KBDOCK – A Case-Based Reasoning Approach for Protein Docking

Dave Ritchie

Team Orpailleur
Inria Nancy – Grand Est

Outline

Basic Difficulties of Modeling PPIs by Docking

The Need to Classify Existing Interactions

The KBDOCK Case-Based Reasoning Approach

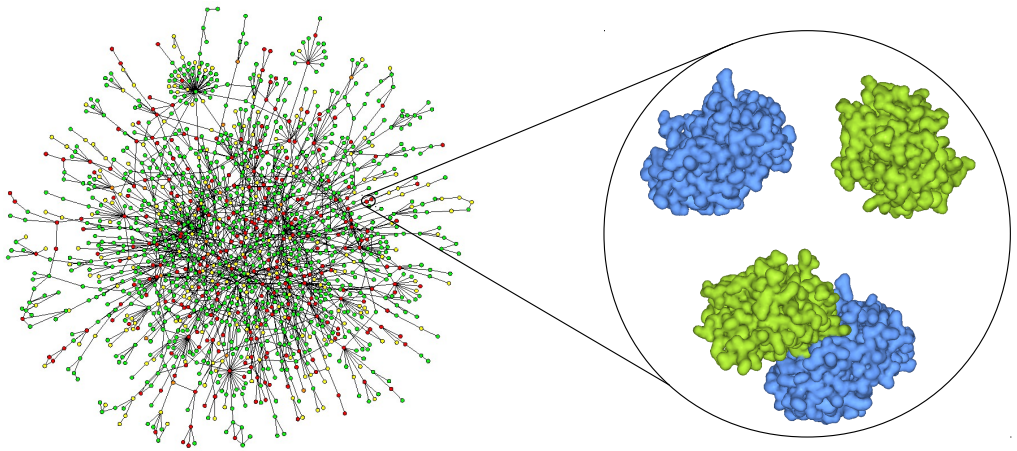
KBDOCK Performance on Selected CAPRI Targets

Demo: Using the KBDOCK Server to Explore DDIs

Practical: Modeling API-A/Trypsin and a TIM-barrel complex

The Protein Interactome

- There probably exist about 25,000 protein-protein interactions
- 3D crystal structures exist for only about 4% of these...
- Can we use existing structures to model unknown interactions?



A Case-Based Reasoning Approach

- PhD thesis project of Anisah Ghoorah (2009–2012)

Difficulties of Modelling 3D PPIs

Ab initio docking algorithms

- Produce thousands of candidate solutions
- Hard to identify acceptable solutions
- Additional challenge: to model protein flexibility

Template-based approaches

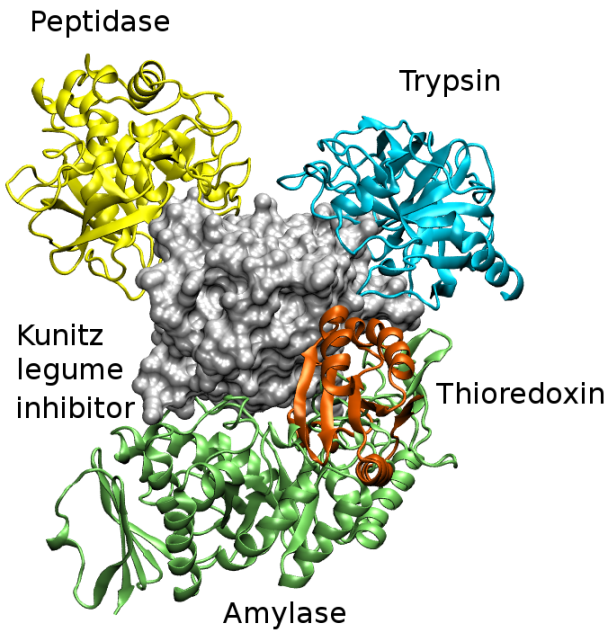
- Need a lot of effort to find suitable templates
- Require full-length templates to exist
- Fail when no templates are available

CAPRI Target 40 (2009) – API-A/Trypsin

- We searched SCOPPI and 3DID for similar 3D interactions
- This helped to identify two inhibitory loops on API-A

