Bioinformatics for crystallographers

Dan Rigden

UNIVERSITY OF LIVERPOOL

Preamble

- Can’t cover all bioinformatics!
- Prof Garratt will cover structure-based function inference
- I will focus on bioinformatics (= prediction)
  - Using newer/less well-known data
  - Majoring on easily available servers/predictions
  - Related to predicting domain composition and interactions. Mainly relevant to construct design and MR (all MR, not just AMPLE!)
- Plus some other bits and pieces

A new source of data: further uses of predicted contacts

- Structures unknown
  - Making better *ab initio* models for MR (AMPLE)
  - Predicting domain boundaries

- Structures known
  - Predicting how proteins interact
  - Validating the content of your crystal structure
  - (Predicting functional sites, highlighting conformational states)

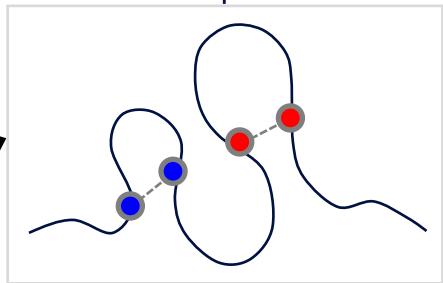
```
>target_seq (s)
GYYELYRRSTIGNSLVDALDTLISDG
RIEASLAMRVLETFDKVVAETLKDNT
QSKLTVKGNLDTYGFCDDVWTFIVKN
CQVTVEDQSVISVDKLRIVACNSKKS
```

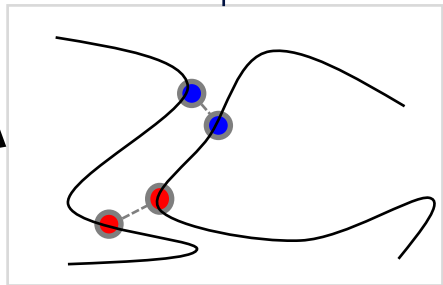
Multiple Sequence Alignment of homologous sequences

```
...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 predicted contacts

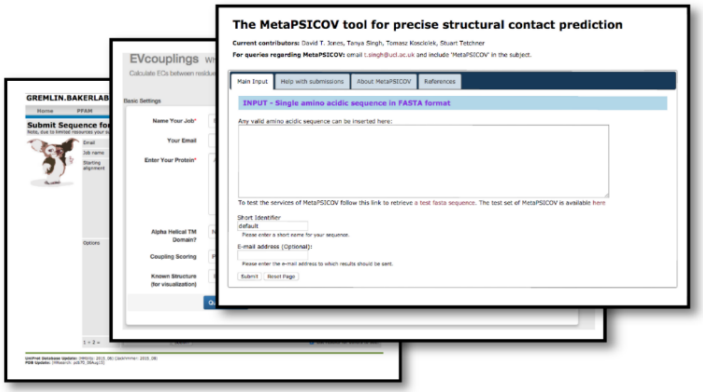




Applications…

Felix Simkovic

```
>seq
FASDGITF
DRSLFFGH
```

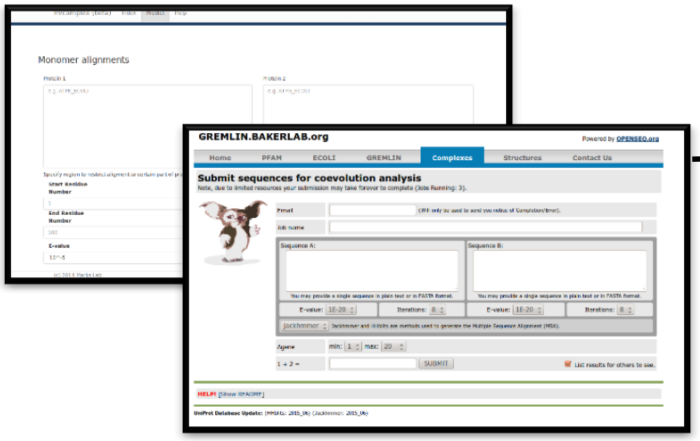


Intramolecular contacts

Intermolecular contacts if a **homo**oligomer

```
>seq1
FASDGITF
DRSLFFGH

>seq2
NMHKLSDF
PLSERWAQ
```



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

Thinking about your target …

Which part to express for crystallisation?

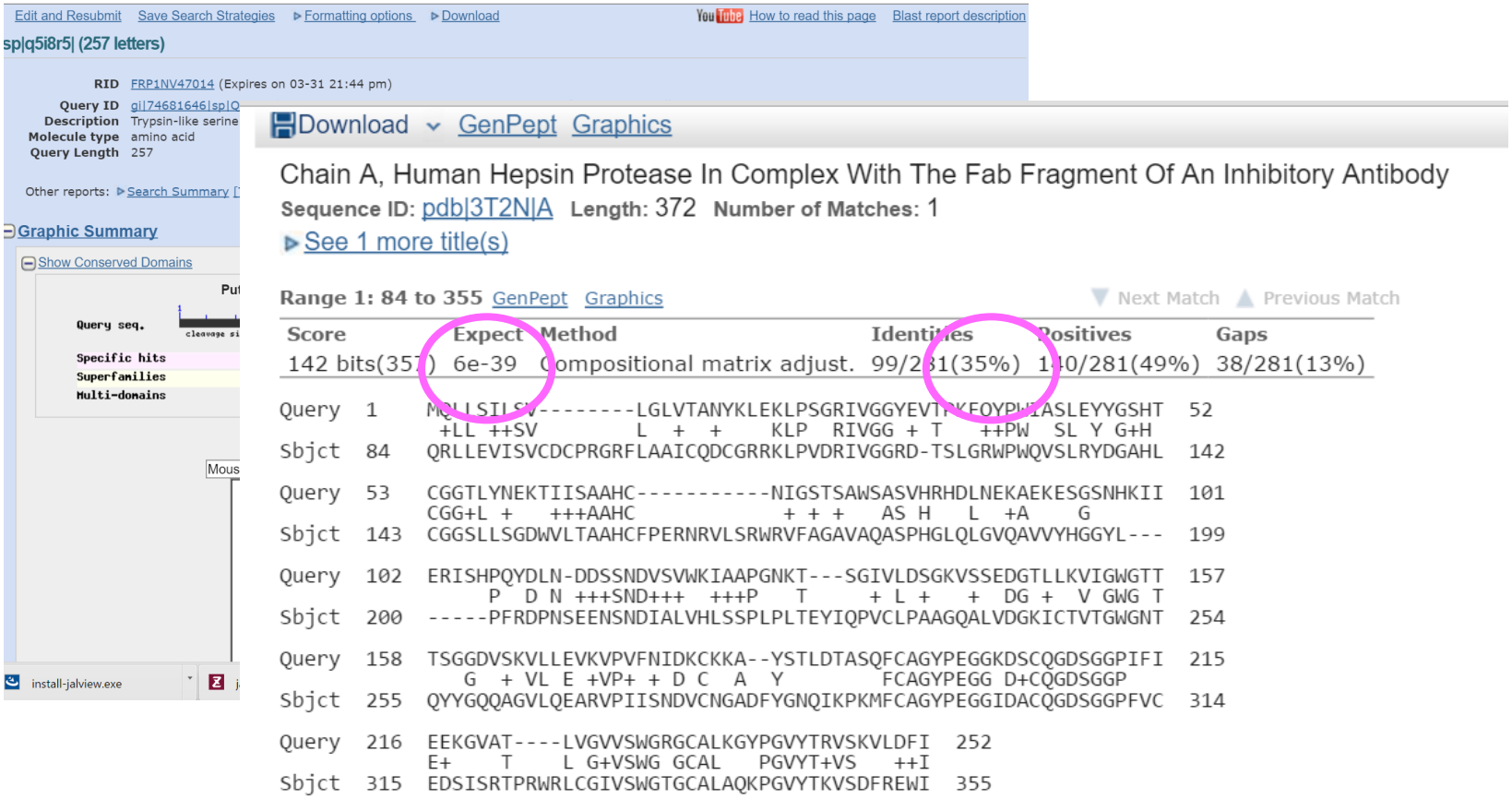
Which parts of your crystallised protein might enable phasing by MR, or experimentally?

Recognising folded domains in your sequence

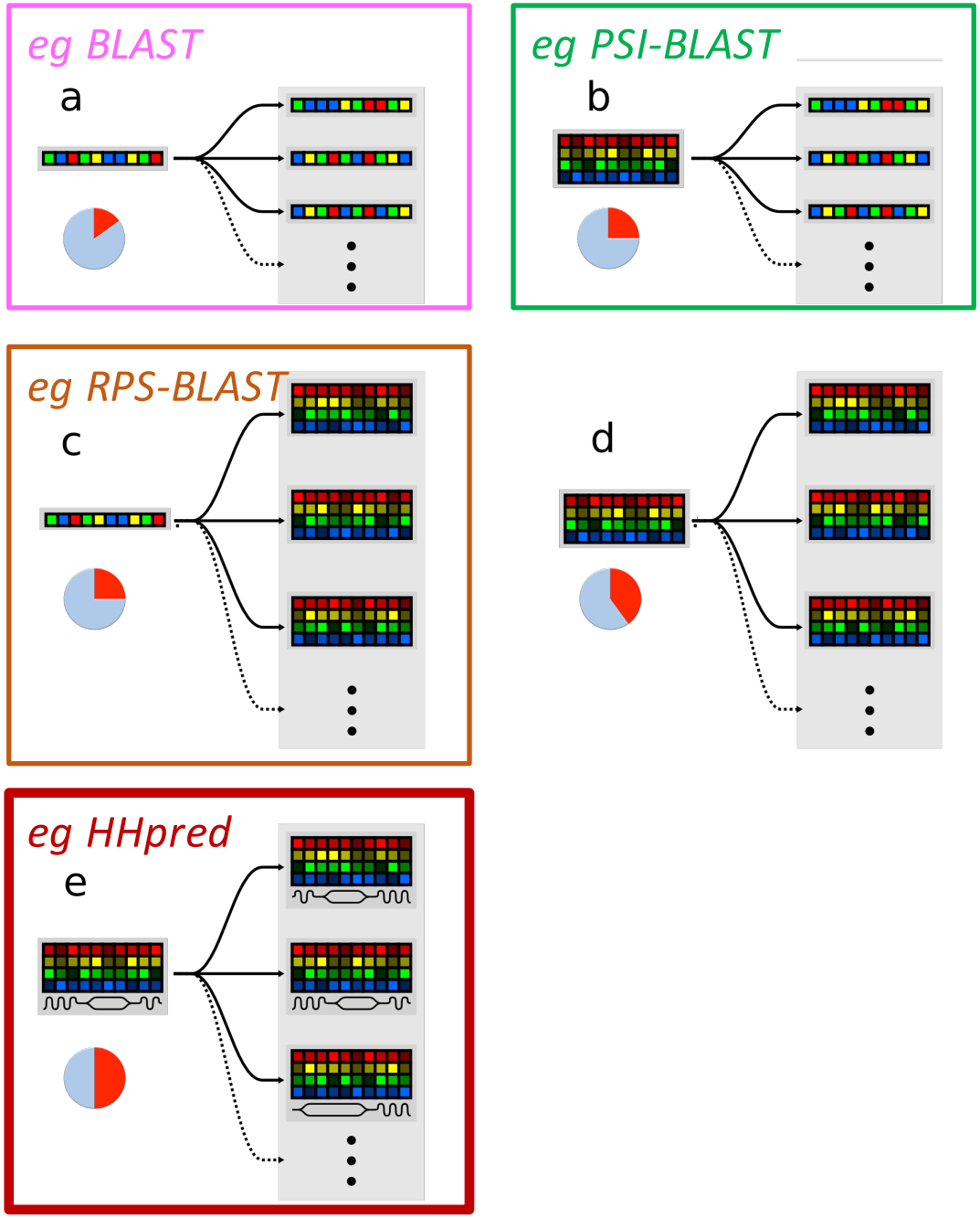
- How novel is your protein target? Recognising distant homology might make it less (or more!) interesting
- You might get extra ideas about its function to guide lab experiments, co-crystallisation, phasing (eg metal-binding sites)
- You might find the whole protein will not express/stay soluble/crystallise 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

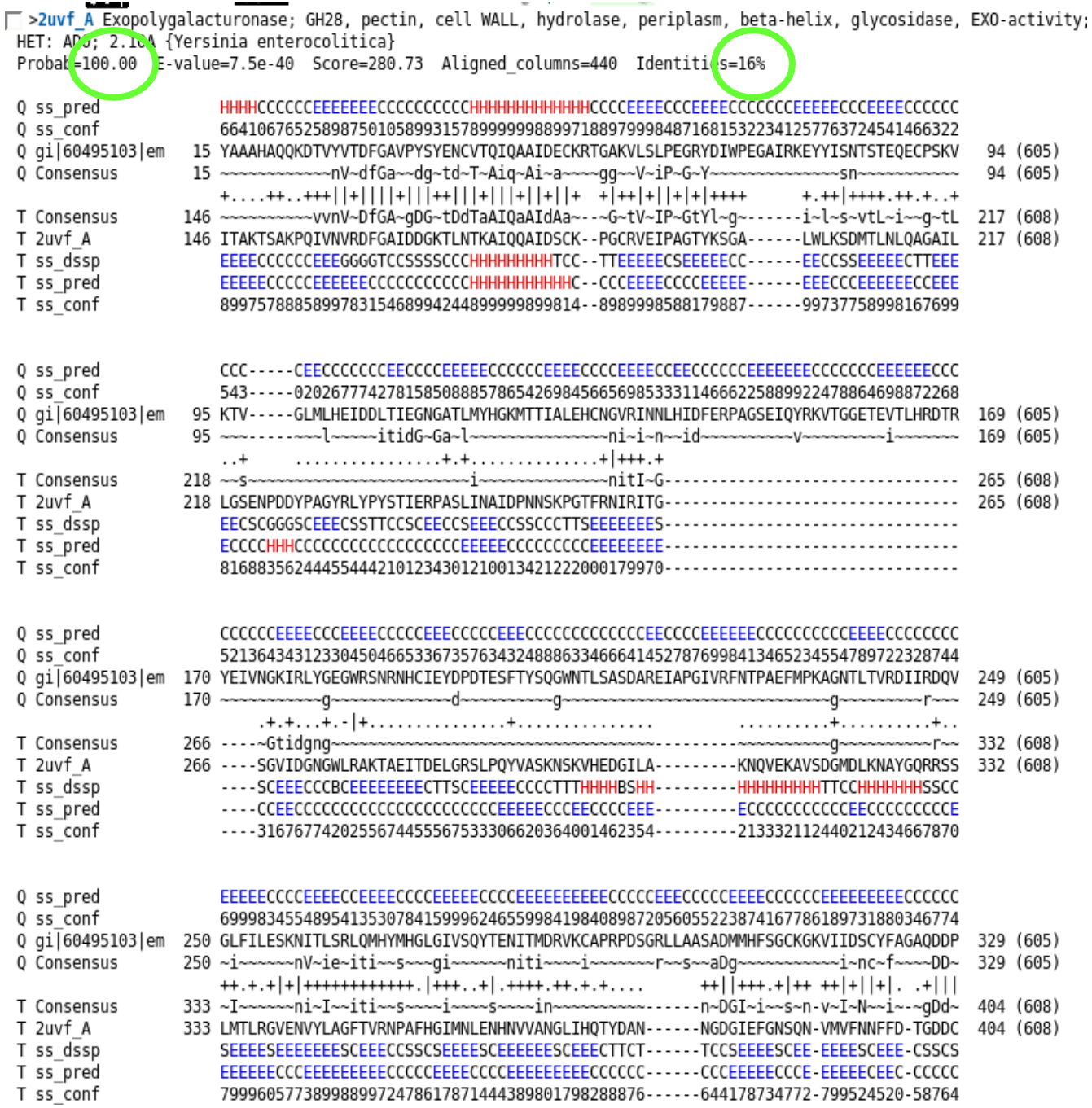


- Harder cases require a more sensitive tool. I recommend HHpred. Used by MrBUMP to find homologues to use as search models

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

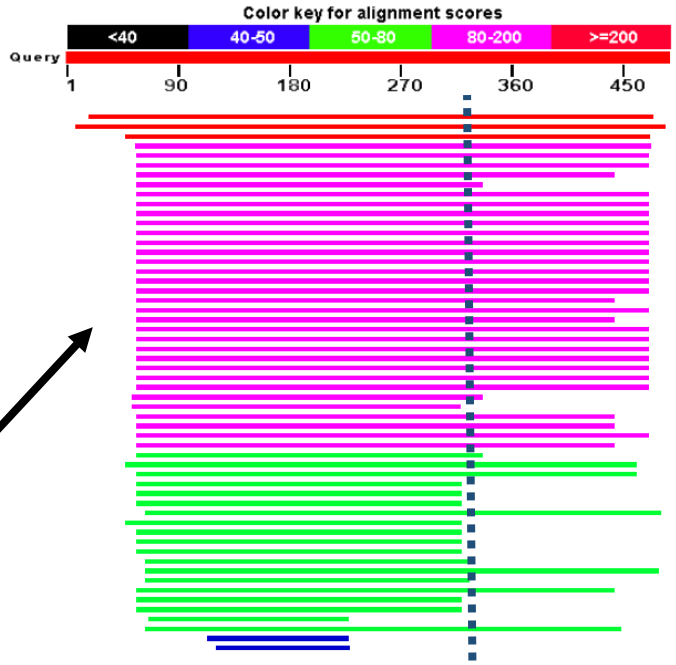
- Probability (0-100) is generally a good guide….
