *et al*. (2015) Structure 23, 2087

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)

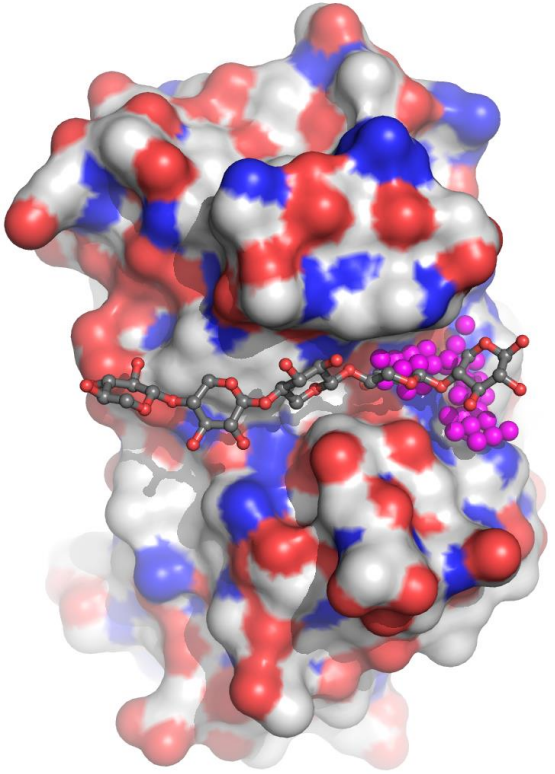
Different probes for different binding sites