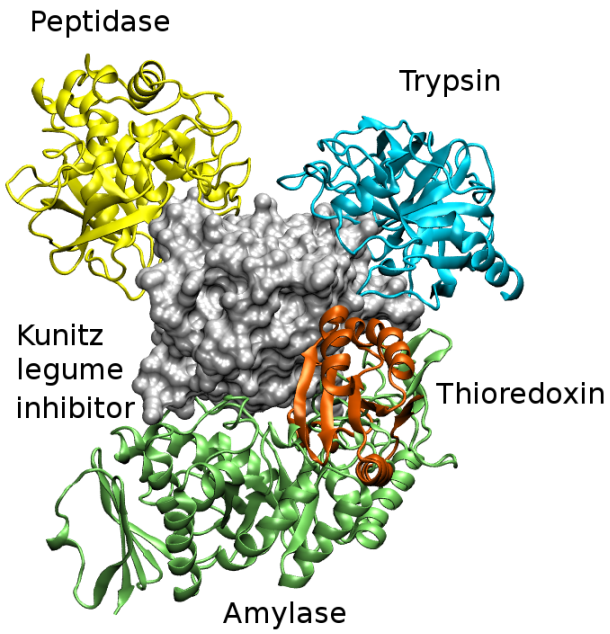
CAPRI Target 40 (2009) – API-A/Trypsin

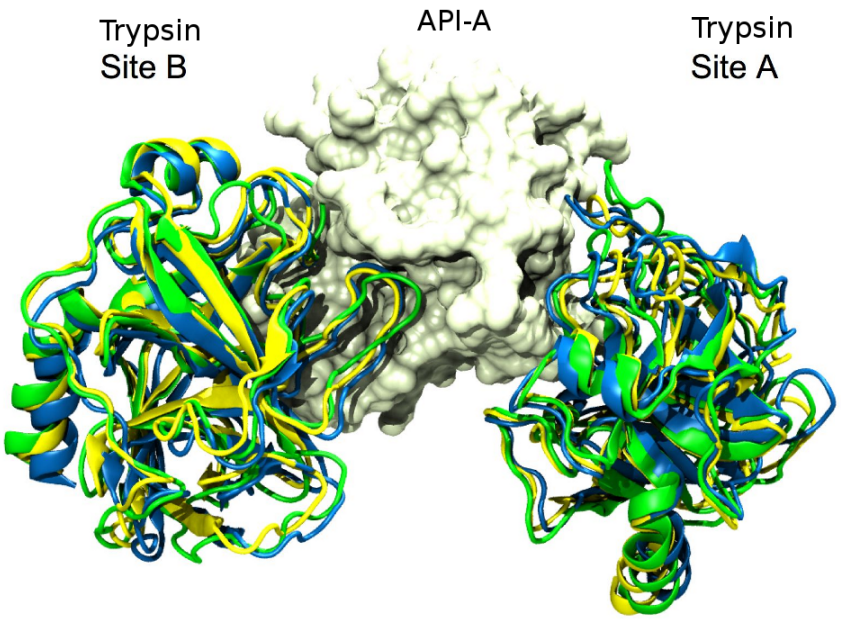
- We searched SCOPPI and 3DID for similar 3D interactions
- This helped to identify two inhibitory loops on API-A



CAPRI Target 40 (2009) – API-A/Trypsin

- We searched SCOPPI and 3DID for similar 3D interactions
- This helped to identify two inhibitory loops on API-A