- …but statistics can mislead for unusual protein sequences eg coiled-coil, low-complexity, Cys-rich
- Consider if the match makes biological sense!
- Look at the matched region
  - Is it a complete domain/structure?
  - If partial, could it reasonably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

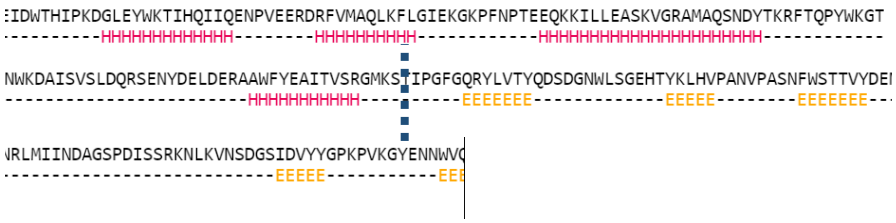
Recognising matches to domains in **sequence** databases, Pfam, Smart etc

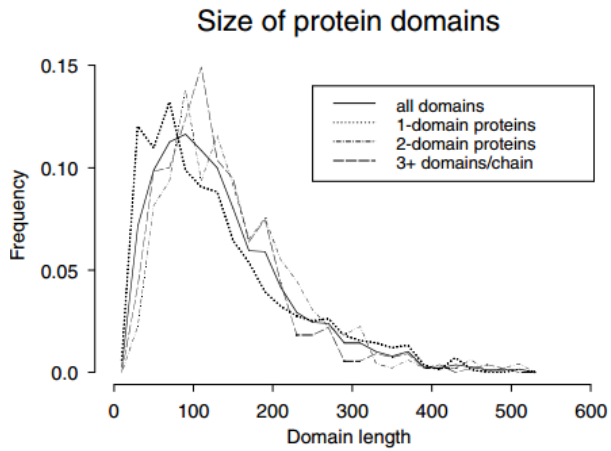
- HHpred is also an excellent, sensitive way to search against these. The same rules for interpretation apply. Matches, even distant ones, can obviously shed light on function
- However, for **structural** purposes need to remember that
  - Many Pfam families don't have structures so that domain limits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

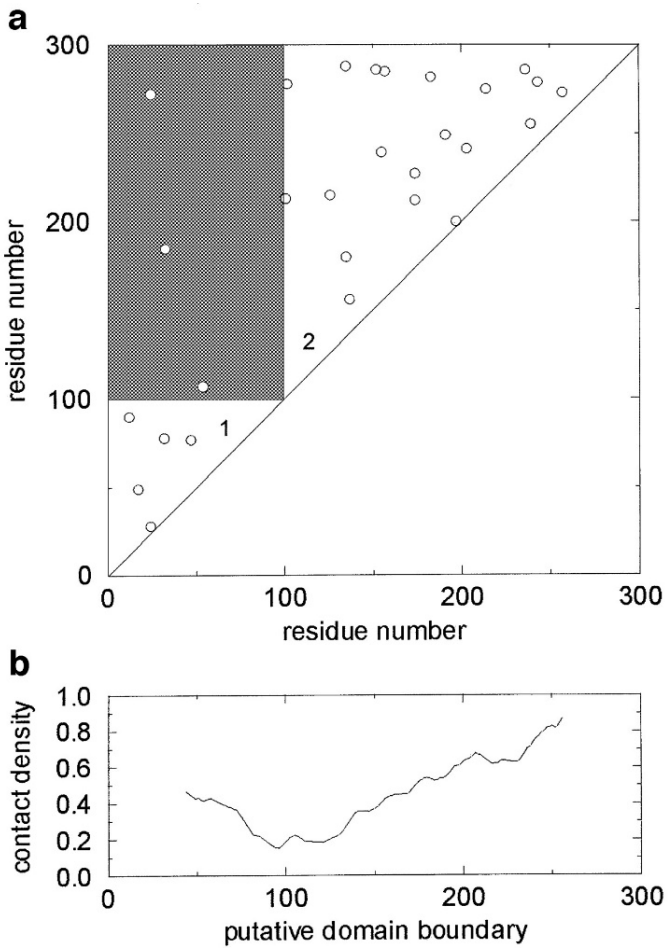


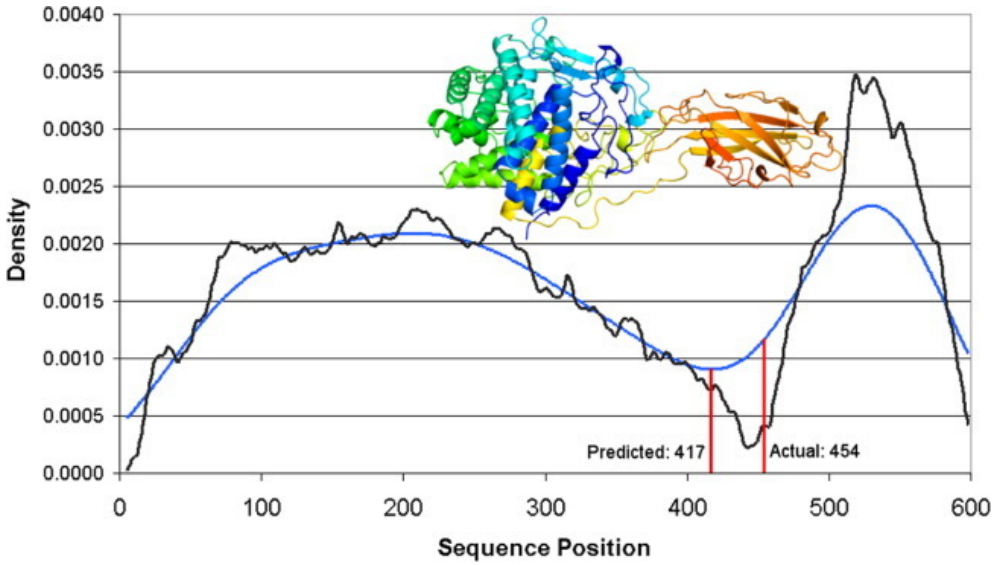




Predicted contacts for defining domains

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





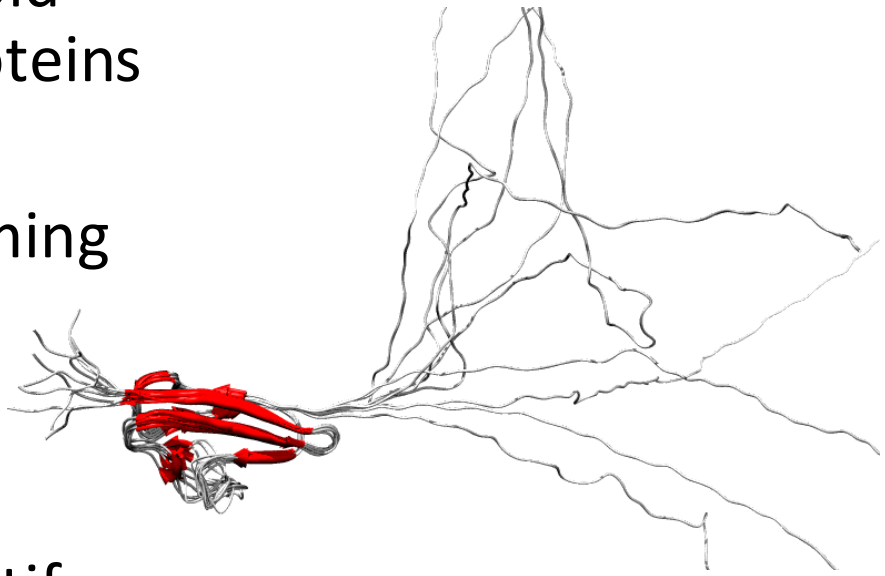
With today's contact predictions, it is now about the best method and works on multiple domains too

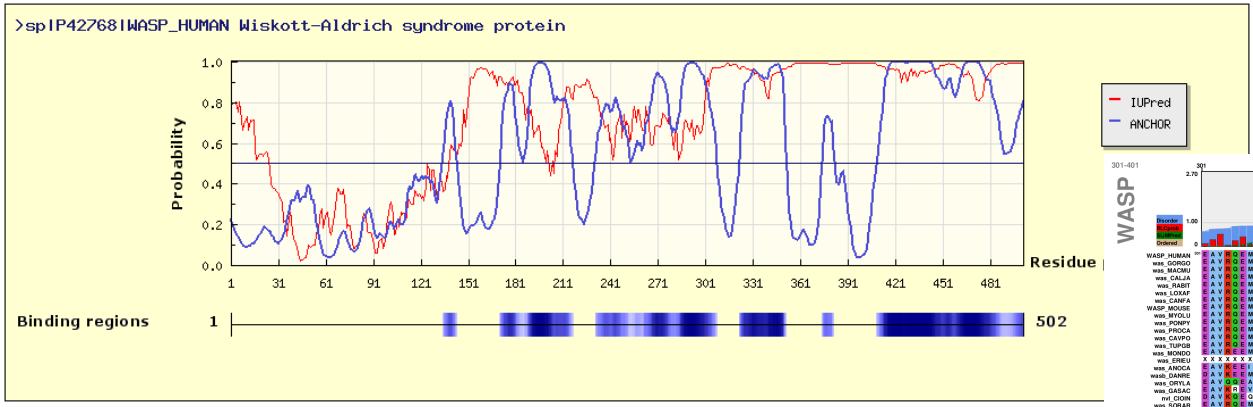
Rigden (2002) Prot. Eng. 15, 65

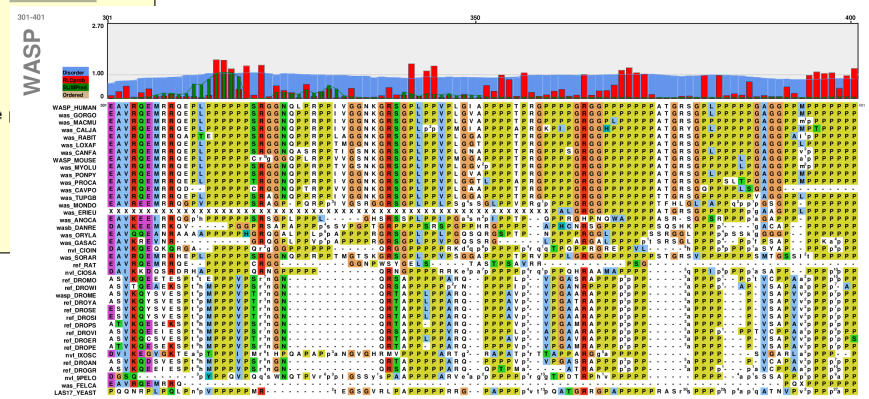
Sadowski (2013) Proteins. 81, 253

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)





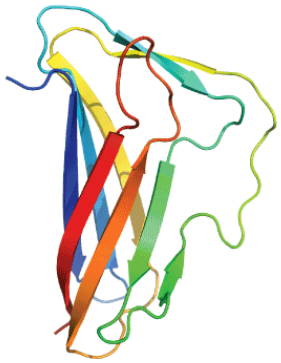


Predicting protein-protein interactions