- 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

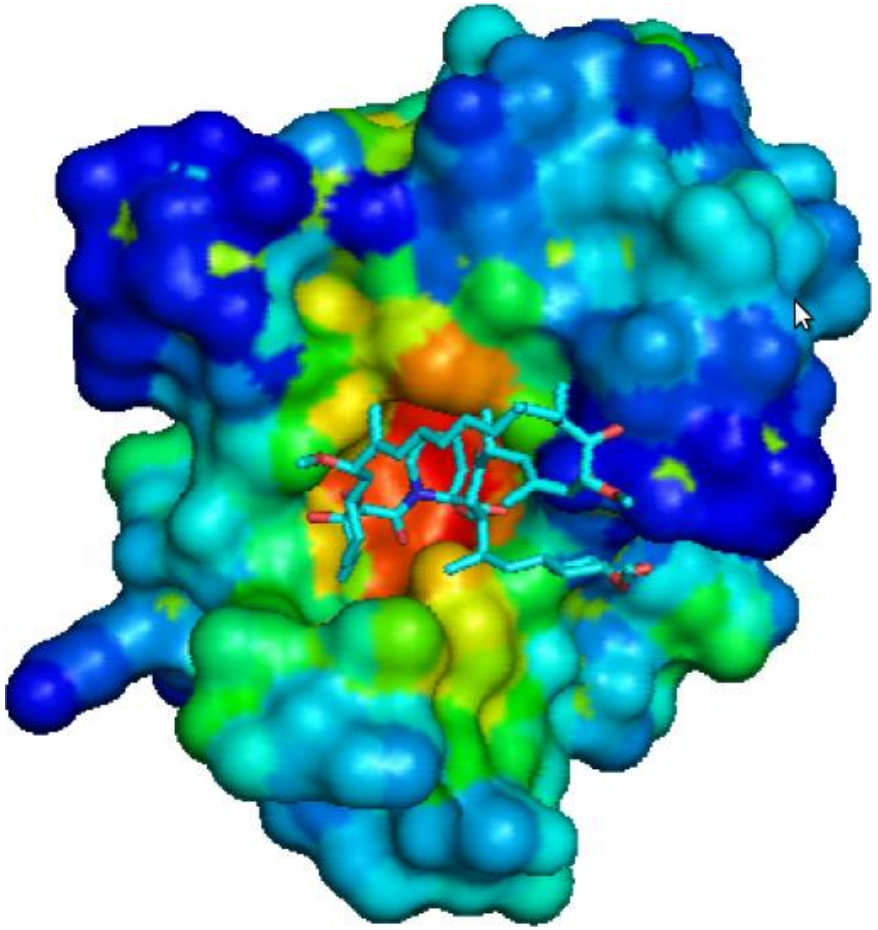


SiteHound

scbx.mssm.edu/sitehound

Binding sites from statistics

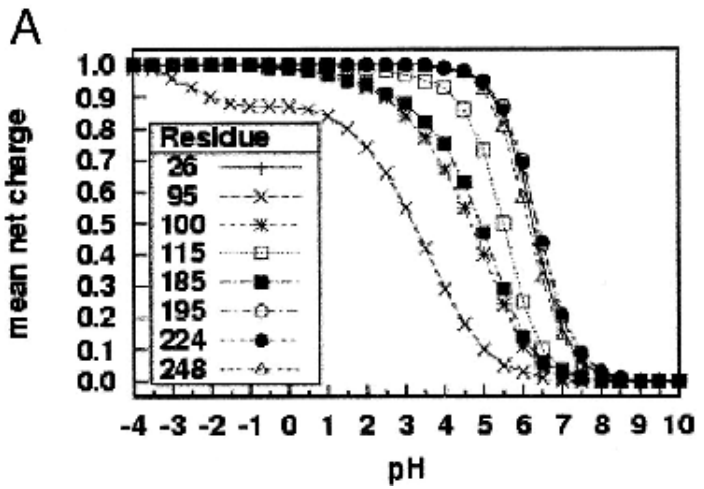
- STP (surface triplet propensities)
- 13 atom types → 455 triplets
- Distribution in binding vs non-binding sites varies
- Designed for small molecules, works on PPIs, including flat surfaces