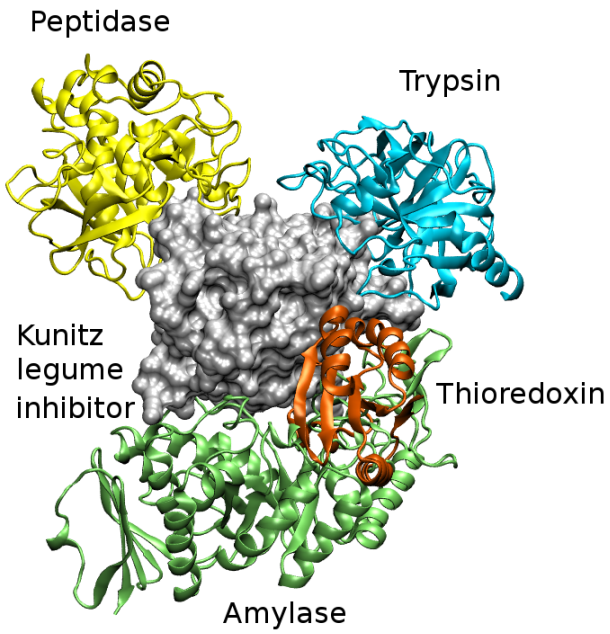


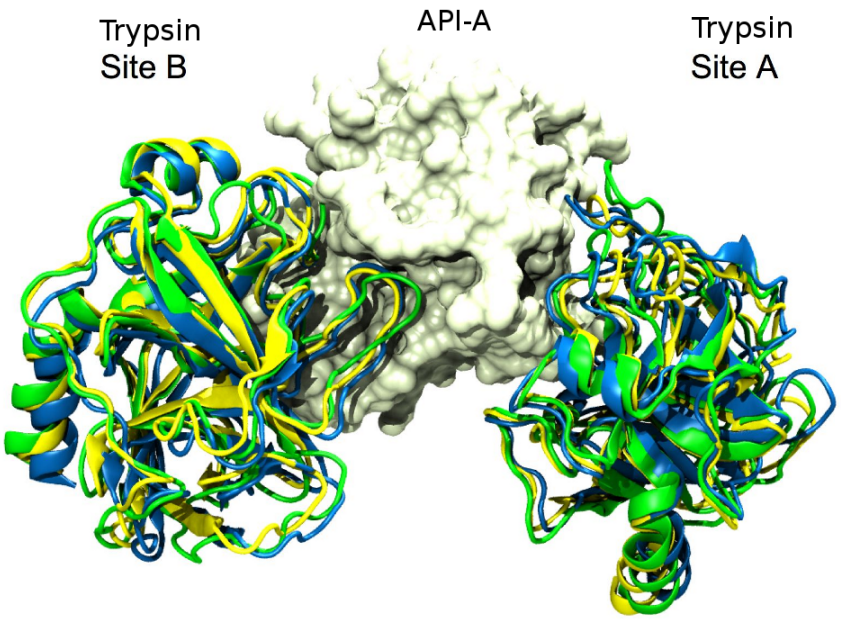


- Using Hex + MD refinement gave NINE "acceptable" solutions

CAPRI Target 40 (2009) – API-A/Trypsin

- We searched SCOPPI and 3DID for similar 3D interactions
- This helped to identify two inhibitory loops on API-A