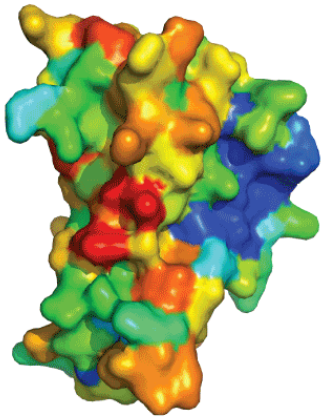
- Relevant to MR eg proteins A and B are cocrystallised but neither alone solves. An accurately predicted complex, being larger, might solve
- Many methods predict complexes based on steric complementarity plus other scoring functions
- Recommendable servers include
  - ClusPro, the best performing docking 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

B. subtilis IPI docking to protease

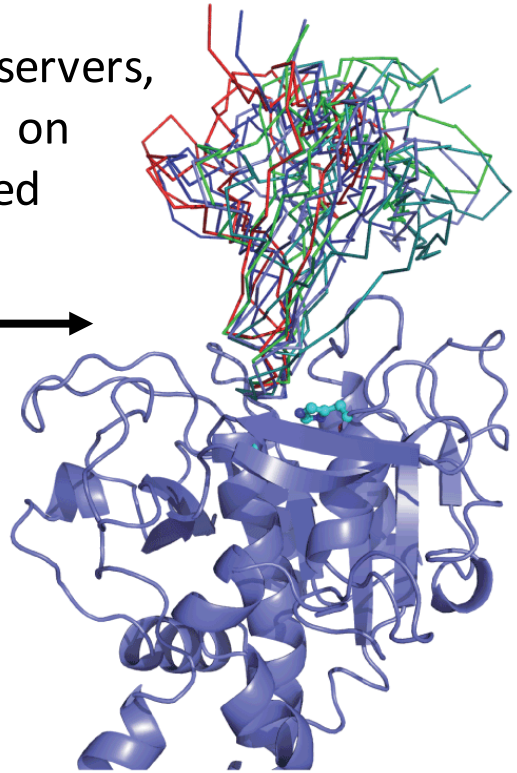
A



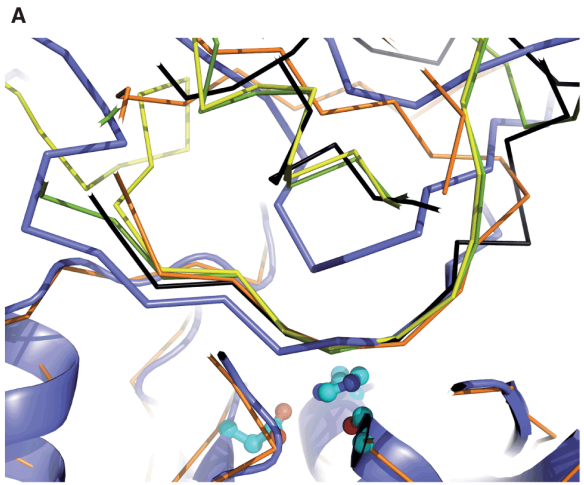
B



Several servers, focusing on conserved region



Final mode resembles other inhibitors

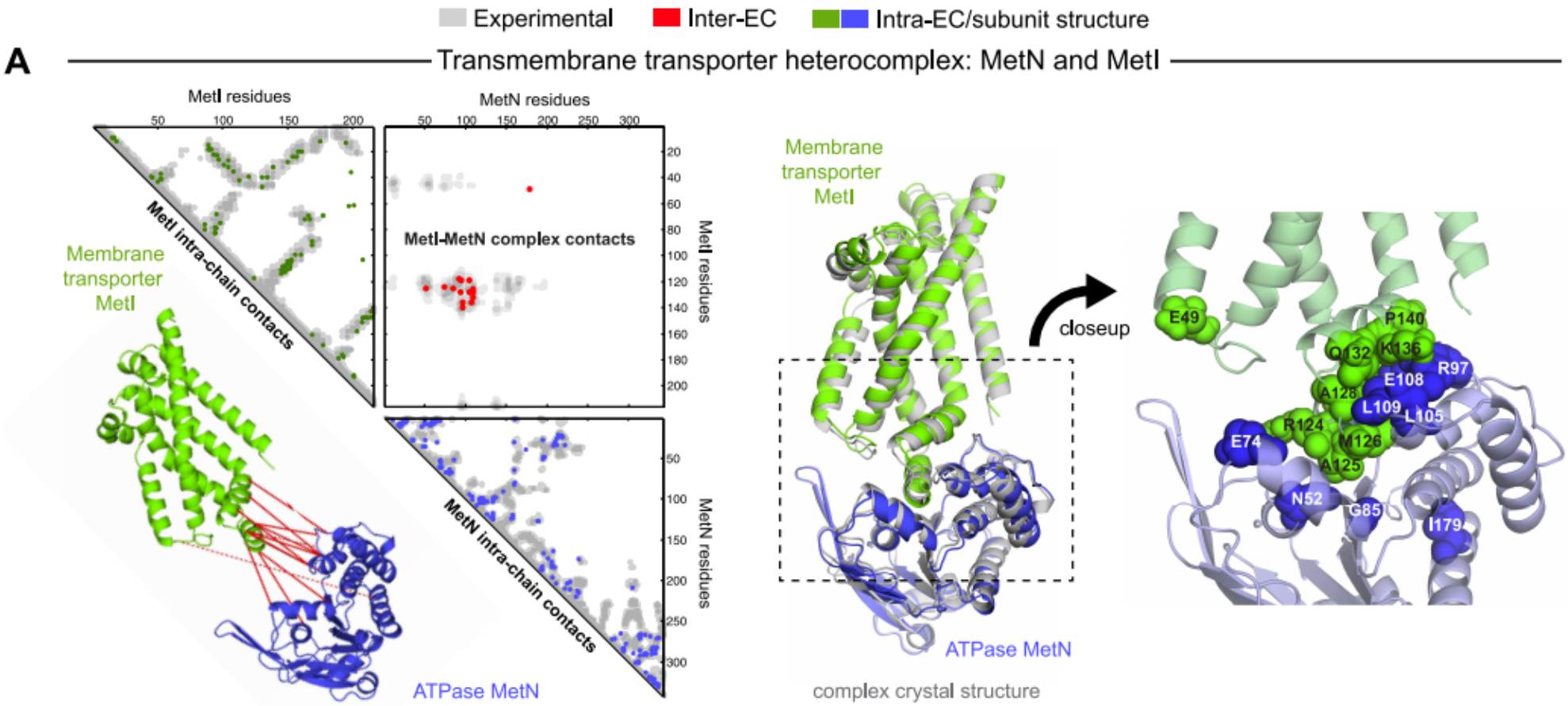


B

3isy_IPI	62	K	M	F	T	Q	A	F	68
2sic	69	V	M	C	P	M	V	Y	75
3qtl	45	S	V	C	T	E	E	Y	51
1oyv	58	N	A	C	T	F	N	C	64
1sbn	41	S	P	V	T	R	D	L	47

Rigden et al. (2013) F1000Research 2:154

Using predicted contacts to help predict complexes



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

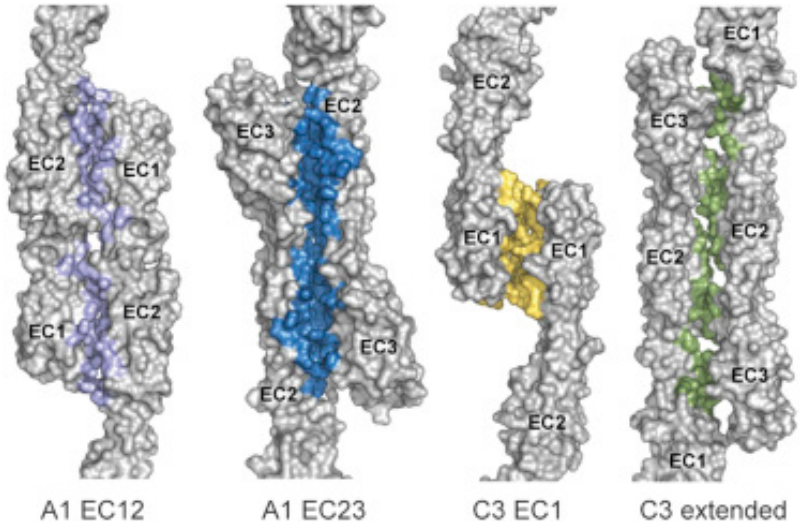
… and once you have your crystal structure…

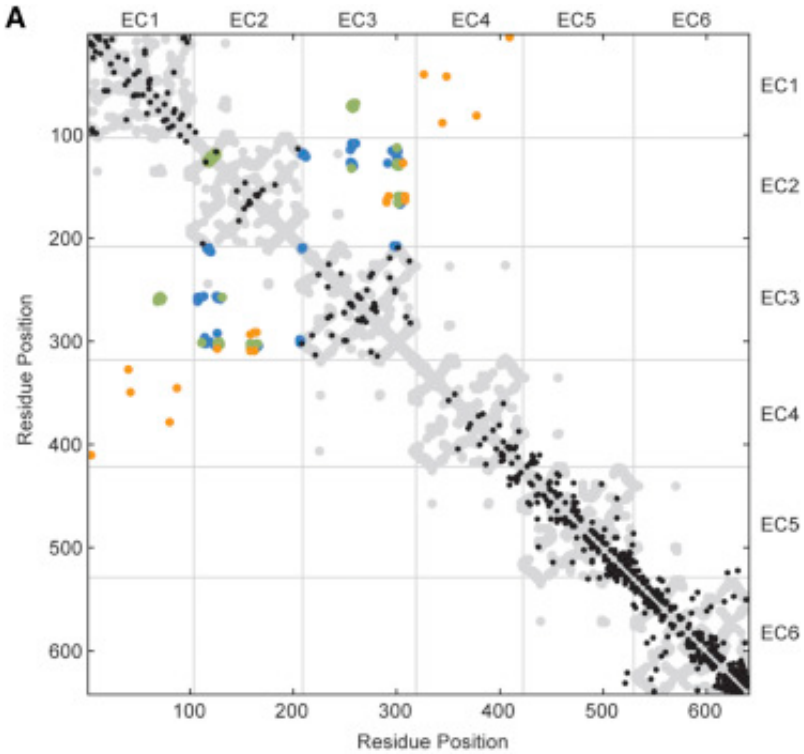
What is the biologically relevant quaternary structure?

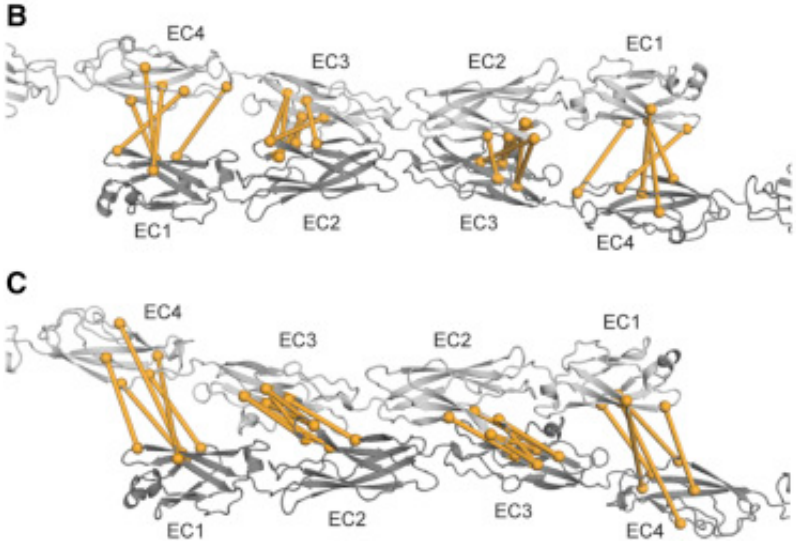
Where are the functional/catalytic sites?

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







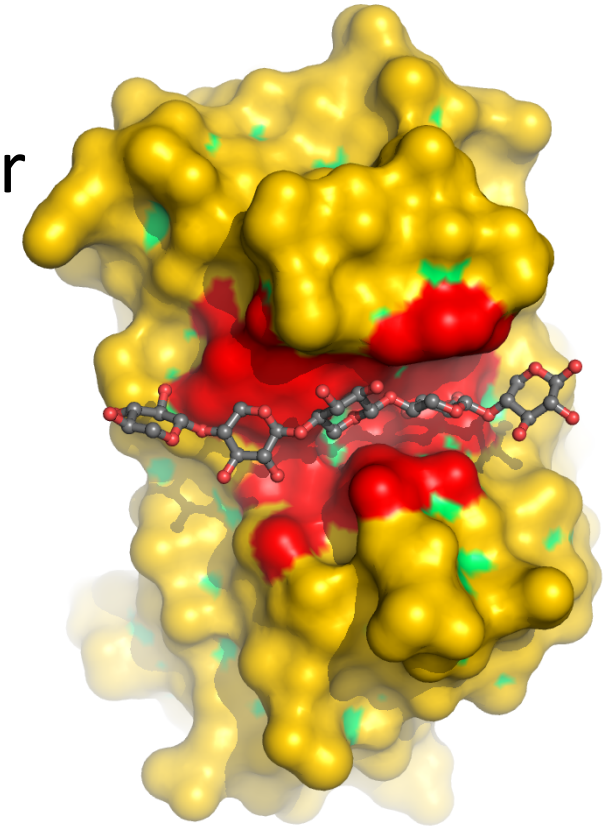
Nicoludis *et al*. (2015) Structure 23, 2087

Some lesser-known structure-based function annotation methods

