Bioinformatics for crystallographers

Dan Rigden

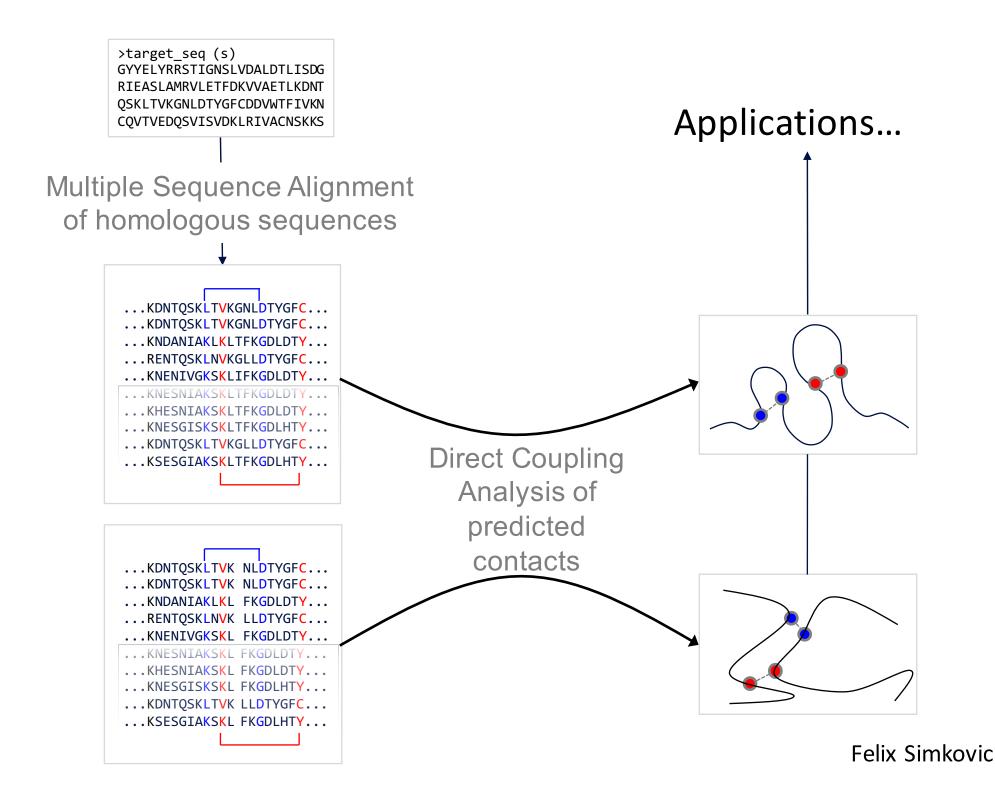


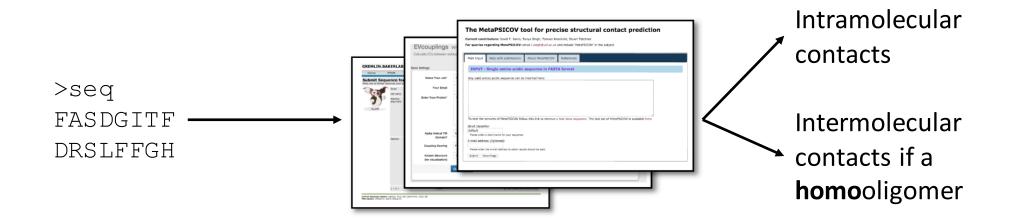
Preamble

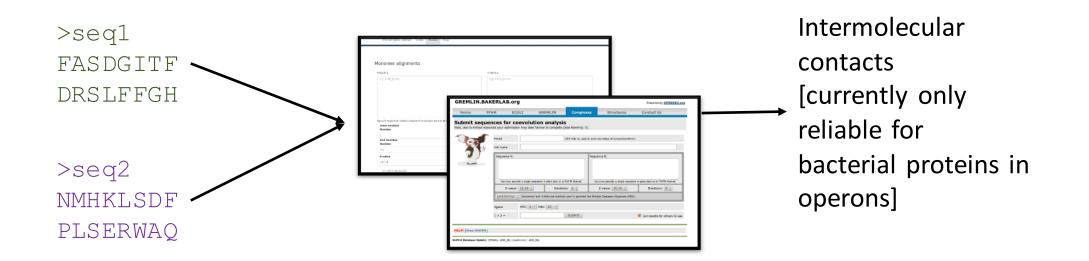
- Can't cover all bioinformatics!
- Prof Garratt will cover structure-based function inference
- I will focus on bioinformatics (= prediction)
 - Using newer/less well-known data
 - Majoring on easily available servers/predictions
 - Related to predicting domain composition and interactions. Mainly relevant 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)







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

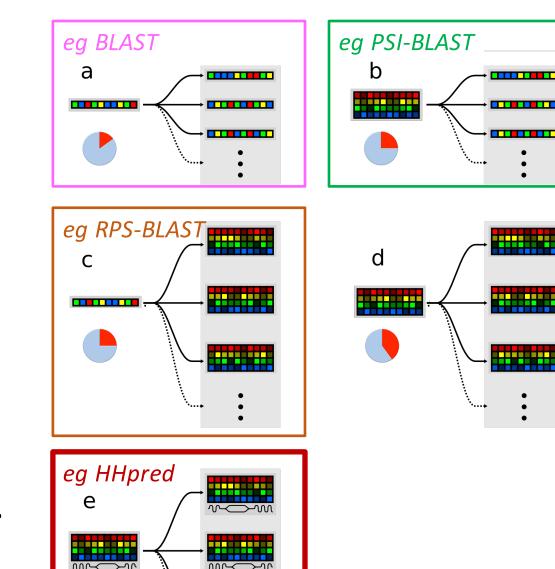
Recognising domains by homology with PDB, SCOP, CATH

• (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



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

HET: APJ; 2.104 {	Yers	turonase; GH28, pectin, cell WALL, hydrolase, periplasm, beta-helix, glycosidase, inia enterocolitica} e=7.5e-40 Score=280.73 Aligned_columns=440 Identiti s=16%	EX0-	activity;
Q ss_pred Q ss_conf Q gi 60495103 em Q Consensus T Consensus T 2uvf_A T ss_dssp T ss_pred T ss_conf	15 146	HHHHCCCCCCCEEEEEECCCCCCCCCHHHHHHHHHHHH	94 217	(605) (605) (608) (608)
Q ss_pred Q ss_conf Q gi 60495103 em Q Consensus T Consensus T 2uvf_A T ss_dssp T ss_pred T ss_conf	95 218	CCCCEECCCCCCEECCCCCEEEECCCCCCEEEECCCEECCCCC	169 265	(605) (605) (608) (608)
Q ss_pred Q ss_conf Q gi 60495103 em Q Consensus T Consensus T 2uvf_A T ss_dssp T ss_pred T ss_conf	170 266	CCCCCCEEEECCCCEEECCCCCEEECCCCCCCCCCCC	249 332	(605) (605) (608) (608)
Q ss_pred Q ss_conf Q gi 60495103 em Q Consensus T Consensus T 2uvf_A T ss_dssp T ss_pred T ss_conf	250 333	EEEEECCCCEEEEECCCCEEEEEEEEEEEEEEEEEECCCC	329 404	(605) (605) (608) (608)