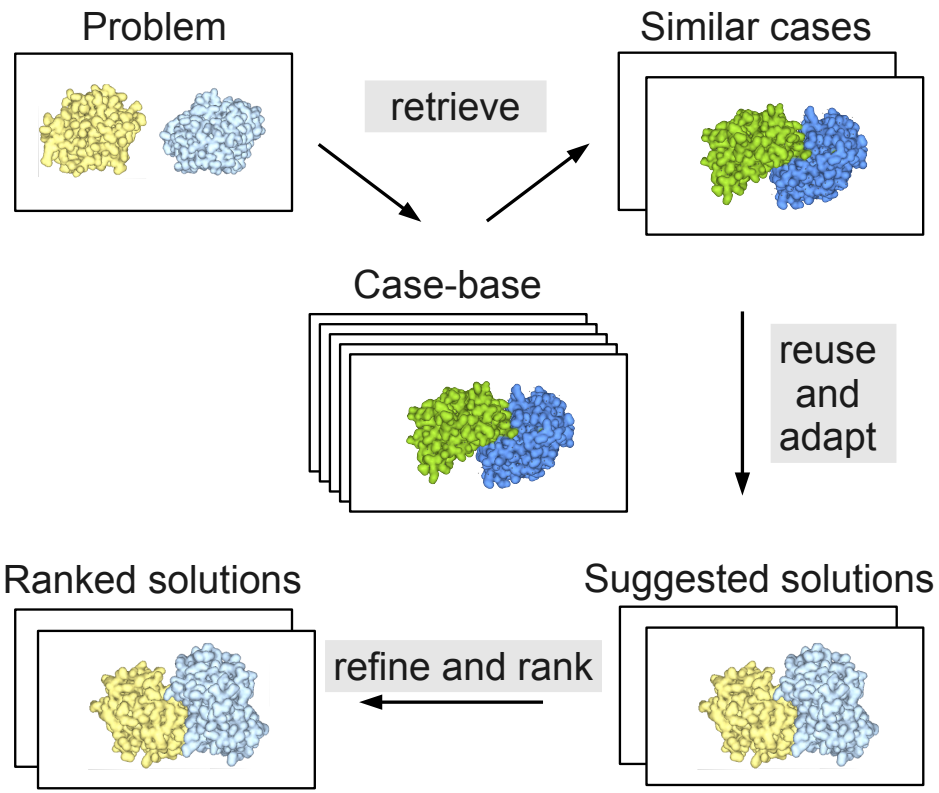




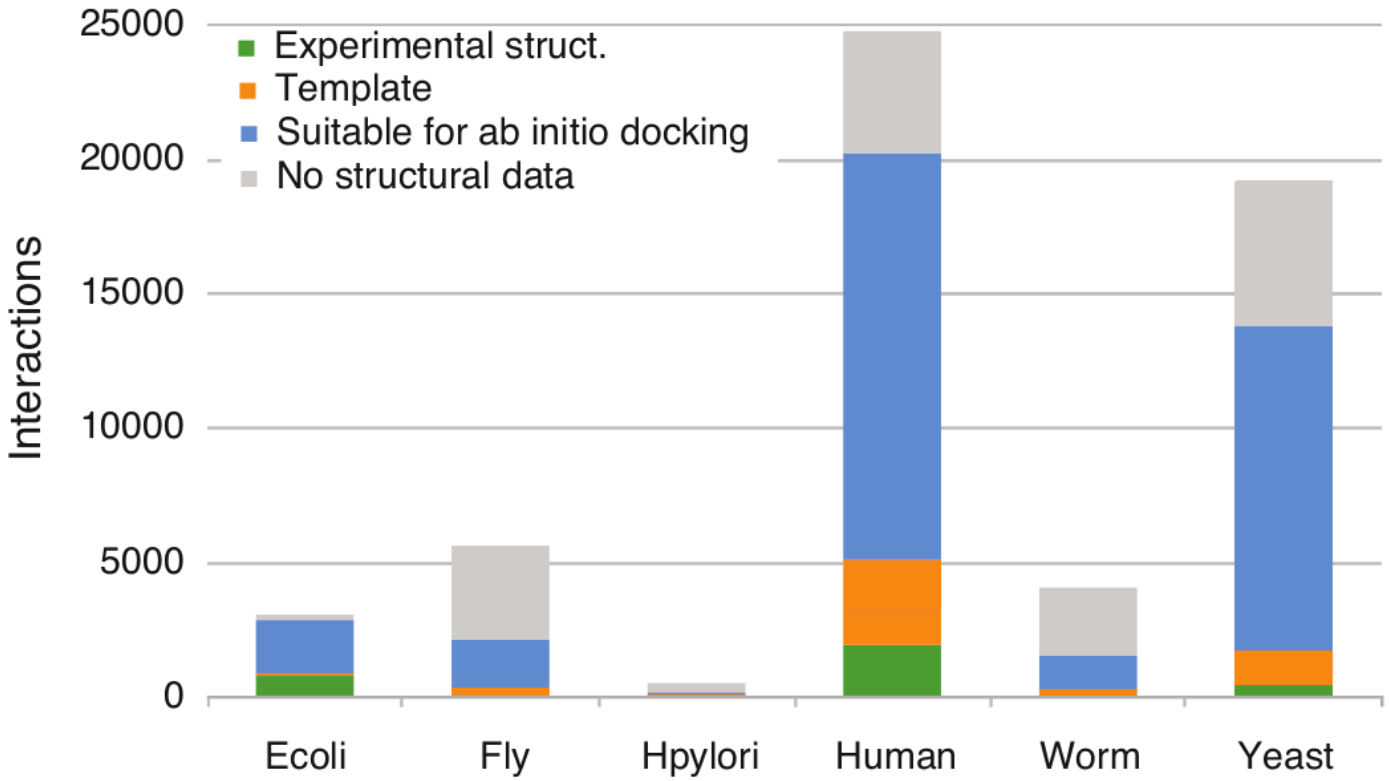
- Using Hex + MD refinement gave NINE "acceptable" solutions
- Anisah's mission: How to automate all this?

Modelling 3D Protein Complexes by Homology

- Case-based reasoning for 3D protein complexes



Current Structural Coverage of PPIs



- Only 8% of the known human PPIs have a 3D structure

Stein *et al.*, *Curr Opin Struct Biol*, 2011

Structural PPI Databases

- There are many ways of representing 3D interfaces
- No unique way to quantify whether two interfaces are similar

Classification	DDIs	Distinct Interfaces
Davis and Sali, 2005 (Pibase)	20,912	18,755
Kim *et al.*, 2006 (Scoppi)	10,080	5,727
Keskin *et al.*, 2004	21,686	3,799
Aung *et al.*, 2008 (PPiClust)	2,634	1,716
Shulman-Peleg *et al.*,2004	64	22

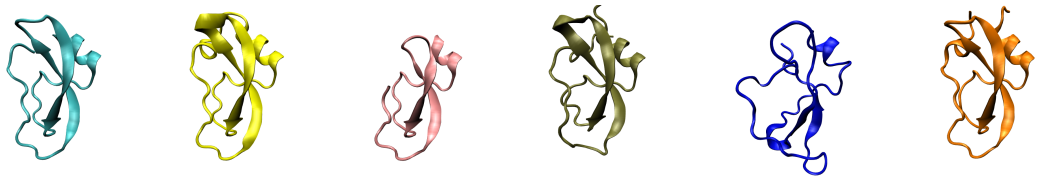
- Can we use such databases for knowledge-based docking?
- How many distinct interface types really exist?