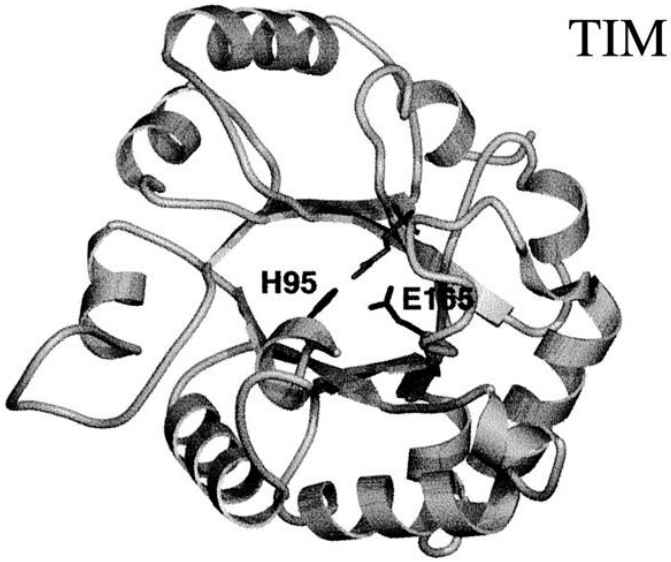
opus.bch.ed.ac.uk/stp

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

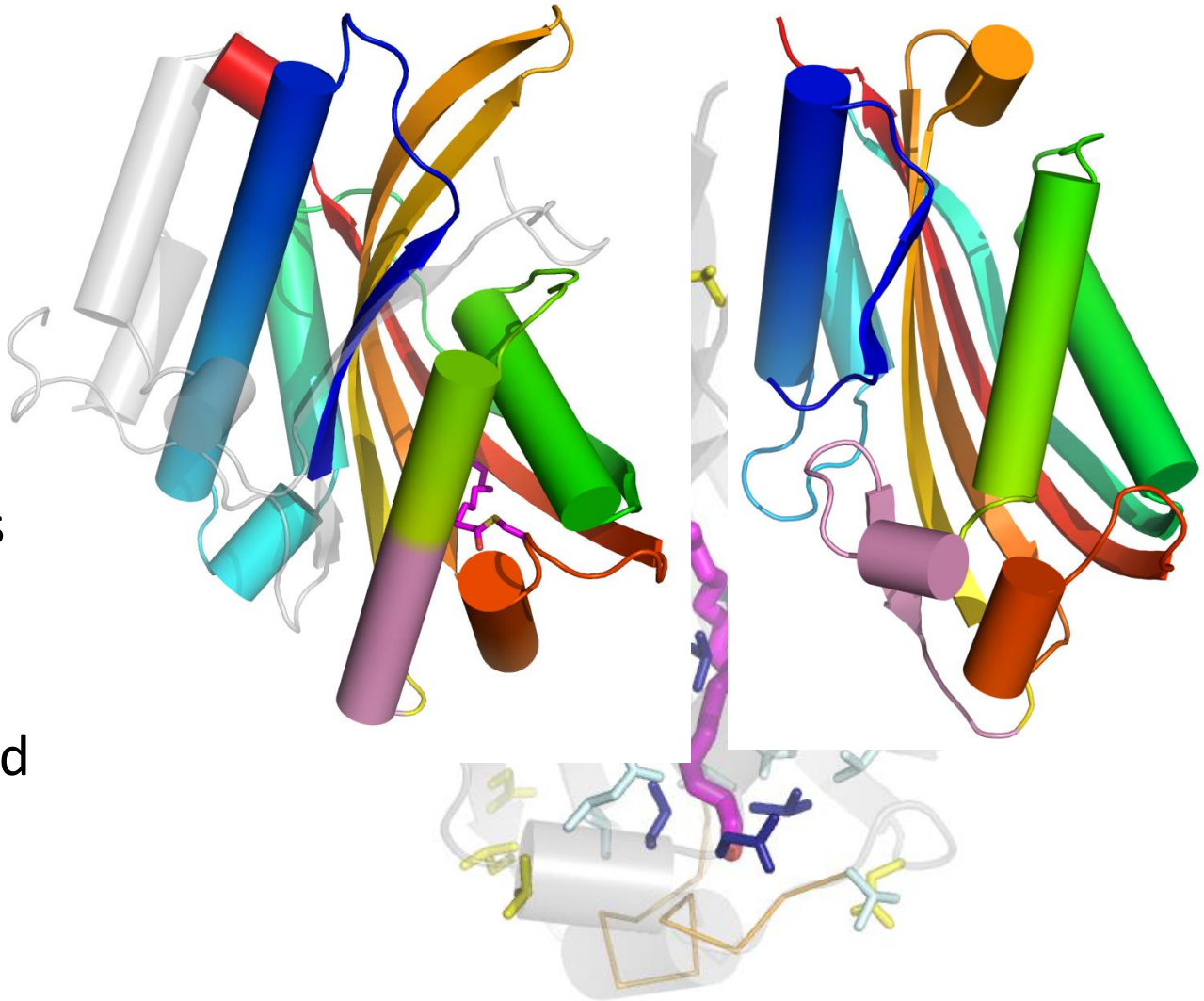


His95 and other residues with atypical pKa cluster at catalytic site

www.pool.neu.edu/wPOOL/index2.jsp

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



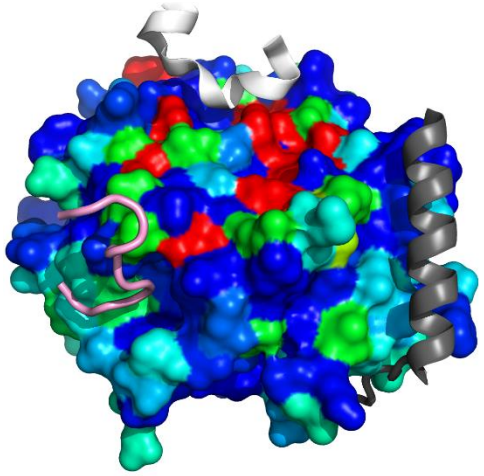
Quevillon-Cheruel et al. (2009) J. Biol. Chem. **35**,2107