HHpred tips and warnings

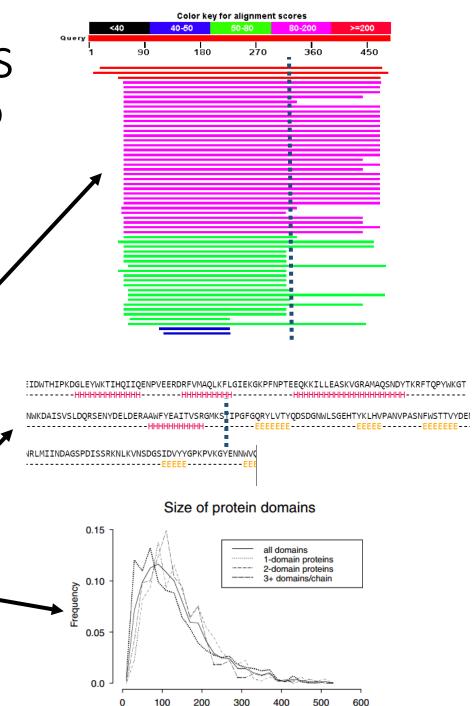
- Probability (0-100) is generally a good guide....
- ...but statistics can mislead for unusual protein sequences eg coiled-coil, low-complexity, Cys-rich
- Consider if the match makes biological sense!
- Look at the matched region
 - Is it a complete domain/structure?
 - If partial, could it reasonably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

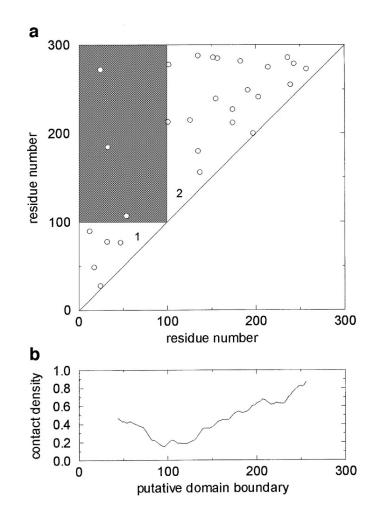
- *ab initio* approaches to domain boundary identification
 - BLAST matches in sequence / databases. Domains are often found in different combinations
 - Secondary structure pattern. PSI-PRED or Jpred4 (faster server)
 - Domain guess by size (server defunct)
 - Contact predictions

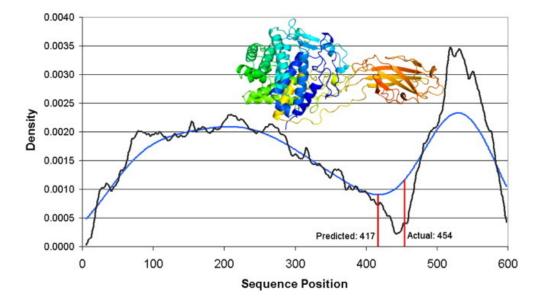


Domain lengt

Predicted contacts for defining domains

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

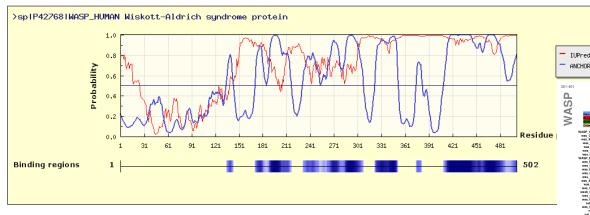


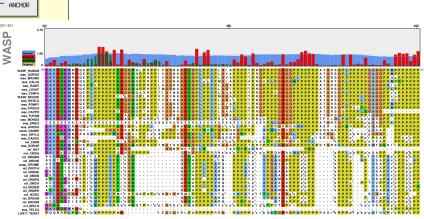


With today's contact predictions, it is now about the best method and works on multiple domains too

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)