The Need for a Structural Classification of DDIs

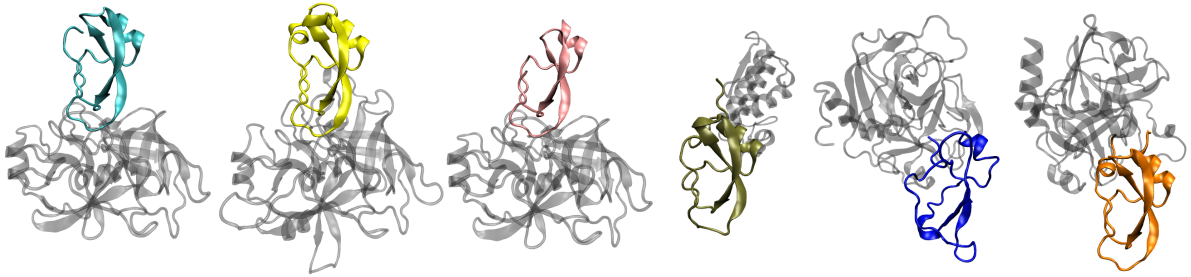
- Pfam classifies sequences into **domain families**

```
P00974 1brb -AG---EPP--YTG--P--CK---A--RI--IRY--FYN---AKAGLCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMRTA--
P00974 1bth ---FCLEPP--YTG--P--CK---A--RI--IRY--FYN---AKAGLCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMRTCG-
P00974 1co7 --------P--YTG--P--CK---A--RI--IRY--FYN-------LCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMR----
P00989 1bun --D-CDKPP--DTK--I--CQ---T--VV--RAF--YYK---PSAKRCV--QF--R-Y--G--G-C--N-G--N---G--NHFKSDHL-CRCECL-
P17726 1kig ---LCIKPR--DWI-DE--CD---S--NEG-GERA-YFR---NGKGGCD--SF--W-I-------C--P-E--DHTGA--DYYSSYRD-CFNACI-
Q8WPI2 2ody ---FCRLPA--DEG--I--CK---A--LI--PRF--YFN---TETGKCT--MF--S-Y--G--G-C--G-G--N---E--NNFETIEE-CQKACG-
```

- Families of similar sequences often have similar structures
- CATH and SCOP classify structures into **structural families**



- KBDOCK introduces **domain family binding sites** (DFBSs)



KBDOCK – Aims and Objectives

- Create a framework to support large scale analyses of protein binding site and interface features
- Use this framework to classify 3D interactions in a compact and re-usable way
- Use this classification as a systematic way to reuse and exploit structural knowledge of existing PPIs to facilitate 3D PPI modelling
- Provide a structural interaction search engine to facilitate 3D PPI modelling, in particular, docking by homology

KBDOCK Statistics

PDB

- Protein Data Bank – $\sim$ 85,000 protein structures (june 2013 snapshot)

Pfam

- Database of protein domain families
- Uses multiple sequence alignments to define domains
- Based on UniProt database
- Contains 14,831 domain families
- Of which, 6,516 have 3D structures in the PDB

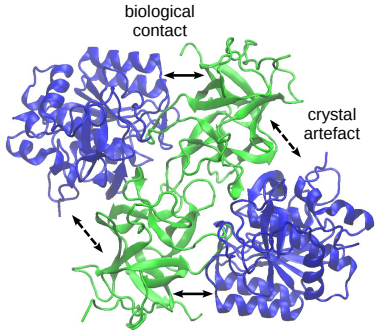
KBDOCK

- Uses Pfam to define domains
- Extracts all DDIs from PDB files
- Some statistics:
  - 231,405 PDB total chains
  - 288,309 total domains
  - 239,494 total DDIs
  - 12,498 inter-chain homo DFBSs
  - 4,001 inter-chain hetero DFBSs
  - 3,021 intra-chain hetero DFBSs
  - 745 intra-chain homo DFBSs
  - 1,213 domain-peptide interactions

Collecting and Annotating Hetero DDIs

Given a PFAM domain of interest:

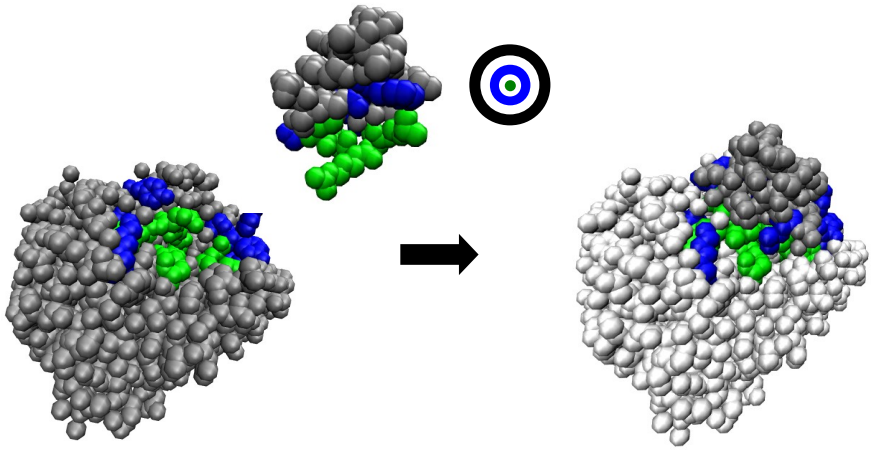
- Classify DDIs into intra, homo and hetero interactions
- Distinguish biologically relevant interactions from crystal contacts
- Eliminate duplicate or near-duplicate interactions
- Identify conserved residue positions to guide multiple structural alignments




```
Consensus  ....C..sh..ptG..s..Cp...s..hh...+a..aYs...spsppCp..pF..h.Y..u..G.C..t.G..N...p..NpFtopcc.CpptC..
1brb       -AG---EPP--YTG--P--CK---A--RI--IRY--FYN---AKAGLCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMRTA--
1bth       ---FCLEPP--YTG--P--CK---A--RI--IRY--FYN---AKAGLCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMRTCG-
1co7       --------P--YTG--P--CK---A--RI--IRY--FYN-------LCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMR----
1bun       --D-CDKPP--DTK--I--CQ---T--VV--RAF--YYK---PSAKRCV--QF--R-Y--G--G-C--N-G--N---G--NHFKSDHL-CRCECL-
1kig       ---LCIKPR--DWI-DE--CD---S--NEG-GERA-YFR---NGKGGCD--SF--W-I-------C--P-E--DHTGA--DYYSSYRD-CFNACI-
2ody       ---FCRLPA--DEG--I--CK---A--LI--PRF--YFN---TETGKCT--MF--S-Y--G--G-C--G-G--N---E--NNFETIEE-CQKACG-
```