Rigden (2009) FEBS Lett. **583**, 1555

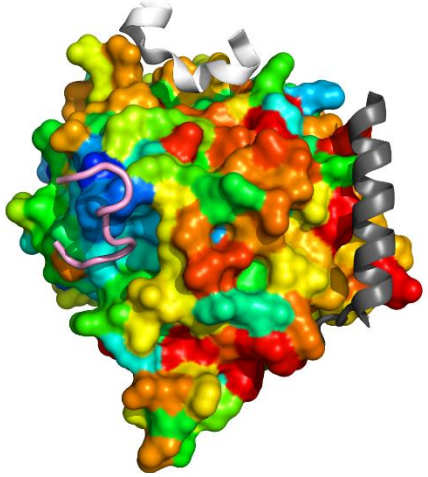
Some servers require thought…

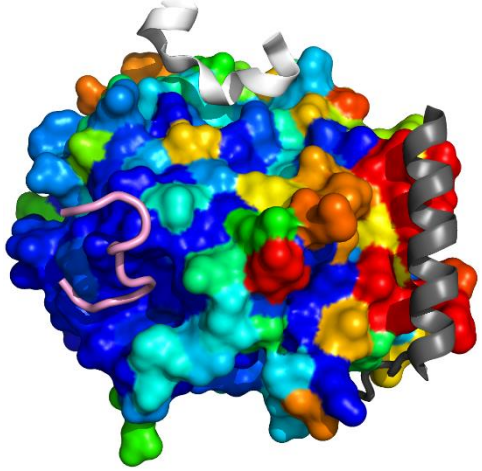
- Consurf maps sequence conservation onto a structure revealing functional sites
- Excellent, general method, but results depend on sequence set chosen for mapping: selecting all or only near relatives gives different results. Either might be more appropriate for you

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





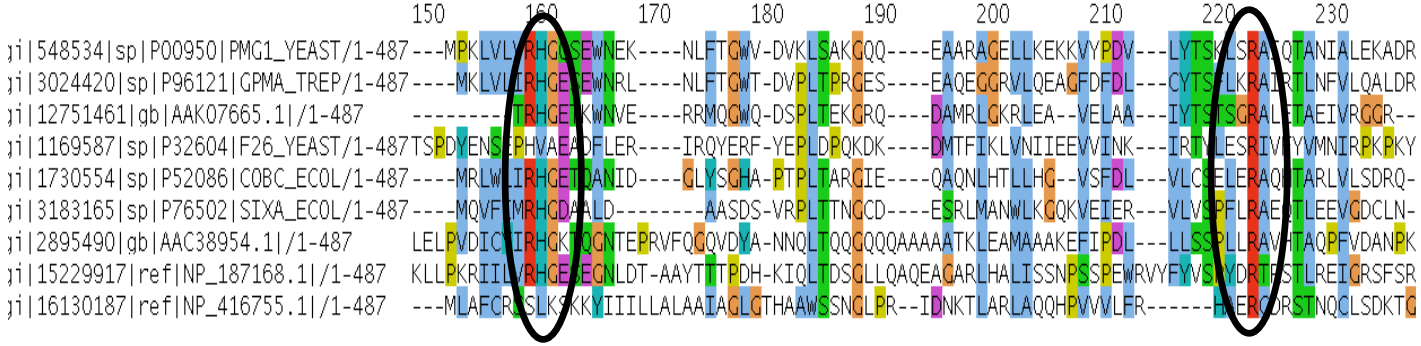
…and finally, you're putting a manuscript together