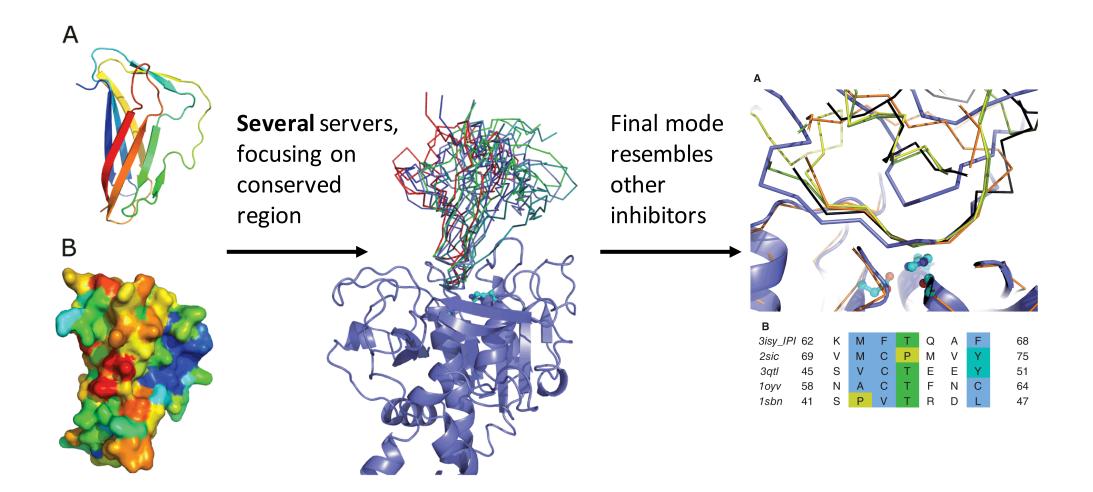




Predicting protein-protein interactions

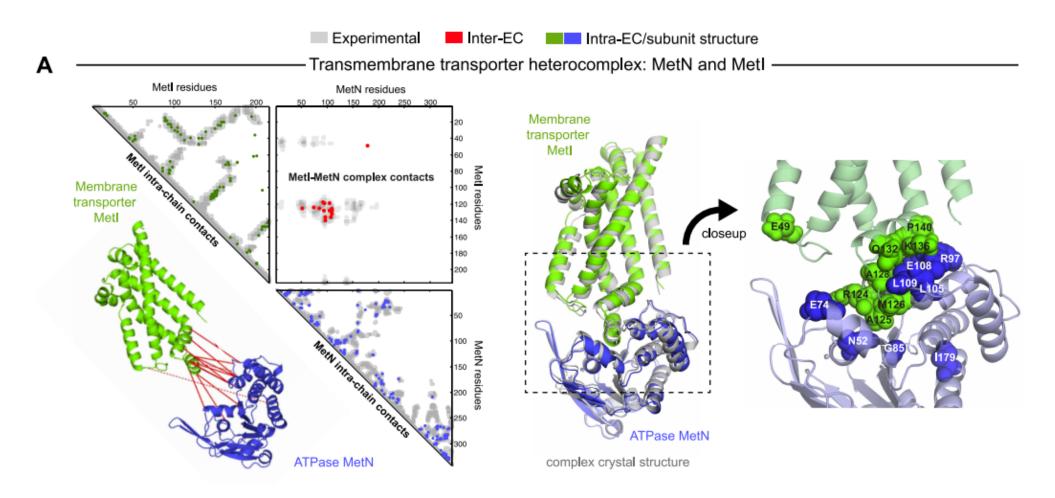
- Relevant to MR eg proteins A and B are cocrystallised but neither alone solves. An accurately predicted complex, being larger, might solve
- Many methods predict complexes based on steric complementarity plus other scoring functions
- Recommendable servers include
 - ClusPro, the best performing docking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



Rigden et al. (2013) F1000Research 2:154

Using predicted contacts to help predict complexes

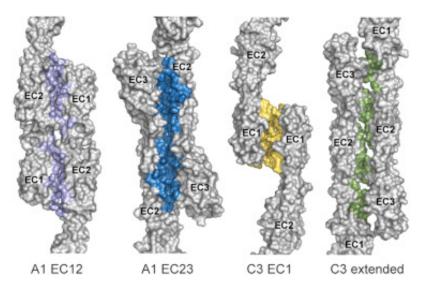


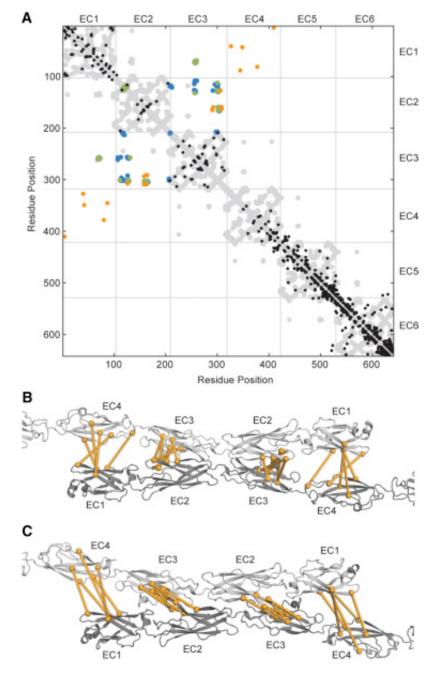
... and once you have your crystal structure...

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





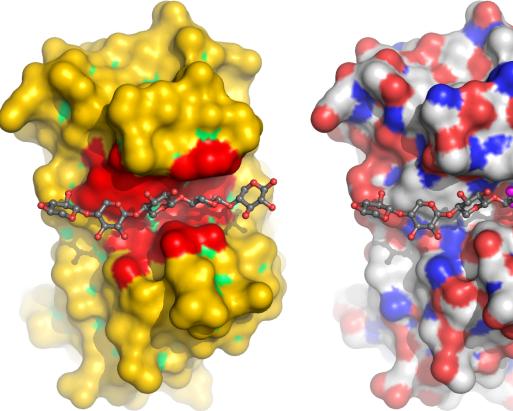
Nicoludis et al. (2015) Structure 23, 2087

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



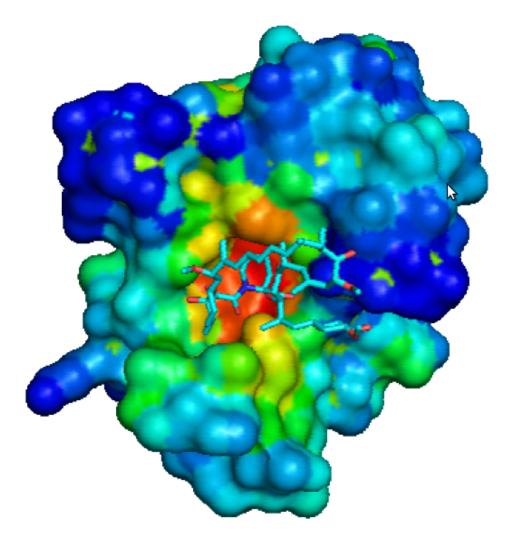
ISMBLab

SiteHound

scbx.mssm.edu/sitehound

Binding sites from statistics

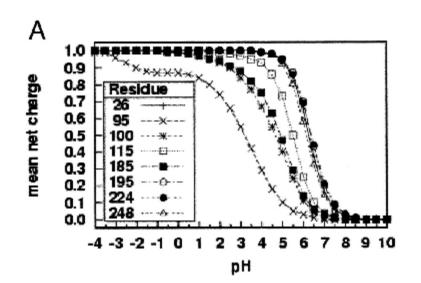
- STP (surface triplet propensities)
- 13 atom types → 455 triplets
- Distribution in binding vs non-binding sites varies
- Designed for small molecules, works on PPIs, including flat surfaces