Identifying Core and Rim Residues

- Core and rim residues form a "target"
- **Core** residues lose 75% of its accessible surface area in the complex
- **Rim** residues lose less than 75%



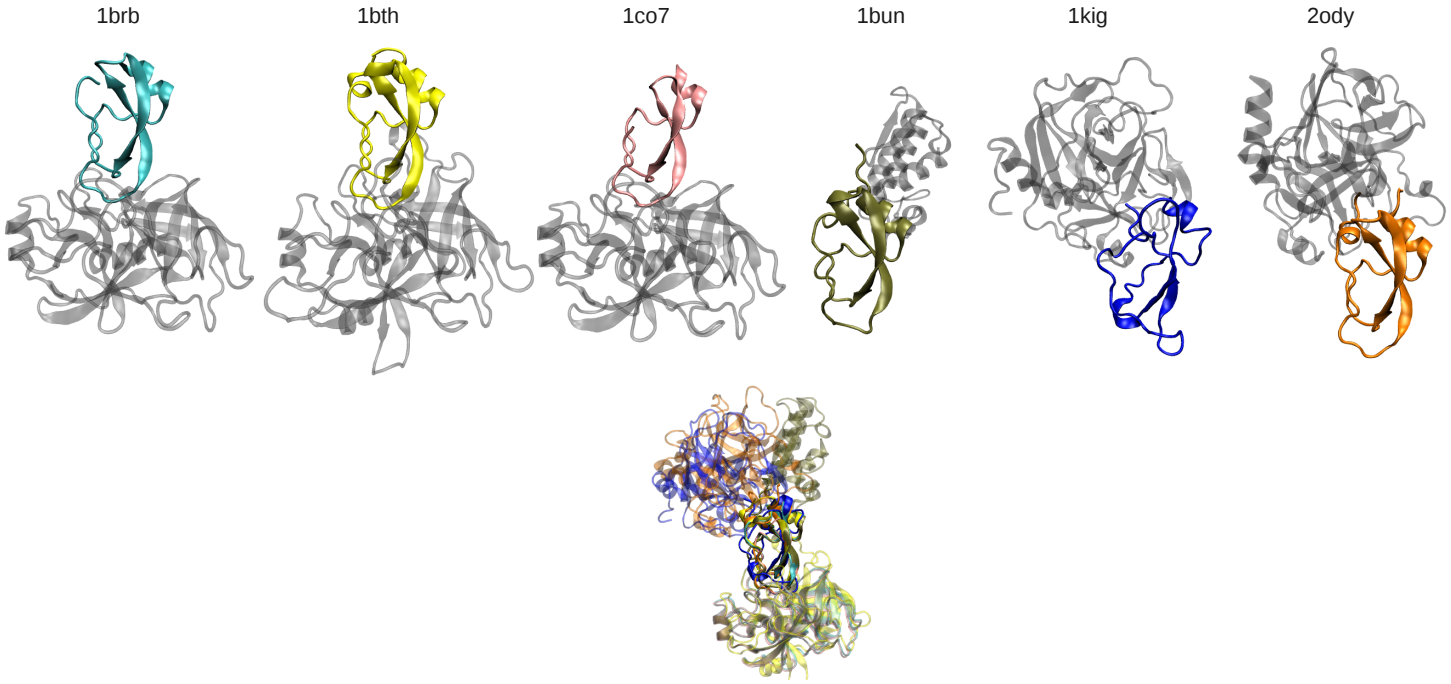
```
Consensus  ....C..sh..ptG..s..Cp...s..hh...+a..aYs...spsppCp..pF..h.Y..u..G.C..t.G..N...p..NpFtopcc.CpptC..
1brb       -AG---EPP--YTG--P--CK---A--RI--IRY--FYN---AKAGLCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMRTA--
1bth       ---FCLEPP--YTG--P--CK---A--RI--IRY--FYN---AKAGLCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMRTCG-
1co7       --------P--YTG--P--CK---A--RI--IRY--FYN-------LCQ--TF--V-Y--G--G-C--R-A--K---R--NNFKSAED-CMR----
1bun       --D-CDKPP--DTK--I--CQ---T--VV--RAF--YYK---PSAKRCV--QF--R-Y--G--G-C--N-G--N---G--NHFKSDHL-CRCECL-
1kig       ---LCIKPR--DWI-DE--CD---S--NEG-GERA-YFR---NGKGGCD--SF--W-I-------C--P-E--DHTGA--DYYSSYRD-CFNACI-
2ody       ---FCRLPA--DEG--I--CK---A--LI--PRF--YFN---TETGKCT--MF--S-Y--G--G-C--G-G--N---E--NNFETIEE-CQKACG-
```