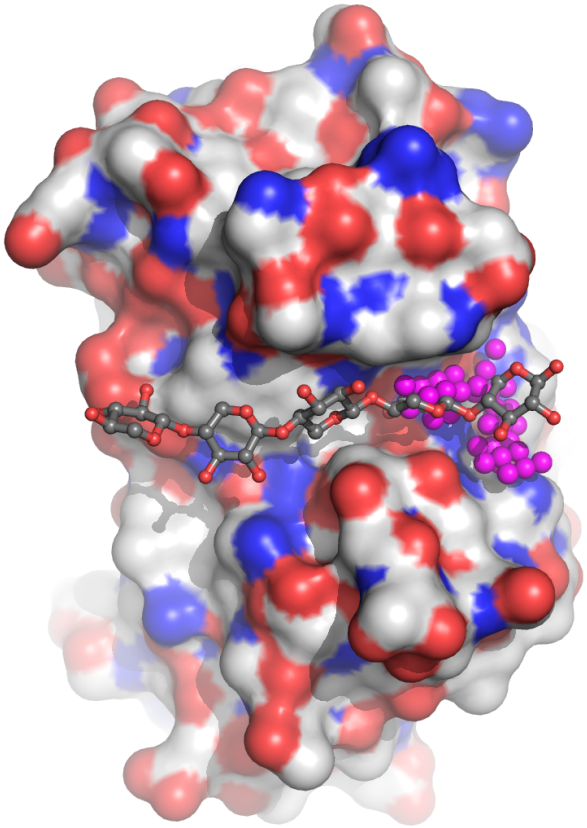
- Finding functional sites is based on their being different somehow to the rest of the protein surface. Important generic methods are based on
  - Shape
  - Electrostatics
  - Physico-chemical characteristics
  - Evolutionary conservation (Consurf)
- Less well-known but valuable characteristics are
  - Statistics of surface atom 'triangles' (STP)
  - Probe interaction energetics (ISMBLab)
  - Rigidity and geometry (EXIA2)
  - 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



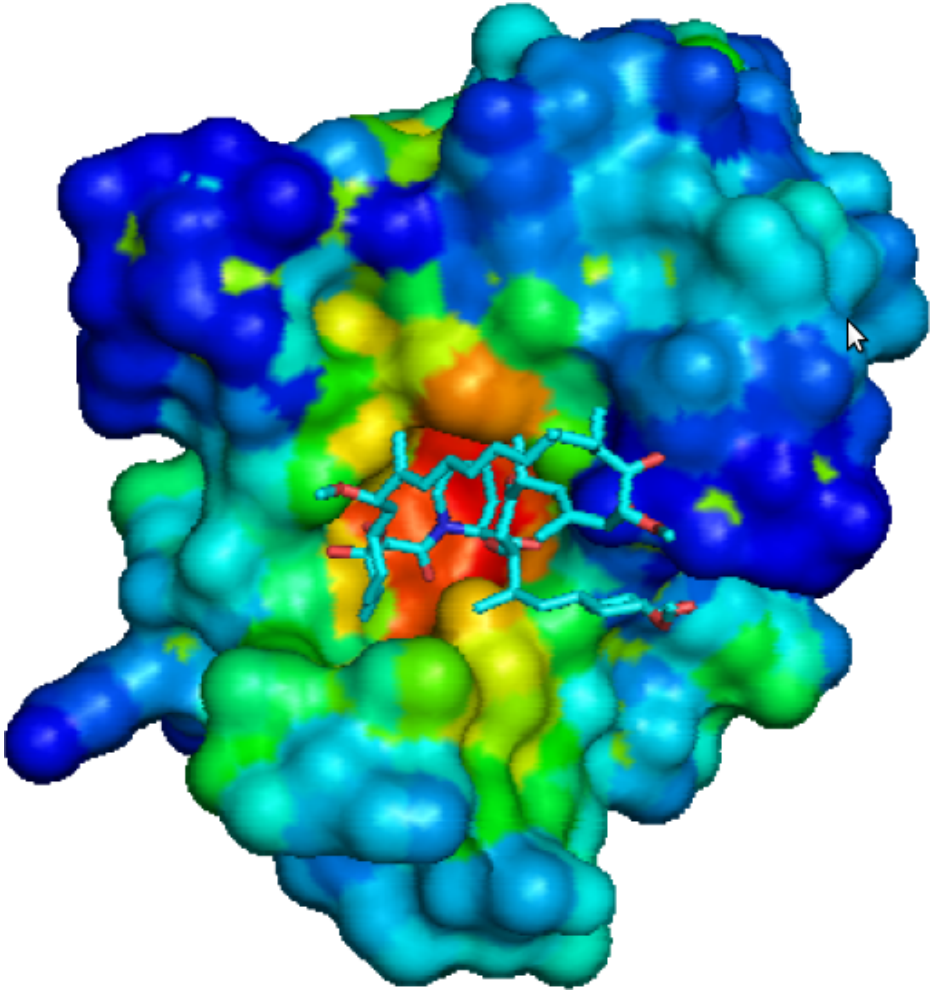
SiteHound

ismblab.genomics.sinica.edu.tw

scbx.mssm.edu/sitehound

Binding sites from statistics

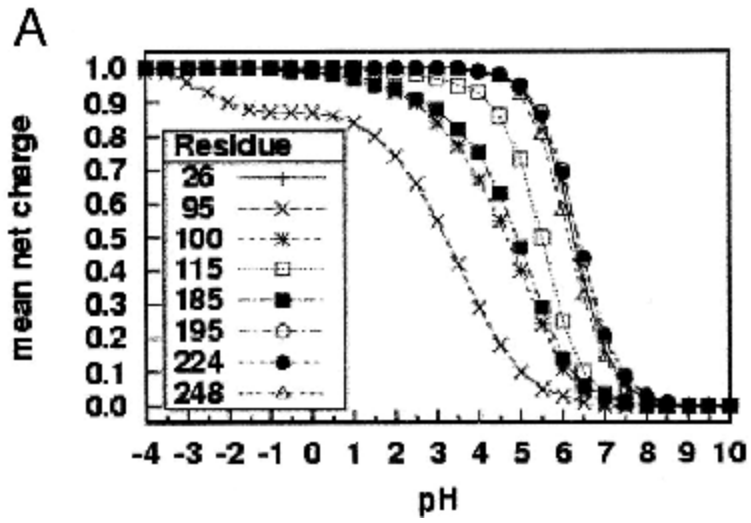
- STP (surface triplet propensities)
- 13 atom types → 455 triplets
- Distribution in binding vs non-binding sites varies
- Designed for small molecules, works on PPIs, including flat surfaces



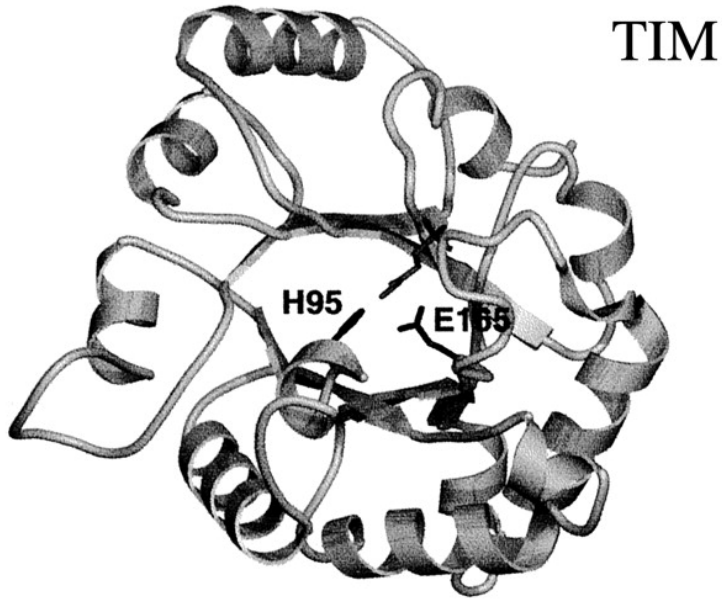
opus.bch.ed.ac.uk/stp

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

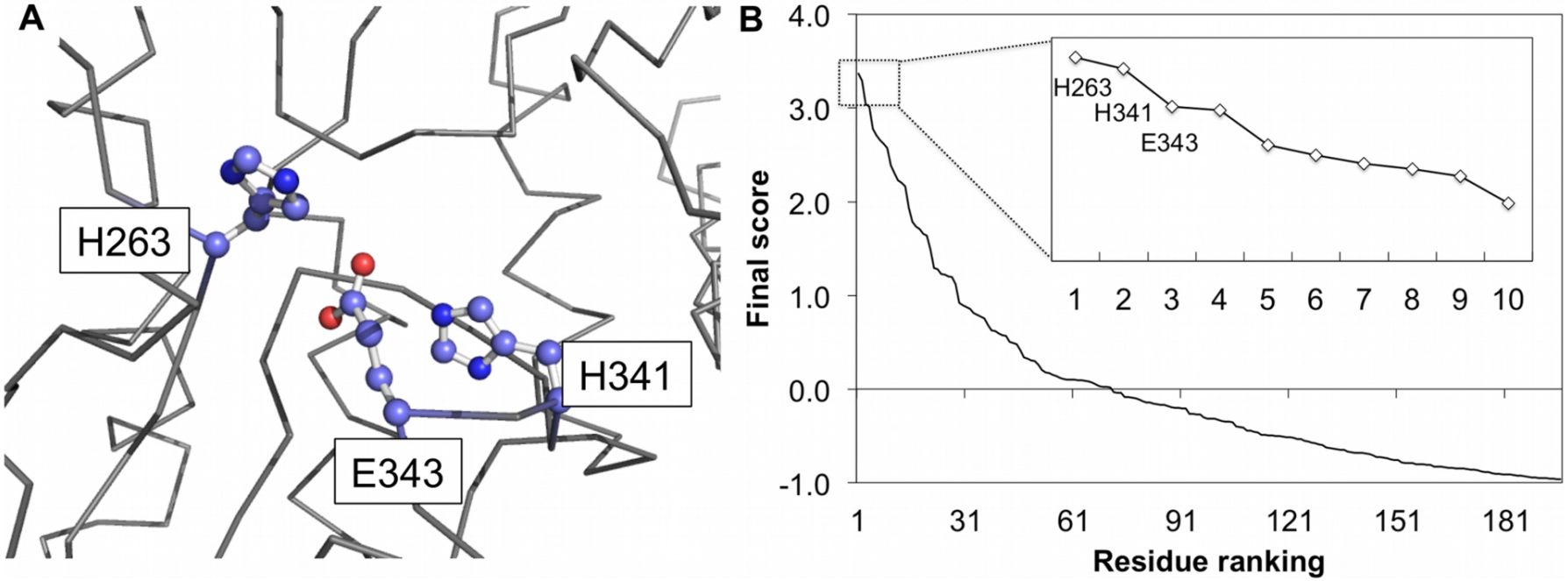


His95 and other residues with atypical pKa cluster at catalytic site

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

EXIA2: side chain orientation and rigidity

- Find points on the surface with many side chains 'pointing at them'
- Weights these further according to predicted rigidity (measured as number of contacts). Catalytic residues tend to be more packed and so more rigid.