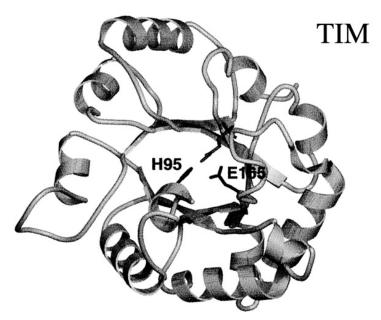
opus.bch.ed.ac.uk/stp

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

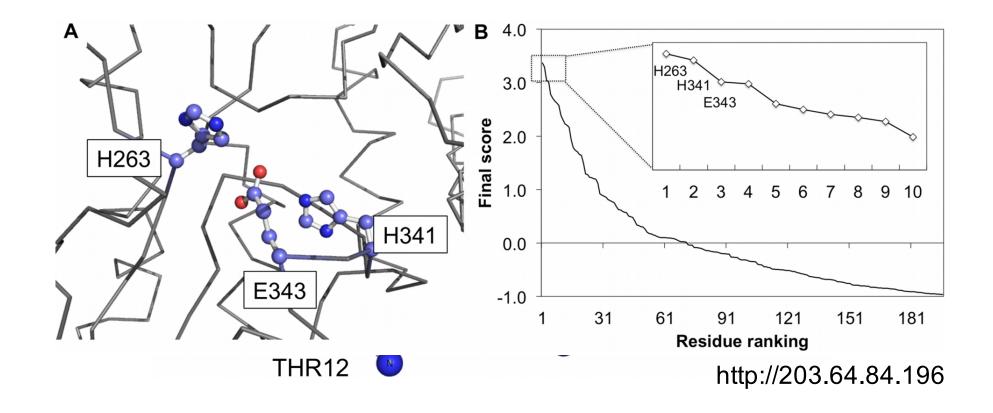


His95 and other residues with atypical pKa cluster at catalytic site

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

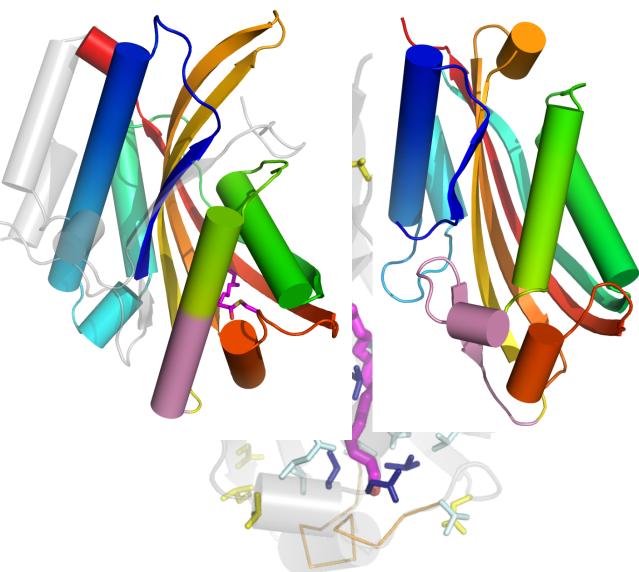
EXIA2: side chain orientation and rigidity

- 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.



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

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

is function in both

Diptera and

Lepidoptera

single protein family

shows only 'pink' site

...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...

150 170 180 190 200 210 230 ij|548534|sp|P00950|PMG1 YEAST/1-487 - FAARACELLKEKI iil30244201sp1P961211GPMA_TREP/1 ji|12751461|gb|AAK07665.1|/1-487 ji|1169587|sp|P32604|F26_YEAST/1-487TS<mark>P</mark>DYFN ji|1730554|sp|P52086|C0BC_EC0L ji|3183165|sp|P76502|SIXA_ECOL/1 ji|2895490|gb|AAC38954.1|/1-487 VDYA-NNOLTOOCOOOAAAAATKLEAMAAAKEFI ji|15229917|ref|NP_187168.1|/1-487 KLLPKRT -AAYTTTPDH-KIOLTDSGLLOAOEAGARLHALTSSNPSSPEWRVYF) ji|16130187|ref|NP_416755.1|/1-487

		220	230	240	250	260	270	280	290	300
but MUSCLE	gi 2895490 gb AAC38954.1 /1-537	LEL <mark>PV</mark> D	IC <mark>YTRH</mark> GKT()GNTE <mark>PRVFQ</mark> C	QVDYANNQ <mark>LT</mark> Q	Q <mark>Q</mark> QQAAAA	ATK <mark>LEA</mark> MAAAI	KEFI <mark>PD</mark> -LL	L <mark>SSP</mark> L_RA	VH <mark>T</mark> AQ <mark>P</mark> FV
	ji 1730554 sp P52086 COBC_ECOLI/1-537	MR	LWLERHGET)A <mark>NID-GLYS</mark> C)HA <mark>P</mark> T <mark>PLT</mark> A	RGIE <mark>Q</mark> AQNL	HTLLH <mark>G</mark> VSFD-	L <mark>V</mark>	LC <mark>S</mark> EL <mark>ER</mark> A	QHTAR-LV
gets it	ji 12751461 gb AAK07665.1 /1-537		<mark>Trhget</mark> k	(WNVE-RRMQ <mark>C</mark>)WQDS <mark>PLT</mark> E	<mark>KG</mark> R <mark>Q</mark> DAMRL	. <mark>G</mark> KRLEAVELA	AI	<mark>ytst</mark> sg r a	L <mark>ET</mark> AE-IV
-	ji 548534 sp P00950 PMG1_YEAST/1-537	M <mark>P</mark> K	lvuv <mark>rhgos</mark> e	WNEK-NLFT)WVDVK <mark>LS</mark> A	KGQ <mark>Q</mark> EAARA	GELLKEKKVY-	<mark>PD</mark> -VL	<mark>yts</mark> k_s <mark>r</mark> a	I TAN-IA
	gi 3024420 sp P96121 GPMA_TREPA/1-537	<mark>M</mark> K	LVU <mark>IRHGES</mark> E	WNRL-NLFT	WTDV <mark>PLT</mark> P	RGES <mark>EA</mark> QEG	I <mark>GRVL</mark> QEA <mark>GF</mark> D	<mark>FD</mark> -LC	<mark>YTS</mark> F_K <mark>R</mark> A	IRTLN-FV
	gi 16130187 ref NP_416755.1 /1-537	<mark>P</mark> V	VVL <mark>FRH</mark> AER(:DRSTNQ <mark>CL</mark> SD)KT <mark>GIT</mark> V	KGT <mark>Q</mark> DAREL	CNAFSADI	<mark>PD</mark> FDL	<mark>YSSN</mark> TV <mark>R</mark> T	I <mark>(</mark> SA
	ji 3183165 sp P76502 SIXA_ECOLI/1-537	<mark>M</mark> Q	VFIM <mark>RH</mark> GDAA	ALDAASD <mark>SV</mark> R-	· <mark>PLT</mark> T	N <mark>G</mark> CD <mark>ES</mark> RLM	IANWLK <mark>G</mark> QKVE-	<mark>I</mark> E-RV	LV <mark>SP</mark> FL <mark>R</mark> A	EDTLE
	ji 15229917 ref NP_187168.1 /1-537	KLL <mark>P</mark> KR	IIL V <mark>RH</mark> GESE	GNLD TAAY <mark>T</mark> T	-T <mark>P</mark> DHKIQLTD	S <mark>G</mark> LL <mark>Q</mark> AQEA	GARLHALISSI	v <mark>p</mark> ss <mark>p</mark> ewrv	/F <mark>YVSP\DRT</mark>	FSTLREIG
	ji 1169587 sp P32604 F26_YEAST/1-537	YVMNIR <mark>P</mark> KPKY	IWLSRHGESI	YNVE-KKIGO	DSSLSE	RGFOYAKKL	EOLVKESA <mark>C</mark> E	INLTV	WTSTLART	OTAN-YL
	5.1									