Chakrabarti and Janin, *Prot Struct Funct Genet*, 2002

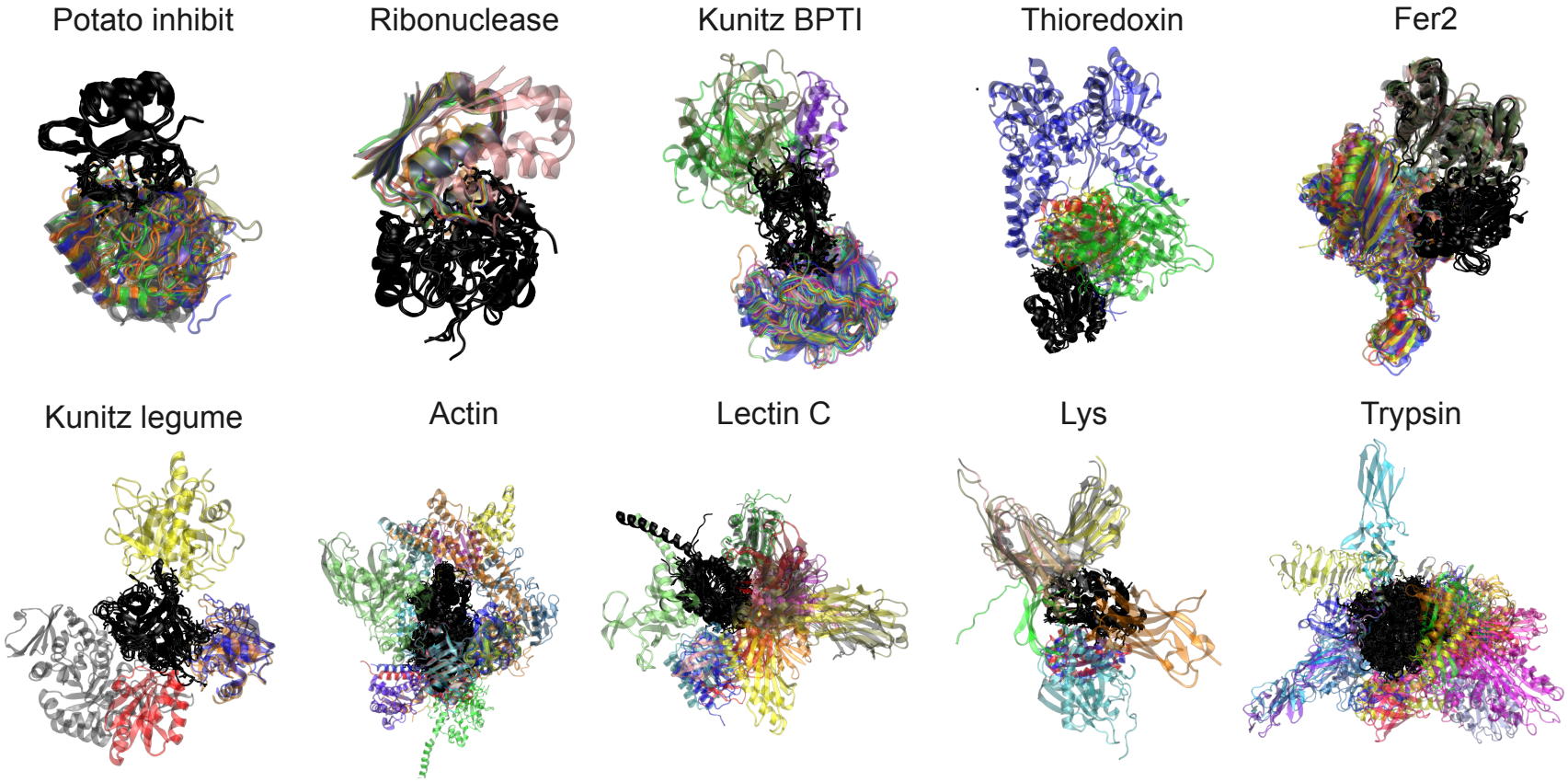
Superposing DDIs in 3D Space – E.g. *Kunitz BPTI*