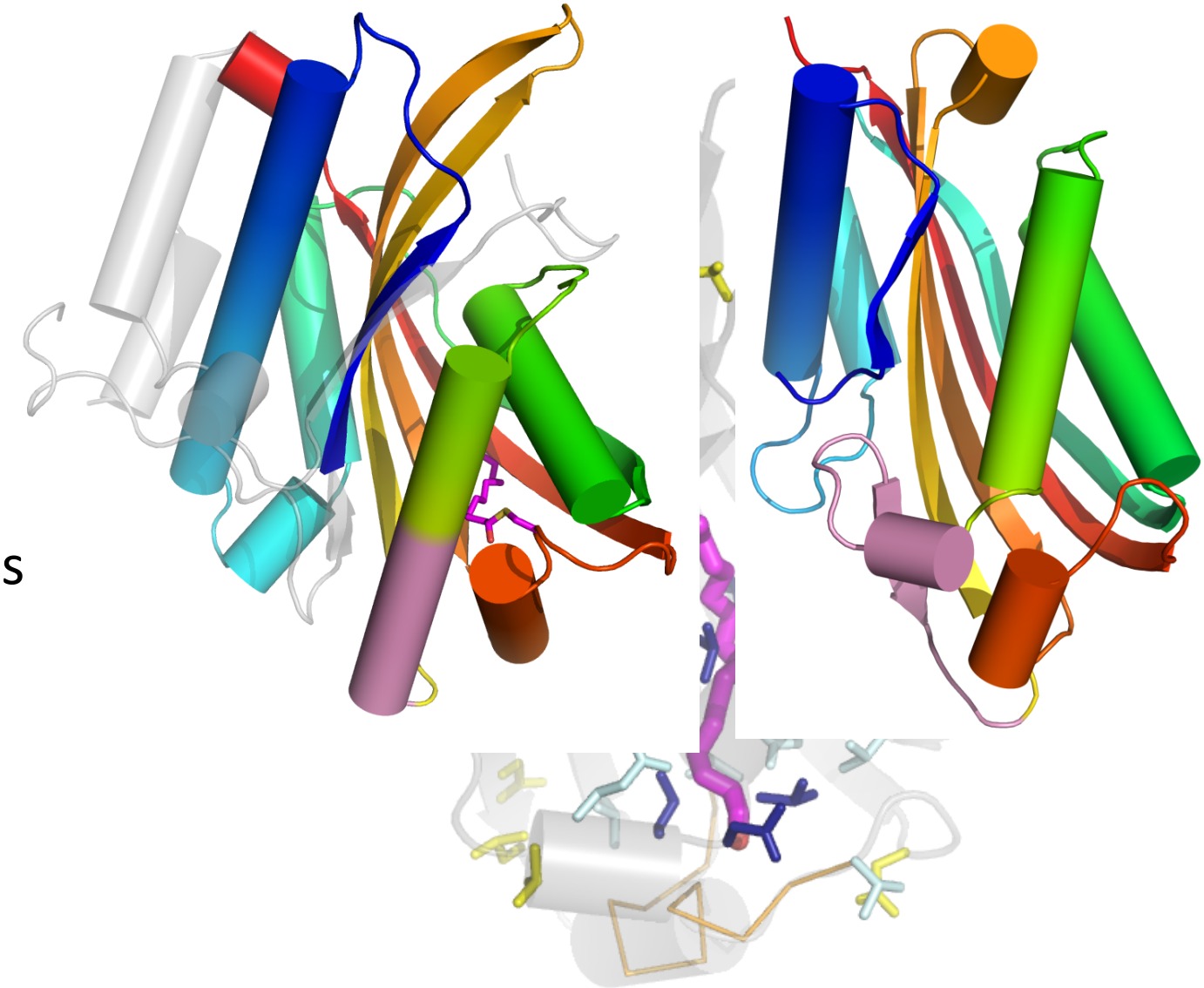


THR12

http://203.64.84.196

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...



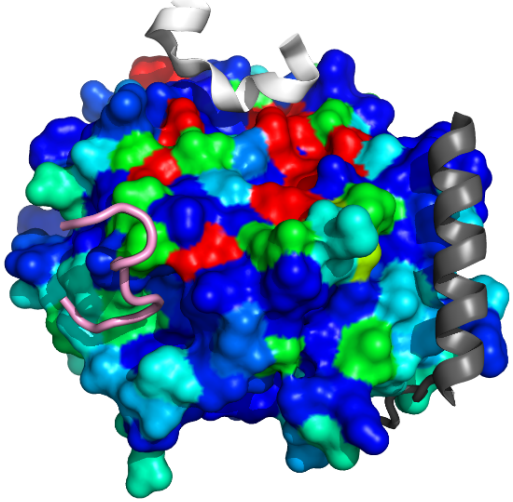
Quevillon-Cheruel et al. (2009) J. Biol. Chem. **35**,2107

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

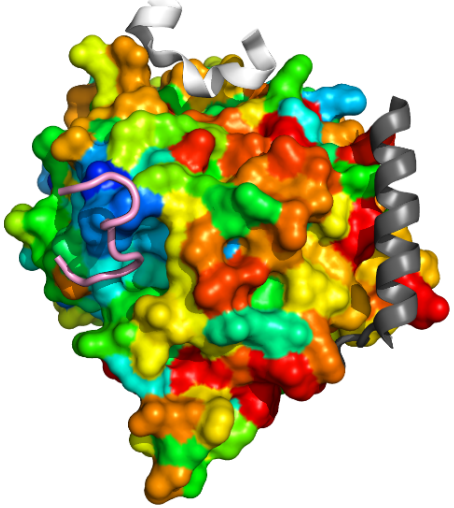
Some servers require thought…

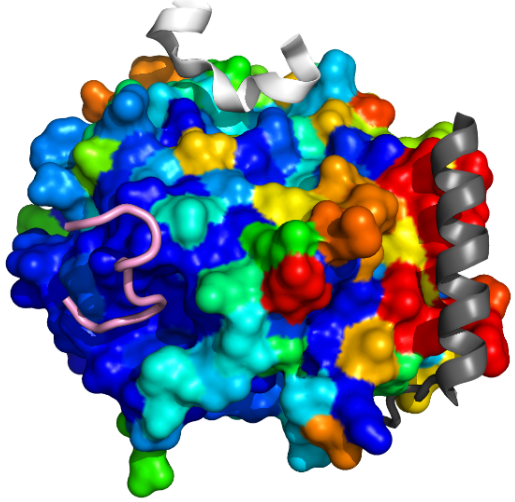
- Consurf maps sequence conservation onto a structure revealing functional sites
- Excellent, general method, but results depend on sequence set chosen for mapping: selecting all or only near relatives gives different results. Either might be more appropriate for you

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





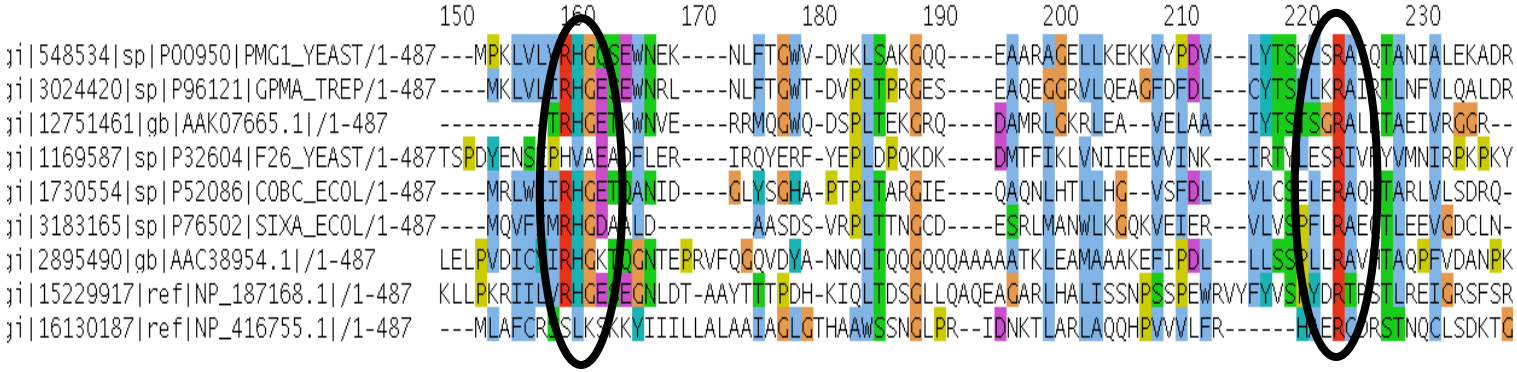
…and finally, you're putting a manuscript together

Calculating and presenting sequence alignments

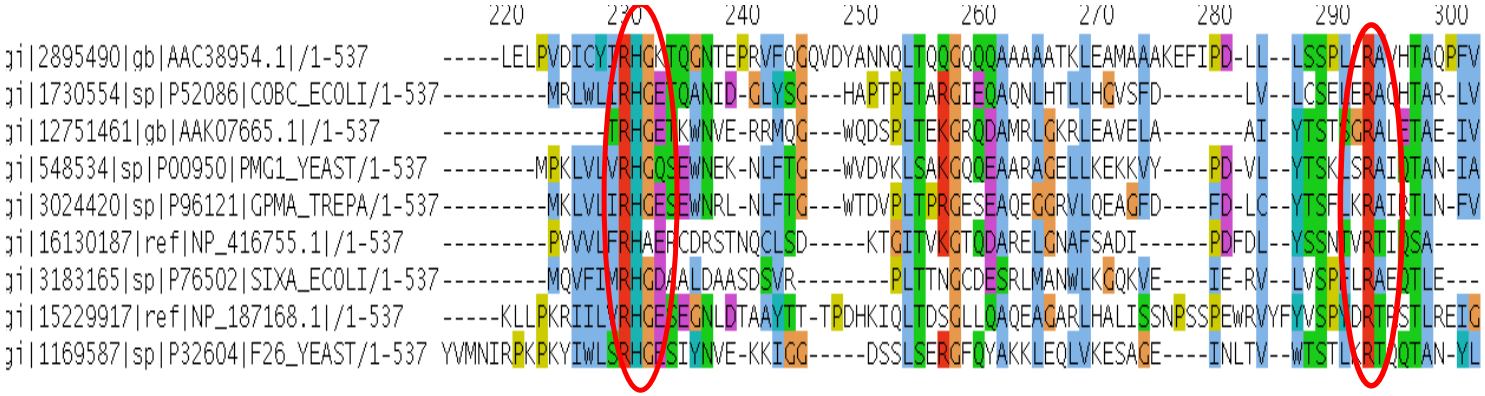
Your sequence alignment

- 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…

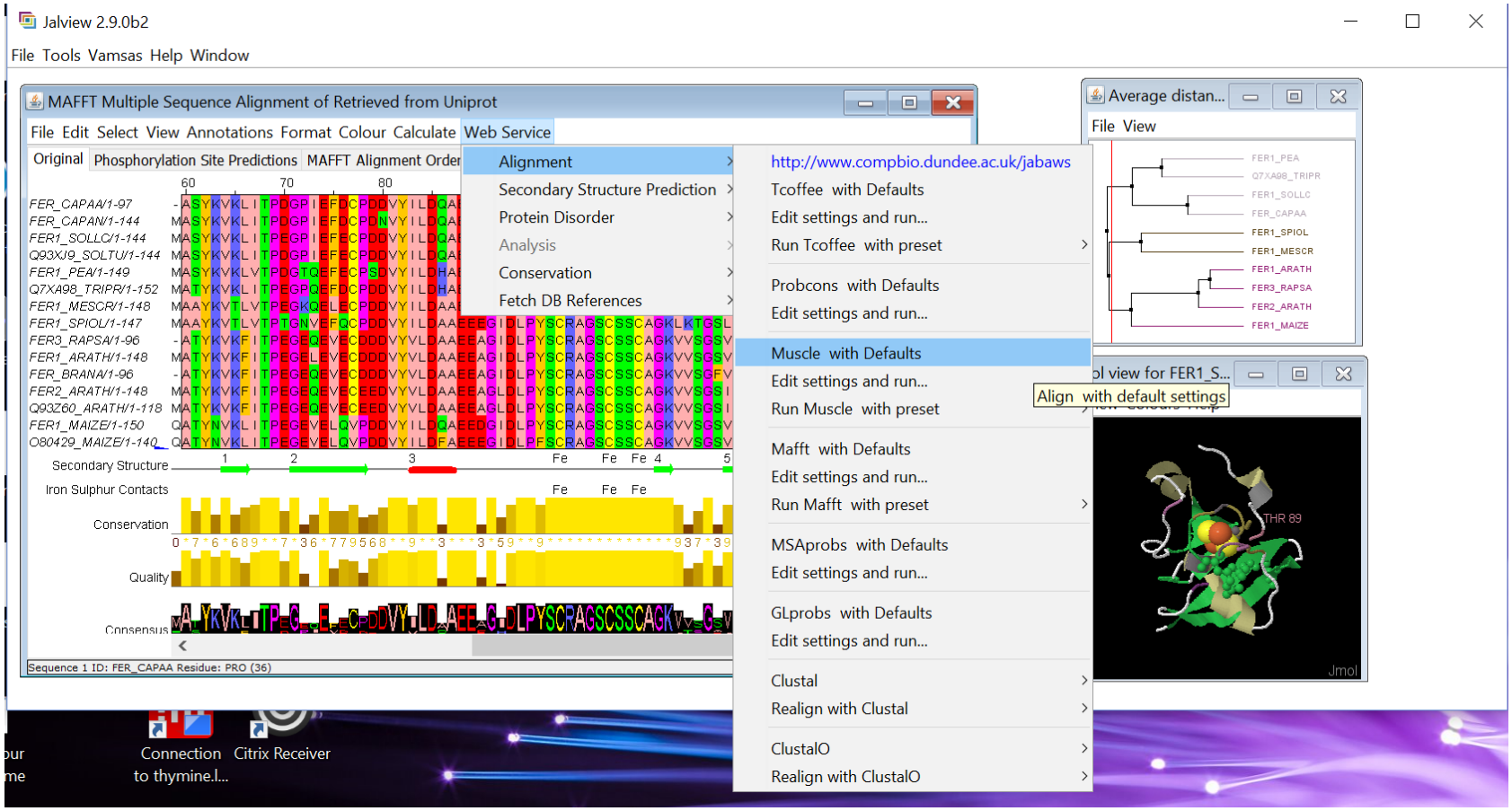


… but MUSCLE gets it



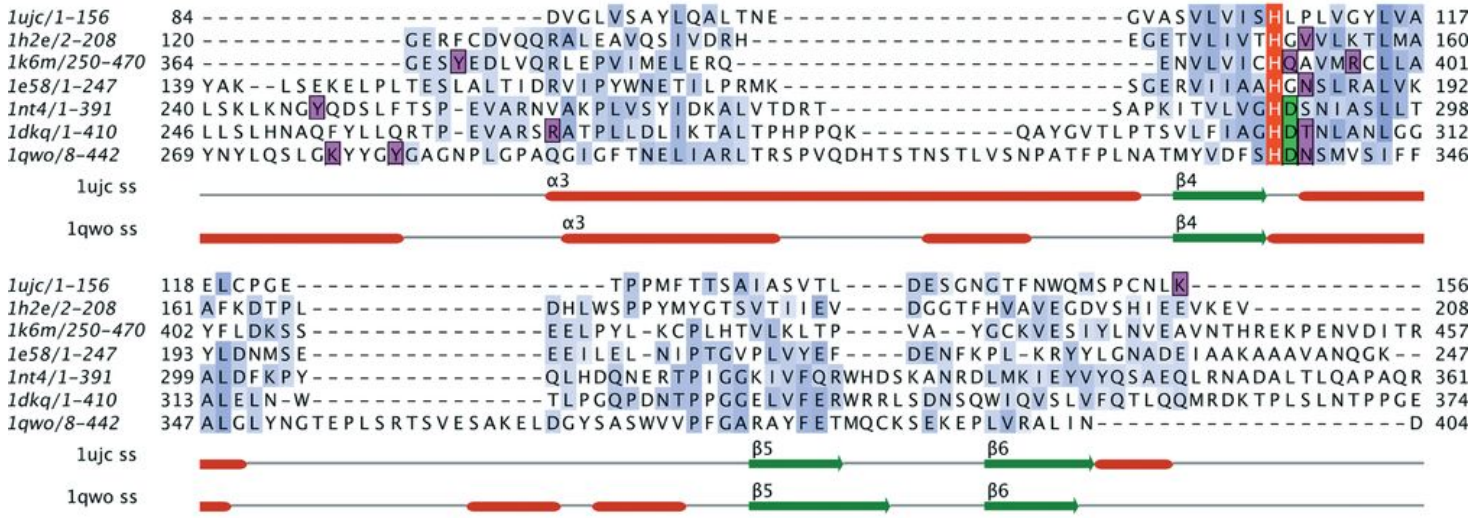
Jalview.org, recommended for sequence alignments

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



Jalview

- Also helps you produce figures like this…



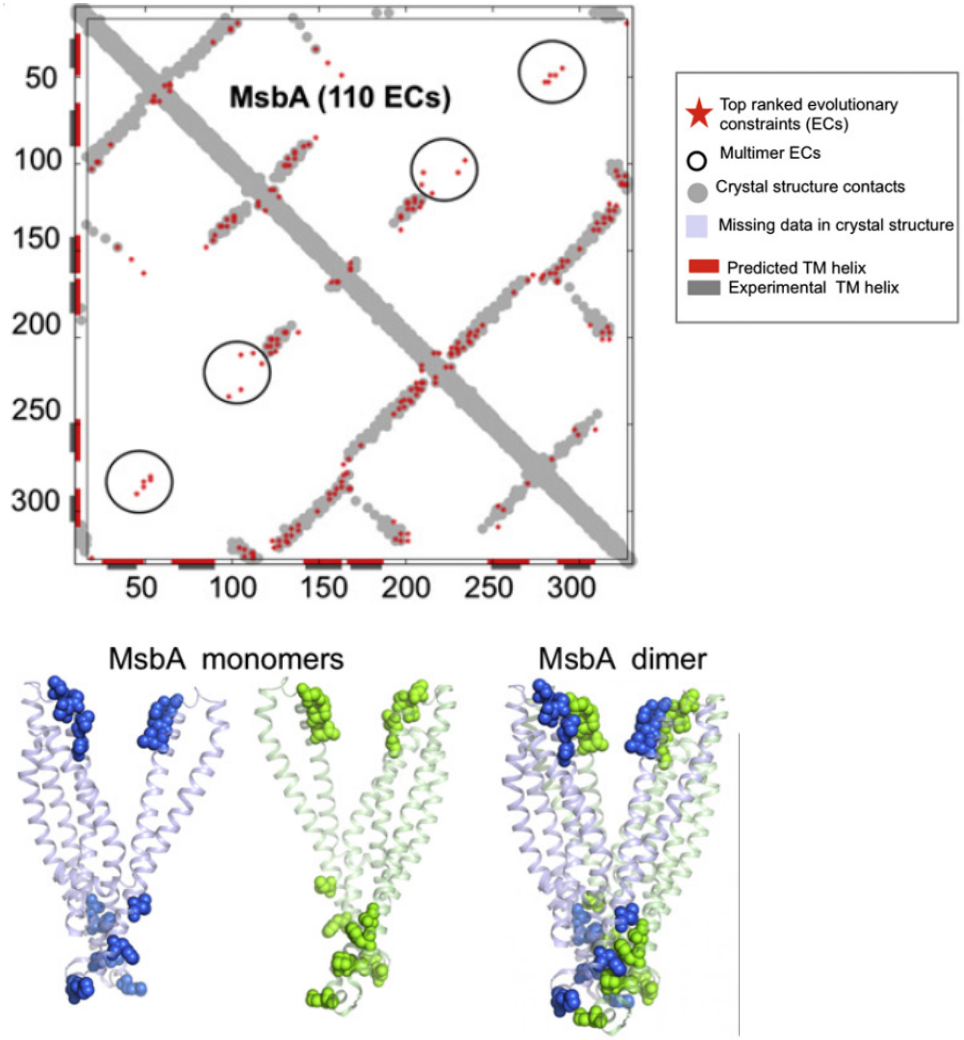
- … rather than like this

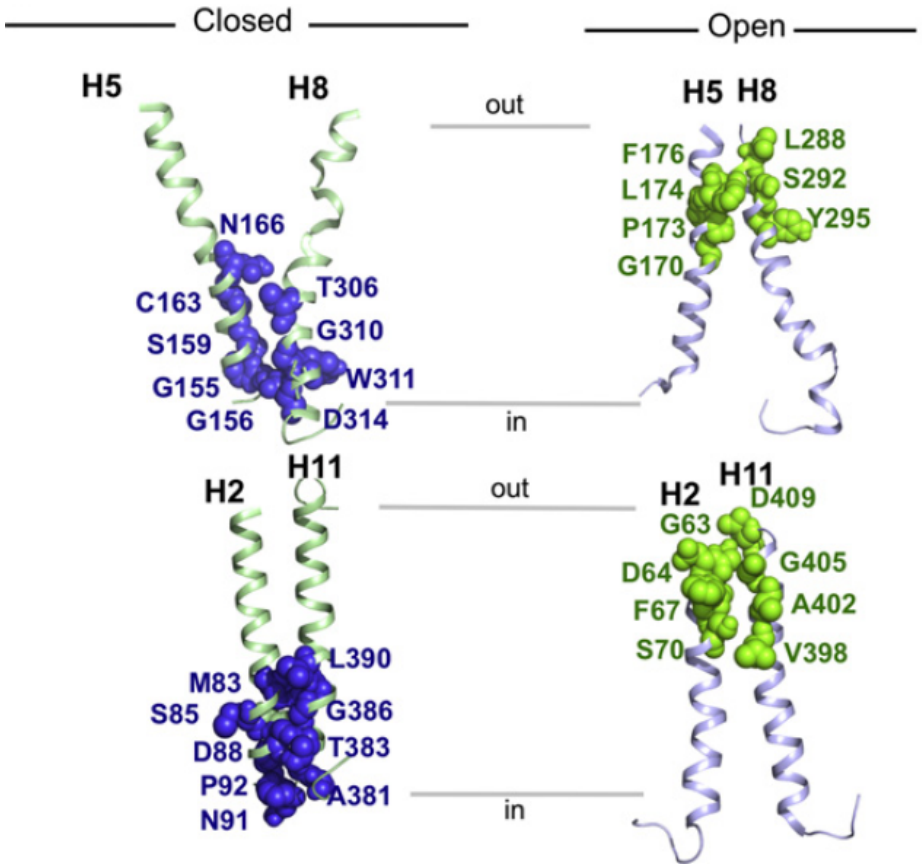
Questions? Feedback?