Jalview.org, recommended for sequence alignments

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

MAFFT Multiple Sequence Alignment of Retrieved fro e Edit Select View Annotations Format Colour Calcu			🕌 Average distan 🗖 🔲 🔀 File View
Piosphorylation Site Predictions MAFFT Alignment 60 70 80 R_CAPAA/1-97 - ASYKVKL I TPDOR I EFDC PDNVYI L R_CAPAA/1-97 - ASYKVKL I TPDOR I EFDC PDNVYI L R_CAPAA/1-97 - ASYKVKL I TPDOR I EFDC PDNVYI L R_CAPAN/1-144 MASYKVKL I TPDOR I EFDC PDOVYI L R/SXI9_SOLC/1-144 MASYKVKL I TPDOR I EFDC PDOVYI L R1_SKU2 MASYKVKL I TPDOR I EFDC PDOVYI L R1_SKU2 MASYKVKL I TPDOR I EFDC PDOVYI L R1_PEA/1-149 MASYKVKL I TPDOR I EFDC PDOVYI L R1_FPA/1-52 MATYKVKL I TPDOR I EFDC PDOVYI L R3_RAPSA/1-96 - - - R1_ARATH/1-148 MATYKVK FI TPEGE DEVECTODOVYVL R2_ARATH/1-148 MATYKVK FI TPEGE DEVECTODOVYVL R2_ARATH/1-148 MATYKVK FI TPEGE DEVECTODOVYVL R2_MARTH/1-148 MATYKVK FI TPEGE DEVECTODOVYVL R2_MARTH/1-148 MATYKVK FI TPEGE DEVECTODOVYVL R2_MARTH/1-148 MATYKVK FI TPEGE DEVECTODOVYVL Secondary Structure	Order Alignment > GA Secondary Structure Prediction > > GA Protein Disorder > GA Analysis > GA Conservation > HA Conservation > AA Fetch DB References > AA Fetch CB Cass Cass Cask vs Ses V Second Vs Ses V AA Fetch CB Cass Cass Cask vs Ses V Fe Fetch Cass Cass Cass Cask vs Ses V Fe Fe Fetch Cass Cass Cass Cask vs Ses V Fe Fe Fetch Cass Cass Cass Cask vs Ses V Fe Fe F	http://www.compbio.dundee.ac.uk/jabaws Tcoffee with Defaults Edit settings and run Run Tcoffee with preset Probcons with Defaults Edit settings and run Muscle with Defaults Edit settings and run Muscle with Defaults Edit settings and run Align Mafft with Defaults Edit settings and run Run Muscle with preset Mafft with Defaults Edit settings and run Run Mafft with preset MSAprobs with Defaults Edit settings and run GLprobs with Defaults Edit settings and run Clustal	<pre>PERI_PEA OZAGE_TRIPE FERI_SOLLO FERI_SOLLO FERI_MESCR FERI_MARTH FERI_MARTH FERI_MARZE DI view for FERI_S With default settings</pre>
Connection Citrix Receiver to thymine.l		Realign with Clustal ClustalO Realign with ClustalO	> >

Jalview

• Also helps you produce figures like this...

lujc/1-156 1h2e/2-208 1k6m/250-470 1e58/1-247 1nt4/1-391 1dkq/1-410 1gwo/8-442	84GVAS VLVISHLPLVGYLVA 1 120GERFCDVQQRALEAVQSIVDRHGETVLIVTHGVVLKTLMA 1 364GESMEDLVQRLEPVIMELERQENVLVICHQAVMRCLLA 4 139 YAKLSEKELPLTESLALTIDRVIPYWNETILPRMKSGERVIIAAHGNSLRALVK 1 240 LSKLKNGMQDSLFTSP-EVARNVAKPLVSYIDKALVTDRTSGERVIIAAHGNSLRALVK 1 246 LLSLHNAQFYLLQRTP-EVARSRATPLLDLIKTALTPHPPQKQAYGVTLPTSVLFIAGHDTNLANLGG 3 269 YNYLOSLGKYYGMGAGNPLGPAQGIGFTNELIARLTRSPVODHTSTNSTLVSNPATFPLNATMYVDFSHDNSMVSIFF 3	60 01 92 98 12
1qw0/8-442 1ujc ss	α3 β4	10
lqwo ss	α3 β4	
1ujc/1–156	118 ELCPGE	
1h2e/2-208	161 AFK DTPLDHLWS PPYMYGTS VTIIEVDGGTFHVAVEGDVSHIEEVKEV2	
1k6m/250-470	402 YFLDKSSEELPYL-KCPLHTVLKLTPVAYGCKVESIYLNVEAVNTHREKPENVDITR 4	57
1e58/1-247	193 YLDNMS E E E I L E L - N I PTGV P L V YE F DE N F K P L - K R Y Y L GNADE I AAKAAAVANQGK - 2	47
1nt4/1-391	299 ALDFK PYQLHDQNERTPIGGK IV FQRWHDSK ANRDLMK I EYV YQSAEQLRNADALTLQAPAQR 3	61
1dka/1-410	313 ALELN-WTLPGOPDNTPPGGELVFERWRRLSDNSOWIQVSLVFQTLQOMRDKTPLSLNTPPGE 3	74
1qwo/8-442	347 ALGL YNGTEPLSRTSVESAKELDGYSASWVVPFGARAYFE TMQCKSEKEPLVRALIN	04
lujc ss	β5 β6	
1qwo ss	β5 β6	

• ... rather than like this

 MASR
 F
 G
 G
 R
 H
 L
 M
 N
 K
 R
 A
 K
 F
 L
 N
 V
 F
 S
 I
 F
 G
 A
 L
 R
 T
 K
 G
 L
 K
 F
 G
 A
 L
 R
 T
 G
 A
 T
 K
 C
 L
 K
 Q
 T
 L
 N
 T
 K
 G
 L
 K
 C
 L
 N
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 R
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N
 N