Calculating and presenting sequence alignments

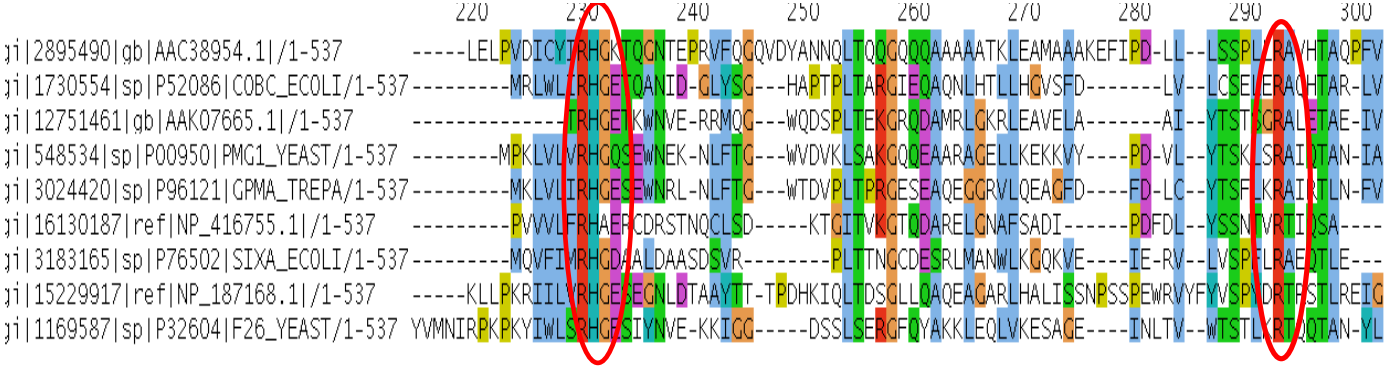
Your sequence alignment

- 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…

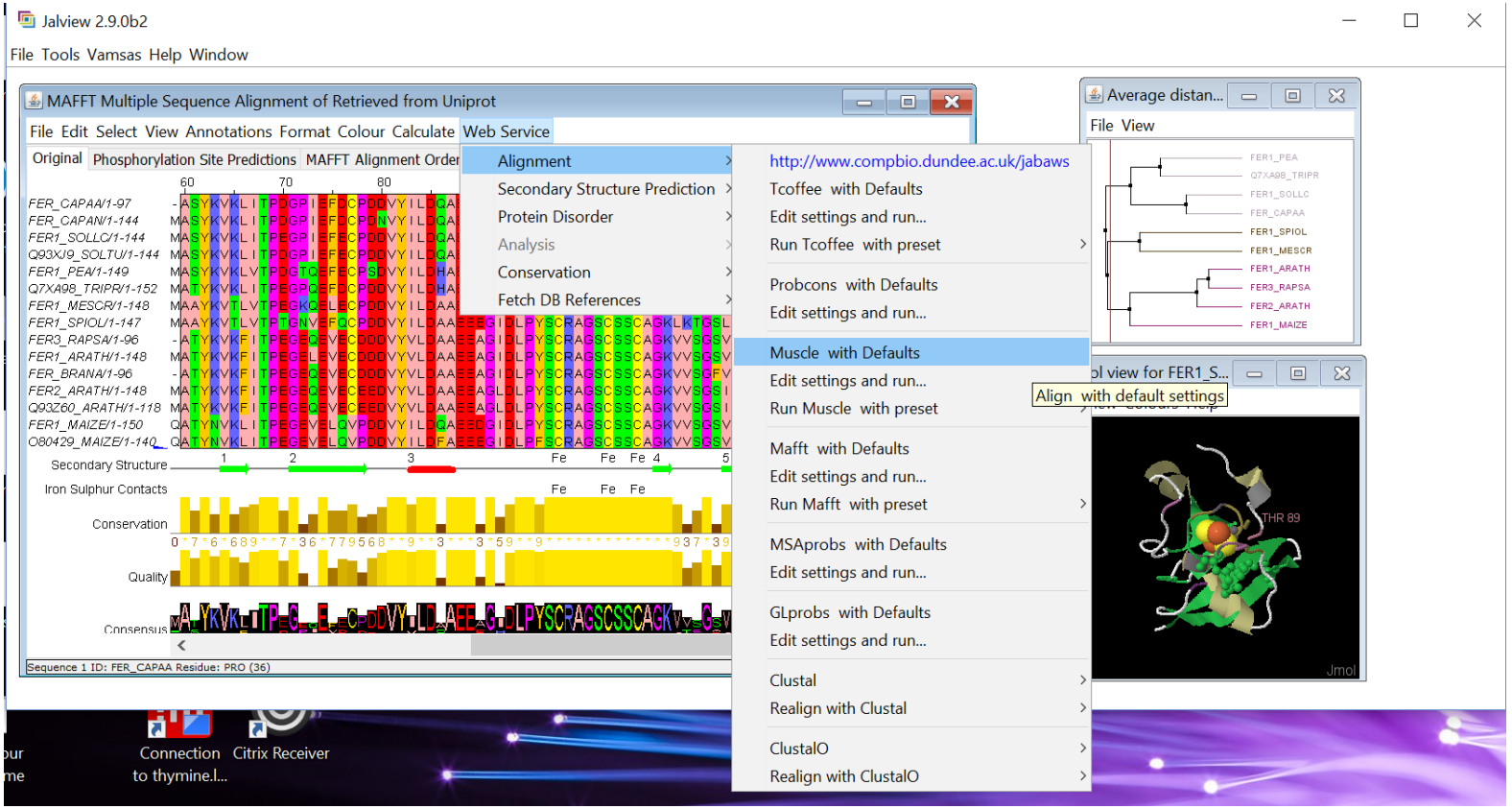


… but MUSCLE gets it