UNIVERSITY OF LIVERPOOL

Contacts also inform on…

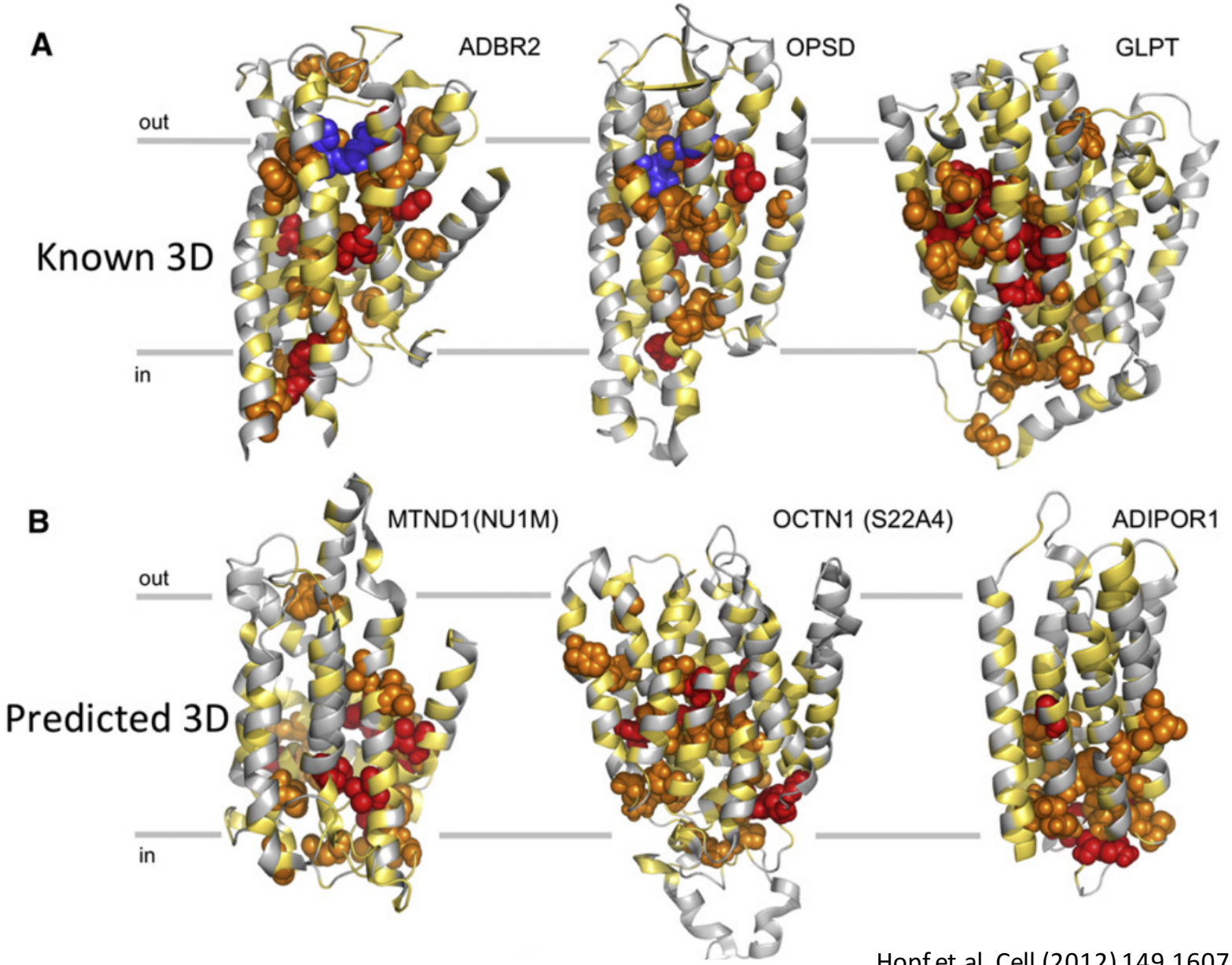
- Oligomerisation
- Conformational change





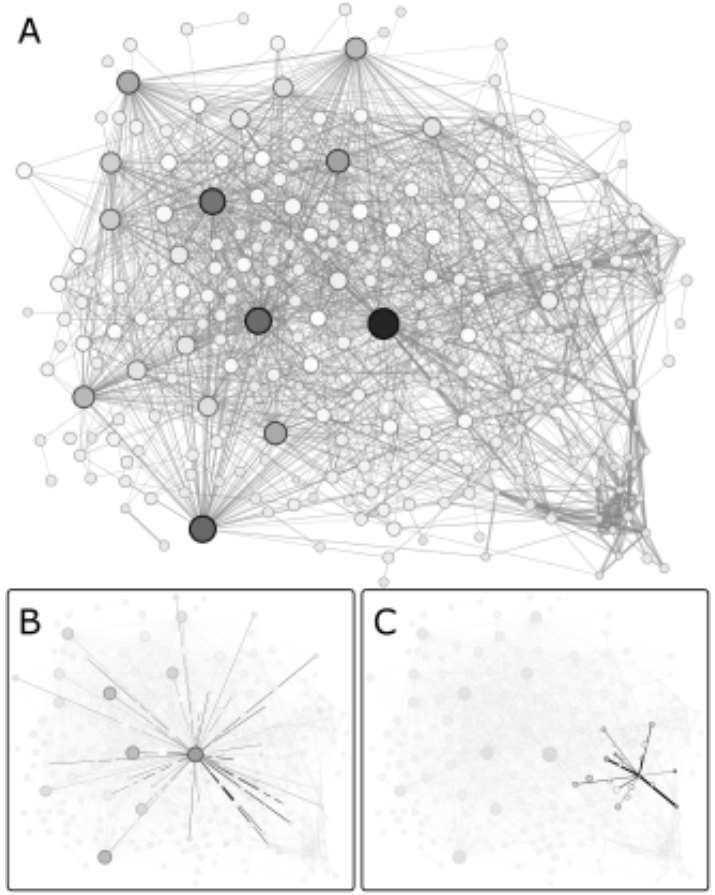
Hopf et al. Cell (2012) 149,1607

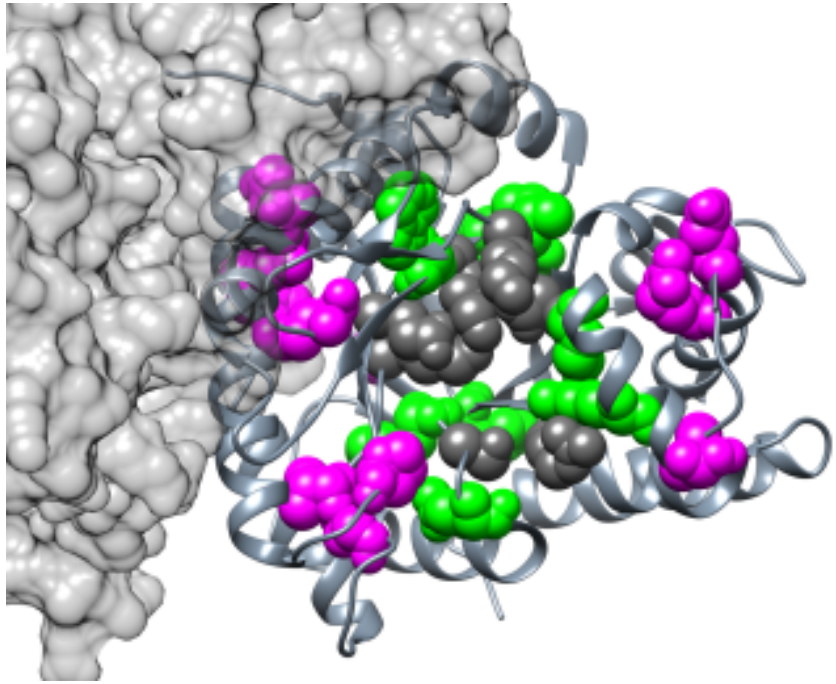
Predicting functional sites



Hopf et al. Cell (2012) 149,1607

Predicting functional sites





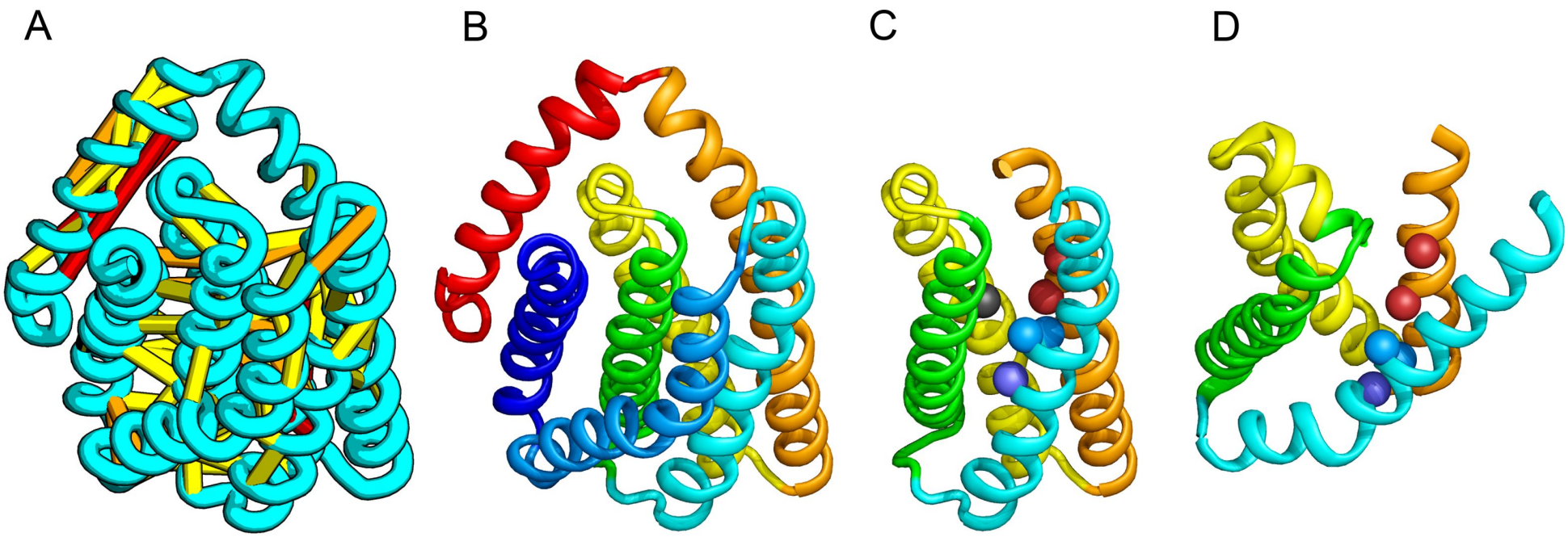
- Considering network structure of contact prediction map gives scores….

- … that pick out important sites more convincingly
- Grey, conserved catalytic (no covariance signal possible!)
- Green, neighbours of catalytic site; Magenta include interface residues

Parente et al. Proteins (2015) 12, 2293

You don’t know the structure

Predicted contacts for folding



Model showing satisfied and unsatisfied contacts

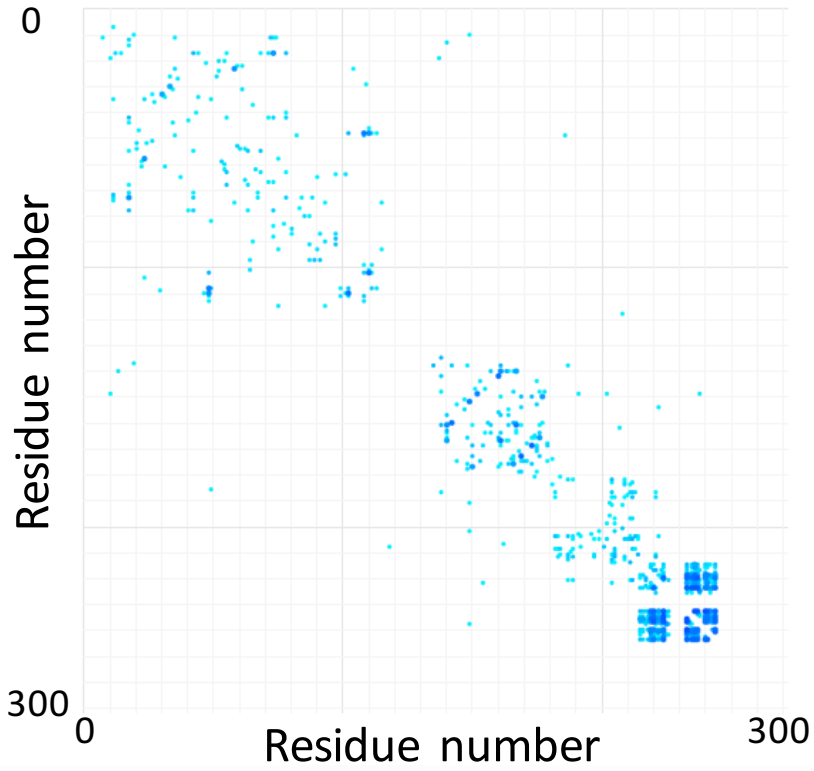
Substructure of model with conserved motifs

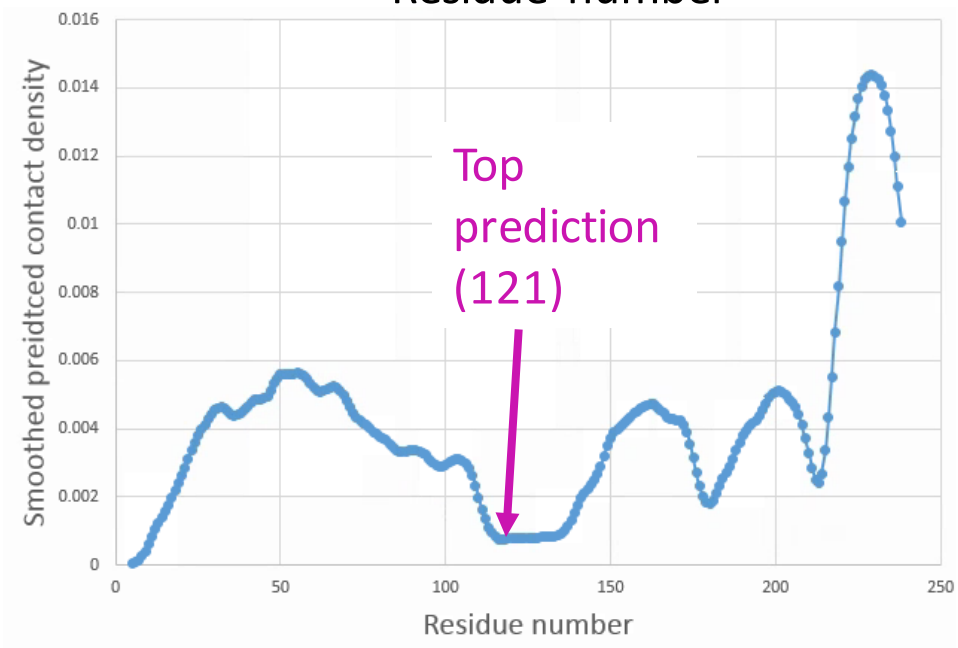
Crystal structure with similar conserved motifs

Ovchinnikov *et al.* (2015) eLife 4, e09248

An example from the original paper

- AL1 geminivirus protein of ~250 residues
- *"… LM1 of this profile lay at residue 132 … the depth of LM1 corresponds to an average error of approximately 19 residues. … the domain definition agrees very well with the functionally defined AL1 origin DNA-binding site domain from residues 1–116."*