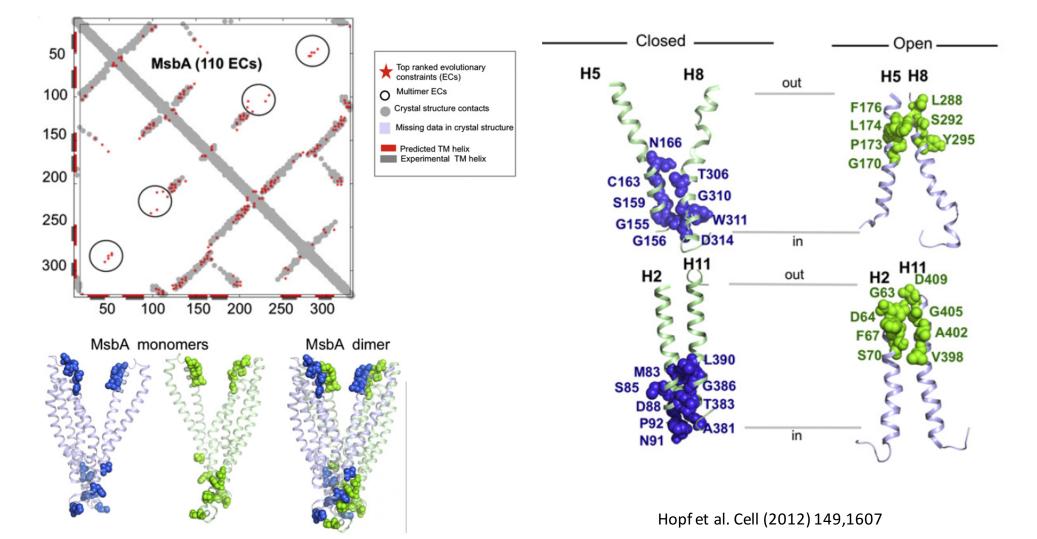
Questions? Feedback?



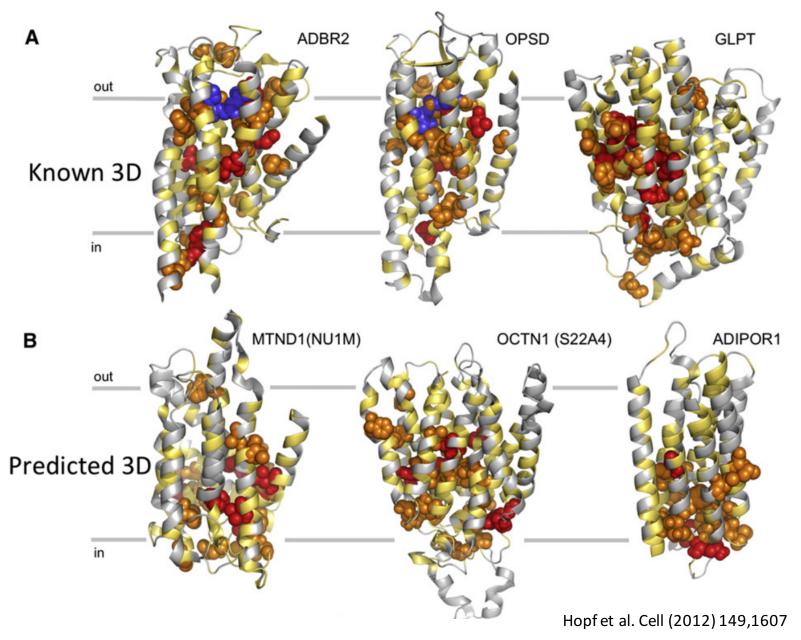
Contacts also inform on...

Oligomerisation

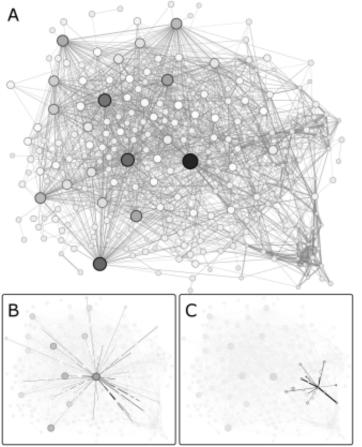
Conformational change



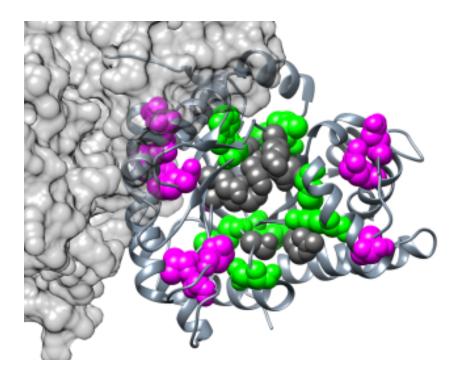
Predicting functional sites



Predicting functional sites



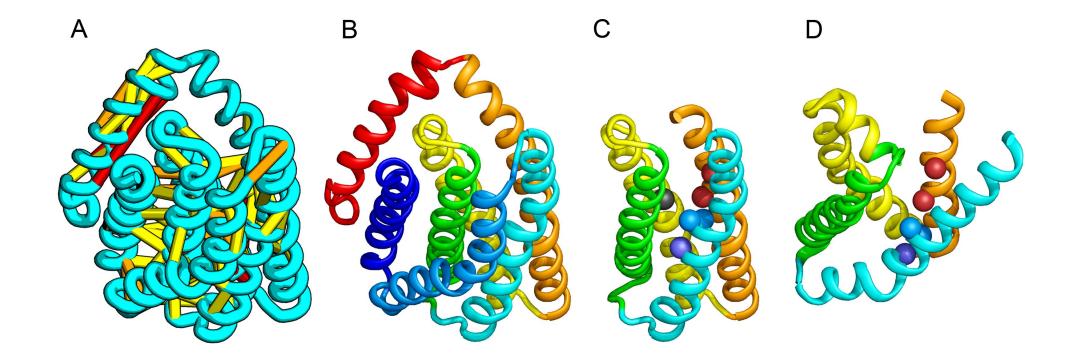
 Considering network structure of contact prediction map gives scores....