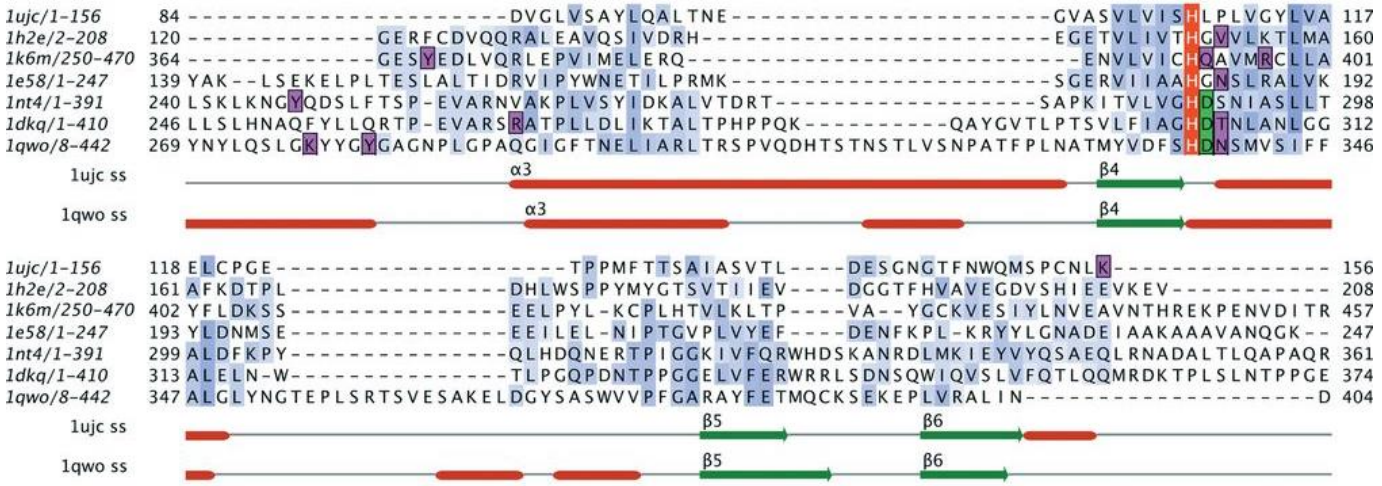
Jalview.org, recommended for sequence alignments

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



Jalview

- Also helps you produce figures like this…



- … rather than like this

Questions? Feedback?

UNIVERSITY OF
LIVERPOOL