- Domain now structurally defined as ~7-118

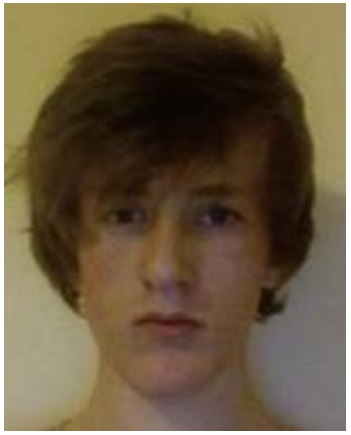




Rigden (2002) Prot. Eng. 15, 65

Sadowski (2013) Proteins. 81, 253

Bacterial competence and ComEC



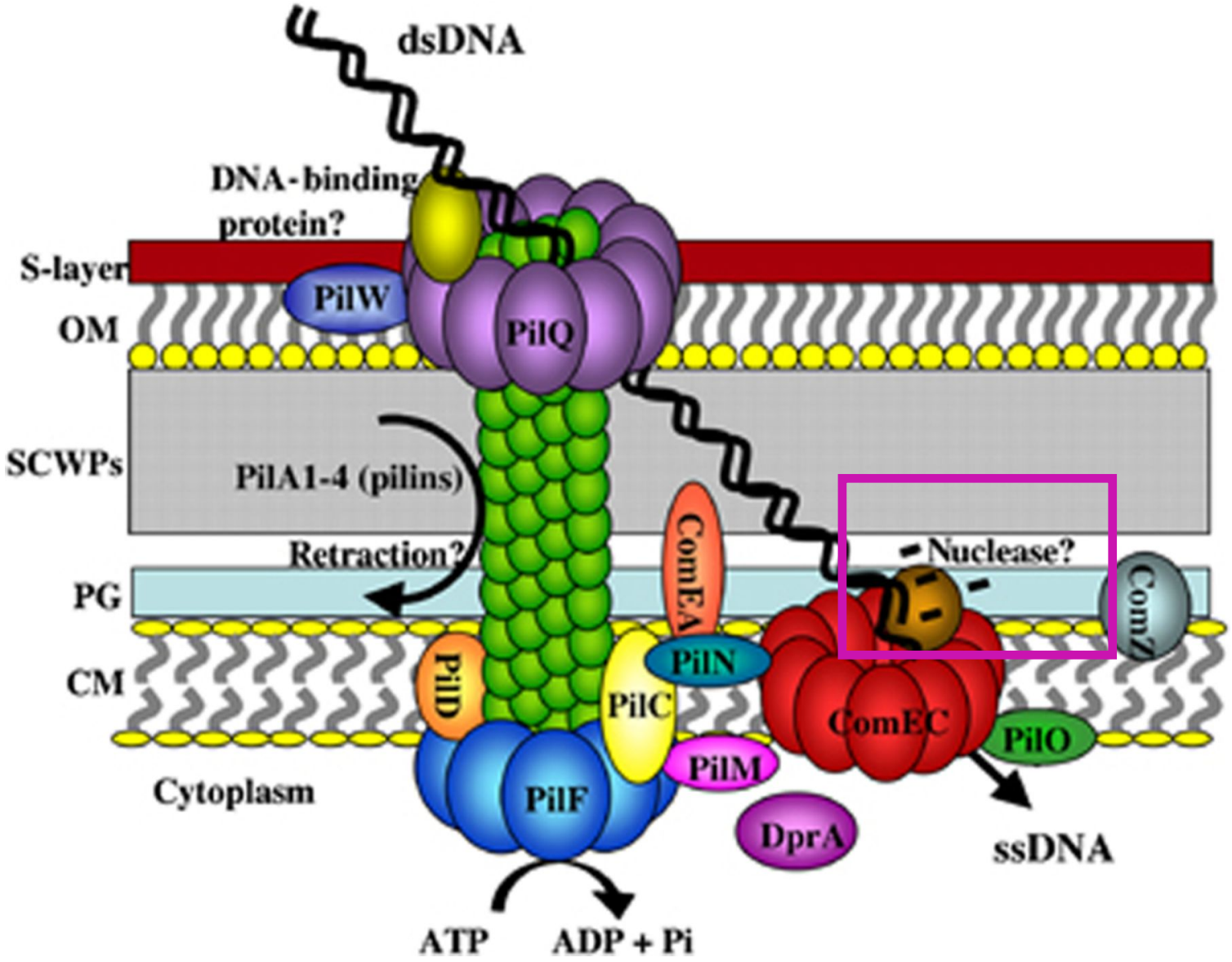
MBiol Project

BBSRC Research Experience Placement

Bacterial competence

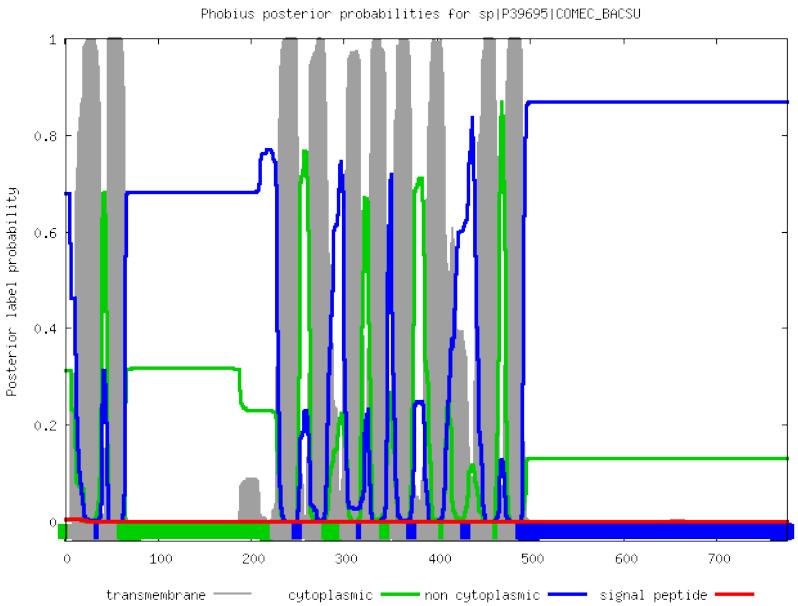
- The innate ability of some bacteria to take up extracellular DNA
- Proteins involved vary between species eg Gram +ve vs -ve
- Many poorly understood
- Bacterial competence involved in antibiotic resistance spread and pathogenicity of some bugs
- One of the proteins most strongly linked to competence is ComEC

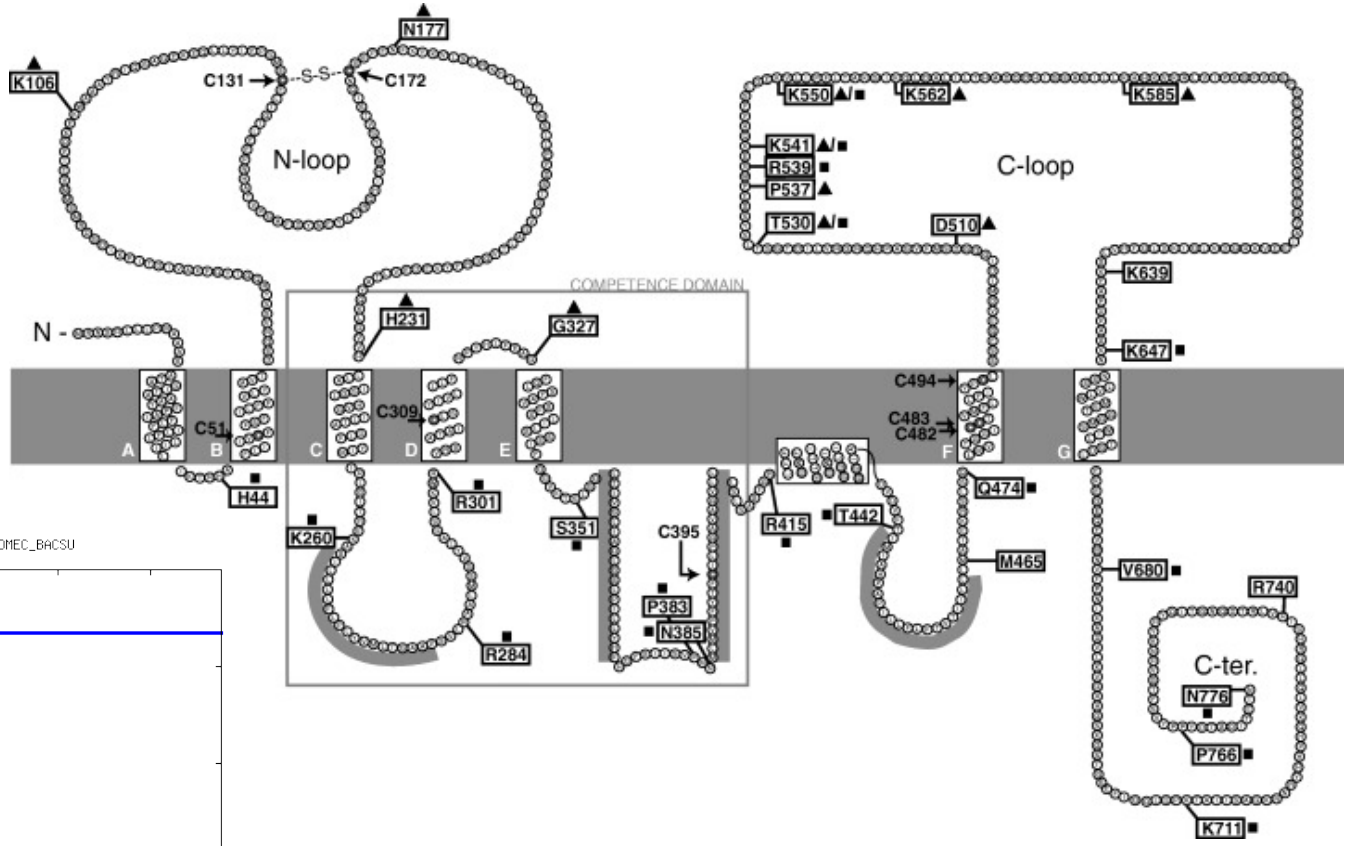
ComEC function



Averhoff (2009) FEMS Microbiol. Rev. 33, 611

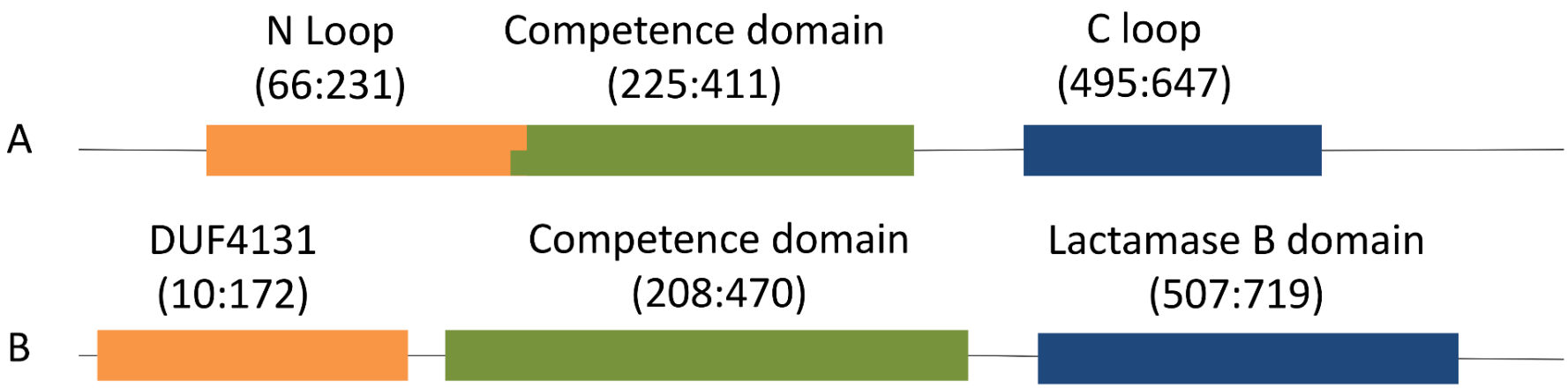
ComEC structure





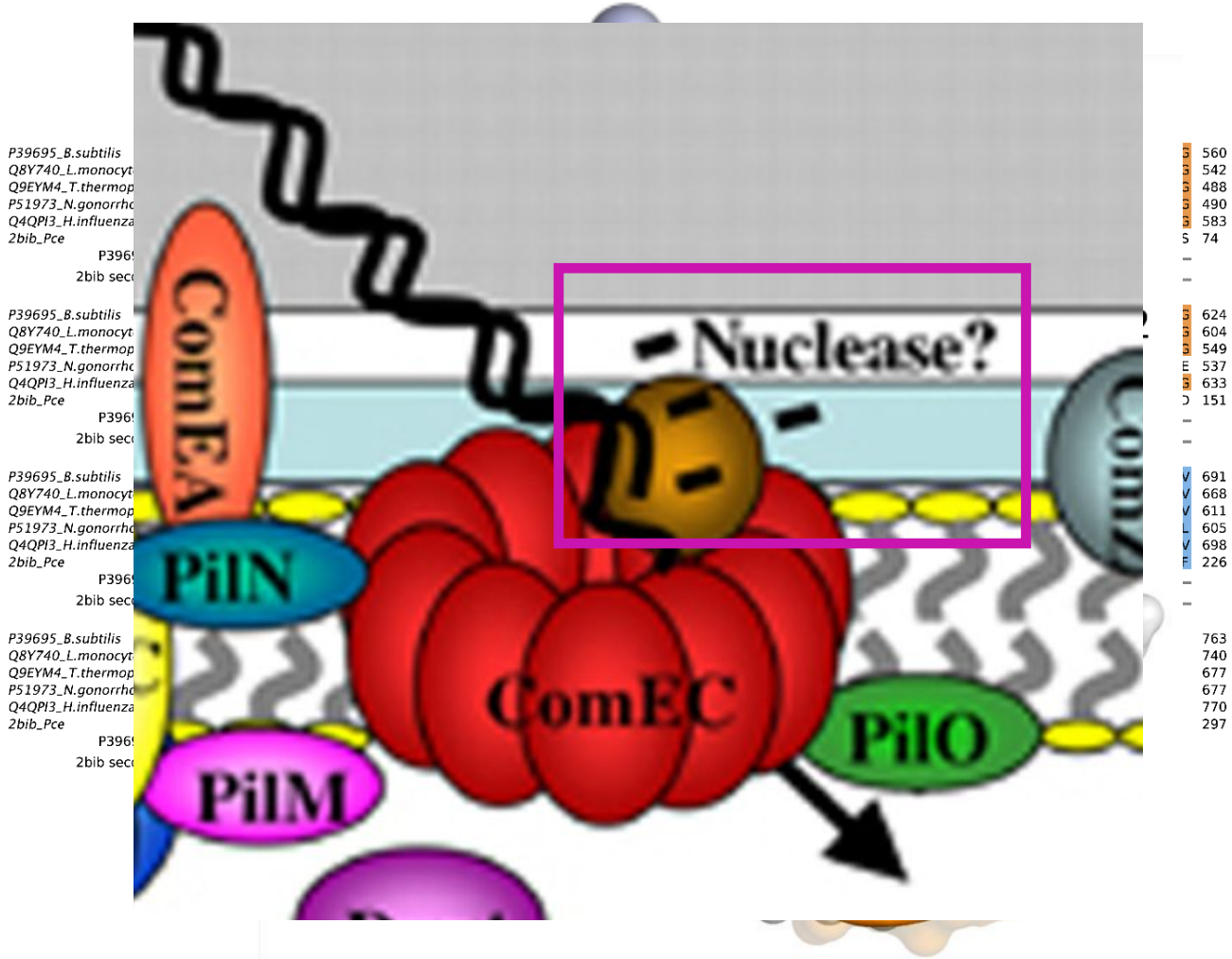
Draskovic & Dubnau (2005) Mol. Microbiol. 55, 881

ComEC domain structure



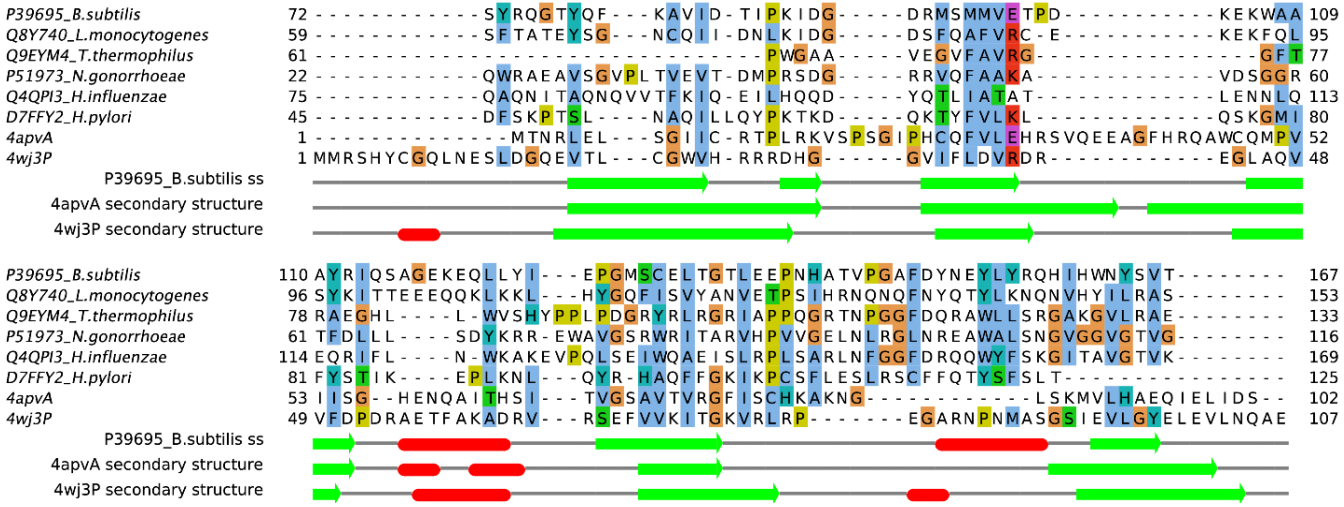
ComEC lactamase-like domain is predicted to be a DNase

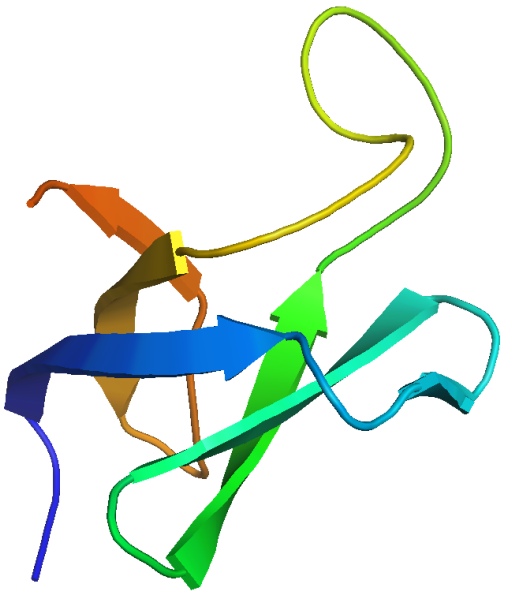
- Positively charged
- Predicted as DNA binding by structure-based predictors
- Accommodates DNA duplex neatly
- We think the mystery nuclease activity is encoded within ComEC!