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

You don't know the structure

Predicted contacts for folding



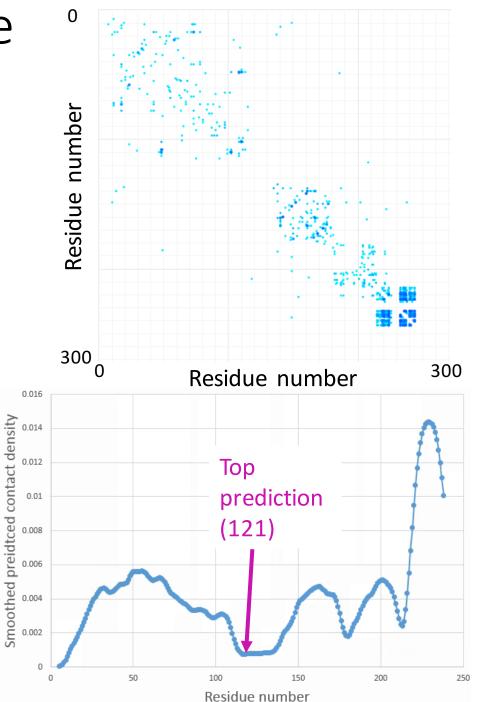
Model showing satisfied and unsatisfied contacts

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



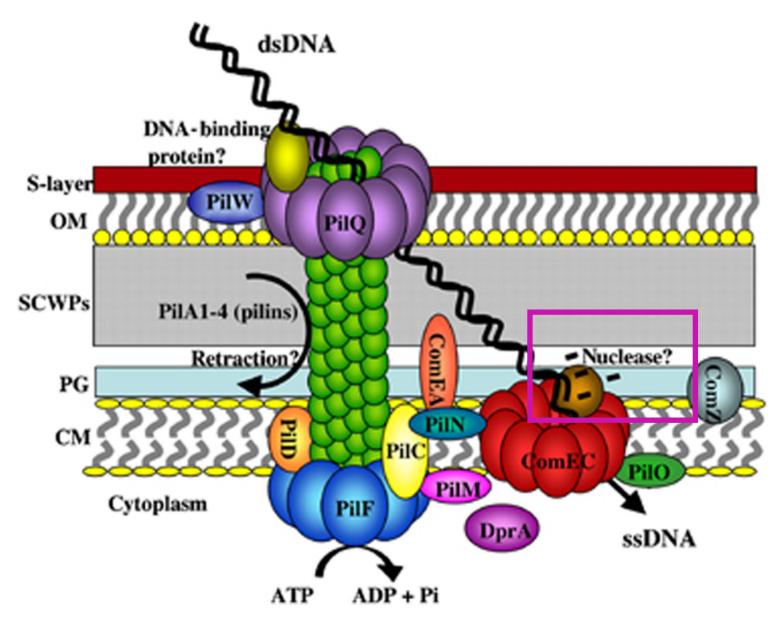


BBSRC Research Experience Placement

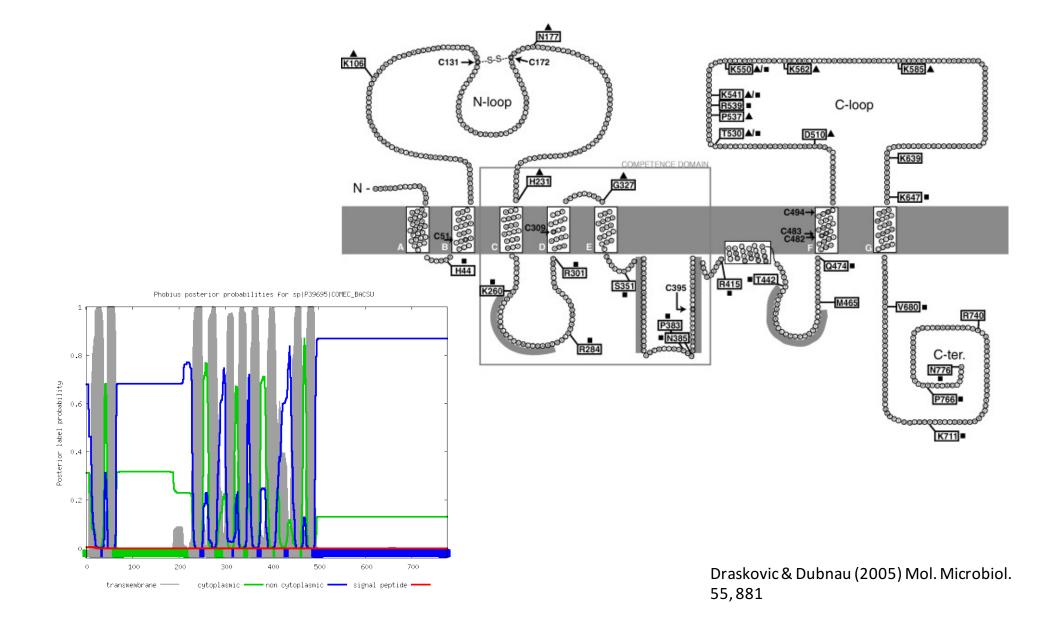
Bacterial competence

- 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

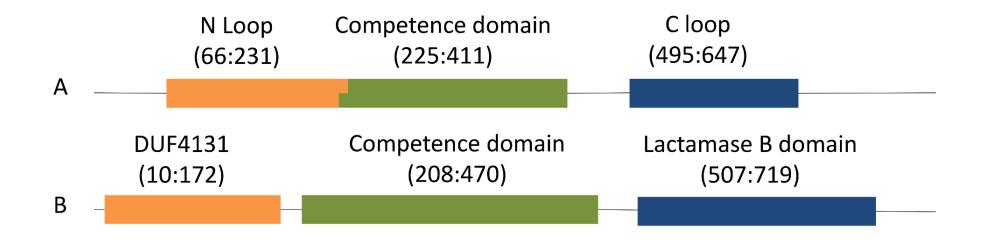
ComEC function



ComEC structure



ComEC domain structure



ComEC lactamase-like domain is predicted to be a DNase

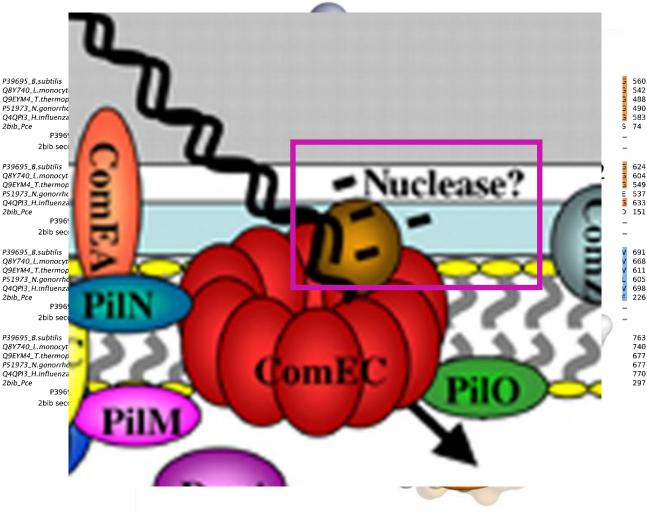
2bib Pce

2bib Pce

2bib_Pce

2bib_Pce

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



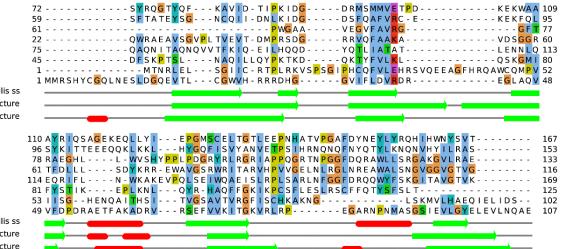
DUF4131 is predicted to be an OB fold

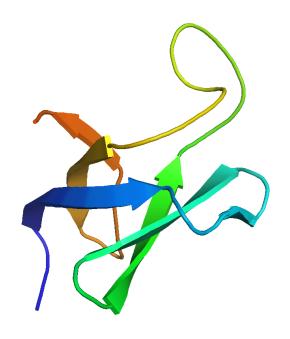
- Pretty clear result by HHpred distant homology detection
- OB folds bind singlestranded nucleic acids or oligosaccharides
- Context suggests former, though structure-based methods do not predict NA binding (trained on dsDNA?)
- But model looks small and not beautifully formed

P39695_B.subtilis Q8Y740_L.monocytogenes Q9EYM4_T.thermophilus P51973_N.gonorrhoeae Q4QP13_H.influenzae D7FFY2_H.pylori 4apvA 4wj3P

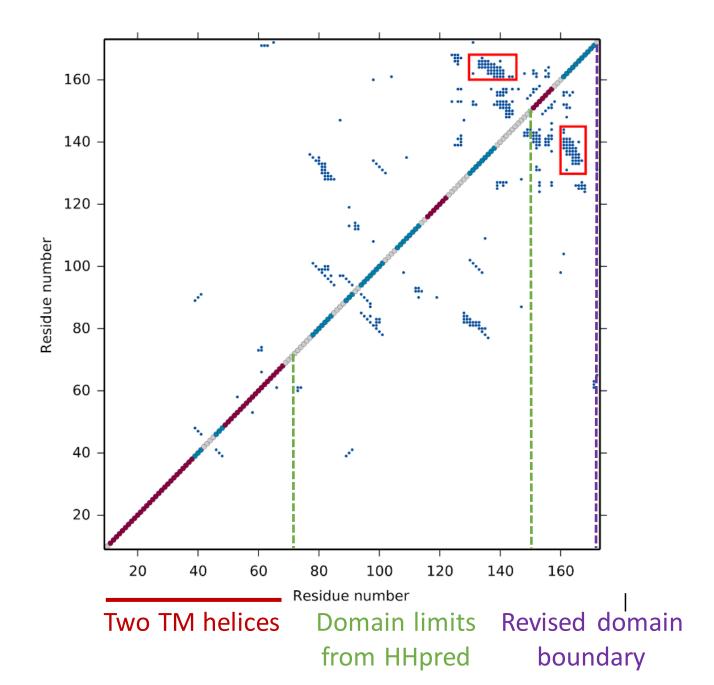
P39695_B.subtilis ss 4apvA secondary structure 4wj3P secondary structure

P39695_B.subtilis Q8Y740_L.monocytogenes Q9EYM4_T.thermophilus P51973_N.gonorrhoeae Q4QPI3_H.influenzae D7FFY2_H.pylori 4apvA 4wj3P P39695_B.subtilis ss 4apvA secondary structure 4wj3P secondary structure

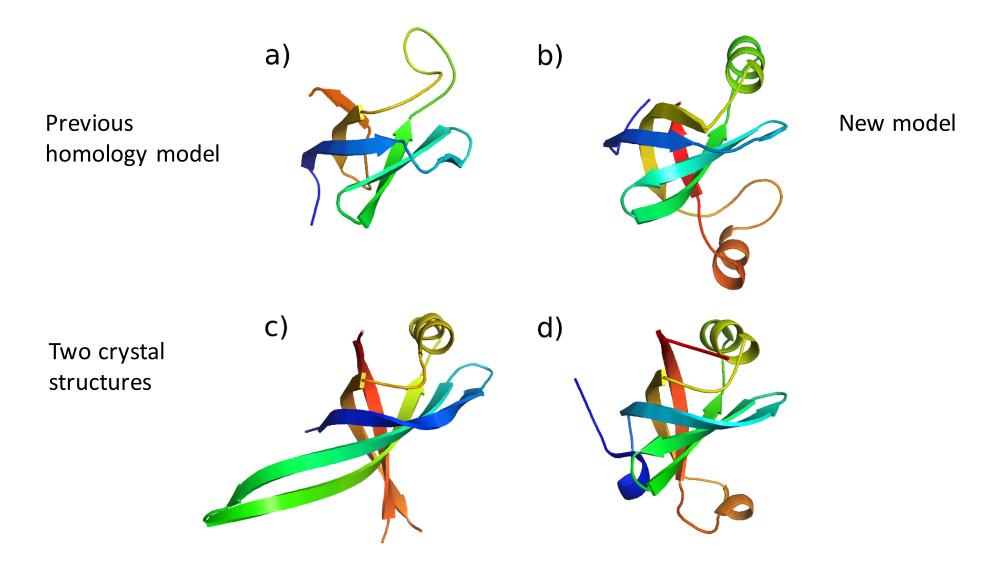




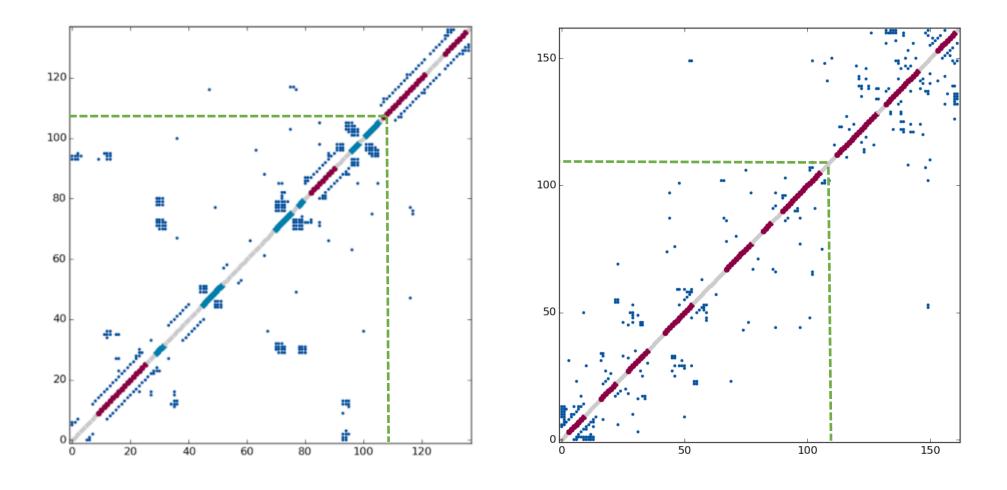
Do we have the right domain boundary?



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



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