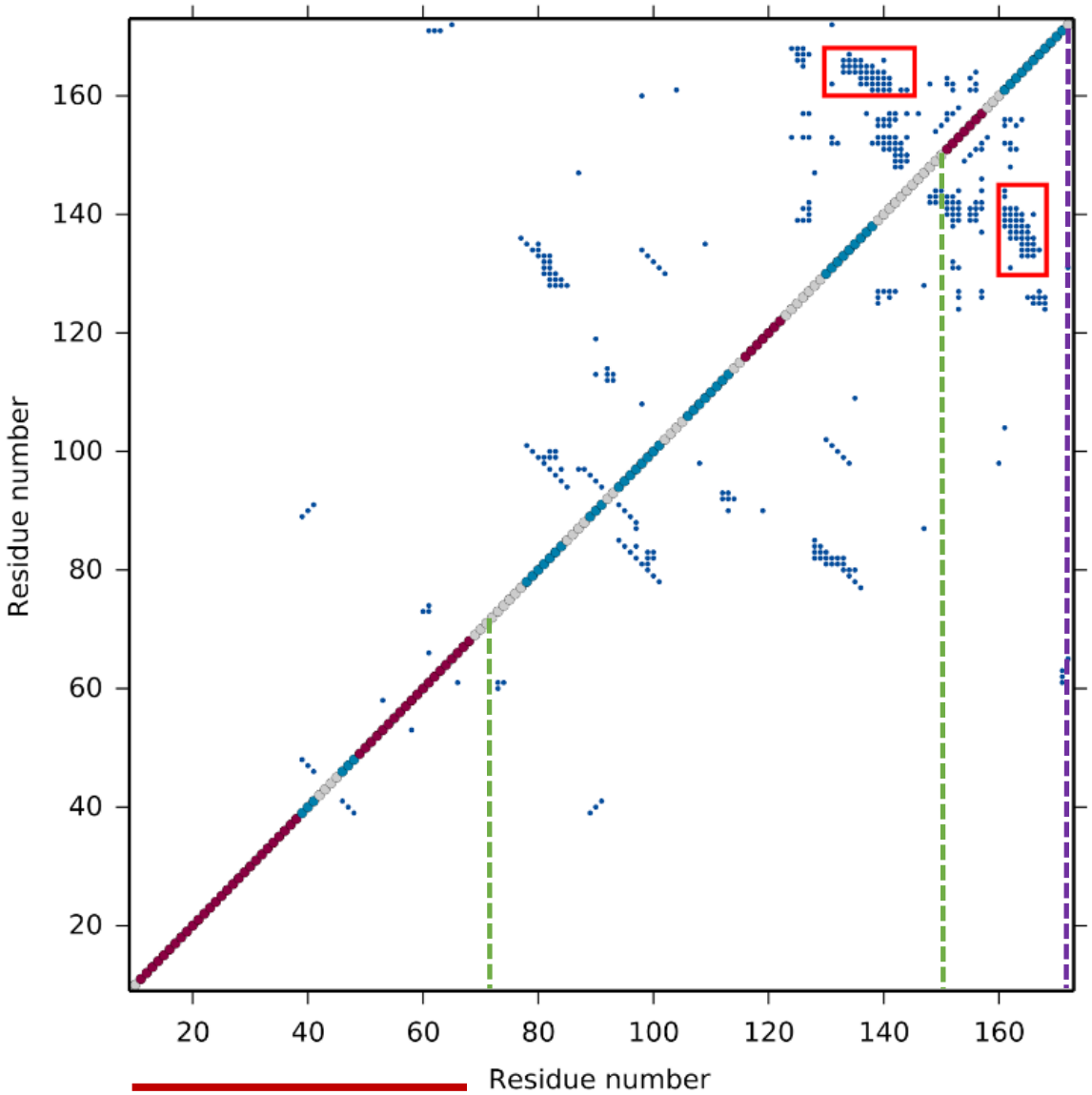
DUF4131 is predicted to be an OB fold

- Pretty clear result by HHpred distant homology detection
- OB folds bind single-stranded nucleic acids or oligosaccharides
- Context suggests former, though structure-based methods do not predict NA binding (trained on **ds**DNA?)
- But model looks small and not beautifully formed





Do we have the right domain boundary?



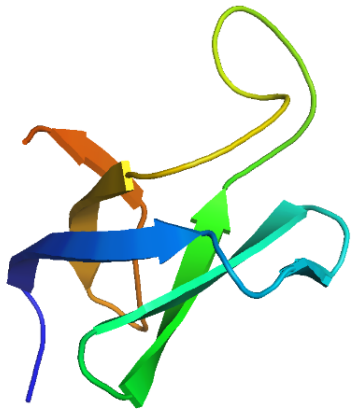
Two TM helices

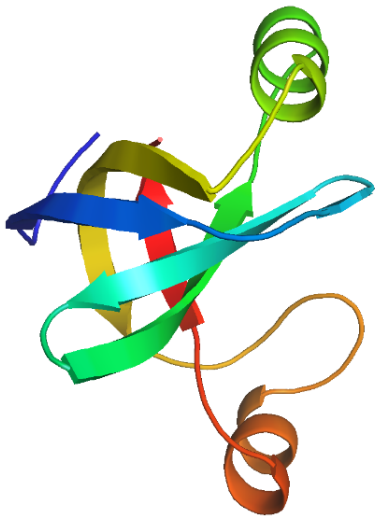
Domain limits from HHpred

Revised domain boundary

A contact-assisted fragment assembly model from Rosetta looks much better

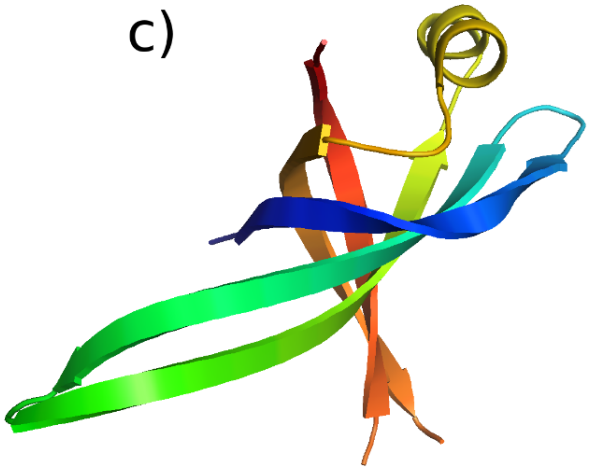
Previous homology model

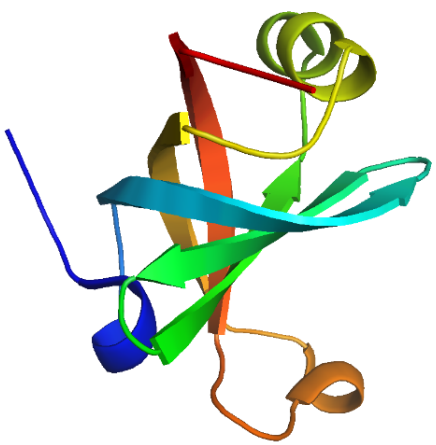




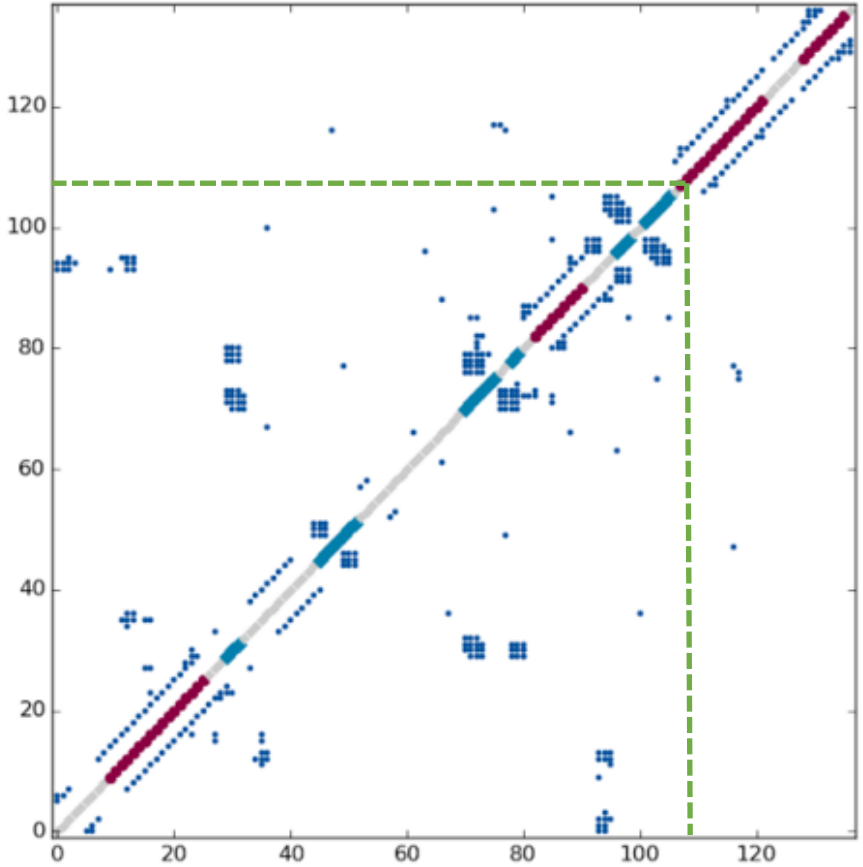
New model

Two crystal structures

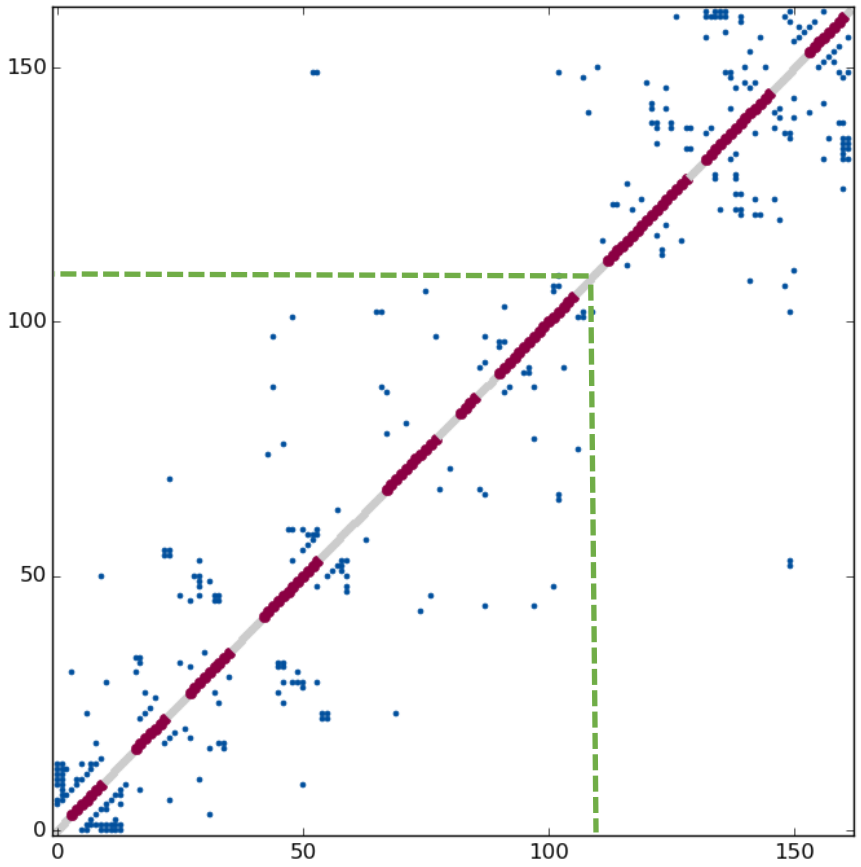




Many large DUFs seem to contain multiple domains



DUF3670 – separate helices at C-terminus?



DUF4158 – two helical domains?

Limitations and opportunities

- Generally need large number (>~1000) of reasonably diverse sequences
- Servers for **single** sequence are available
  - Evcouplings (a day)
  - RaptorX (two days)
  - Gremlin (1-2 hours)
- Only one server for **two** sequences and doesn't work well
- Only one server for **folding** from a sequence, but it's slow and is not the top method
- We have tools for single sequences and folding installed locally (Felix). No local method for two sequences available

Limitations and opportunities

- Get better domain definitions for cloning, bioinformatics
- Predict folds *ab initio* for large families (AMPLE)
- Rank interfaces in crystal structures and docking results

- ? Predict functional sites
- ? Filter true from false positives in Y2H, affinity tagged complexes
  - ??Genome-scale *in silico* interactomes
- ? Supplement incomplete NMR data