For each Pfam domain family:

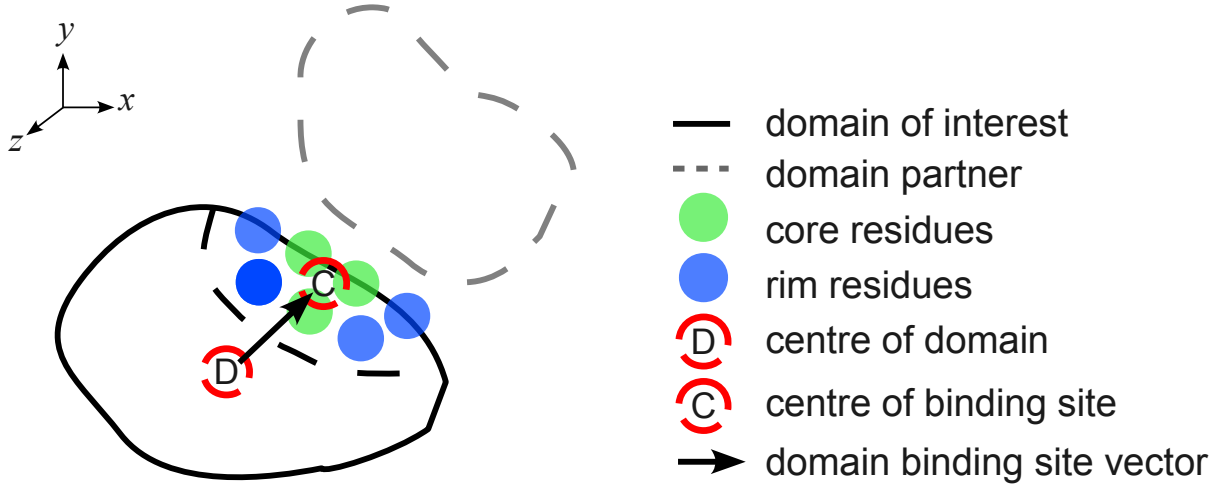
- Place all members and their interaction partners in a common frame
- Use conserved residue positions to guide structural alignment
- This reveals the overall spatial distribution



Ten Selected Domain Family Superpositions



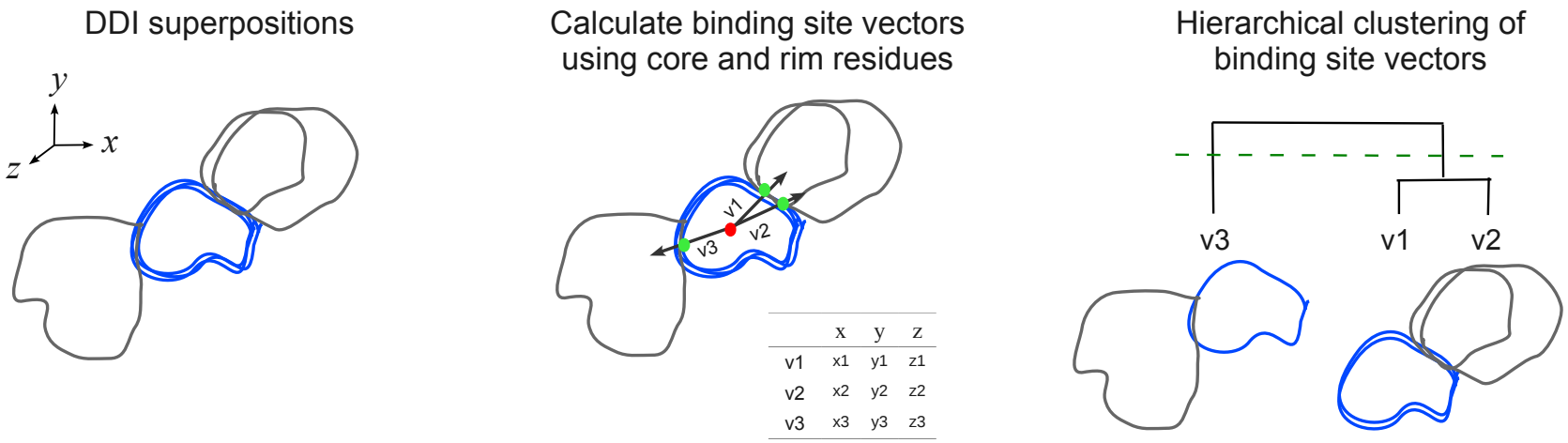
Defining Binding Site Direction Vectors



- D_i = centre of mass of domain
- C_i = geometric centre of binding site
  - calculated as a weighted average of 75% core and 25 % rim residues
- $D_i = \frac{\vec{C}_i - \vec{D}_i}{|\vec{C}_i - \vec{D}_i|}$ = binding site direction vector

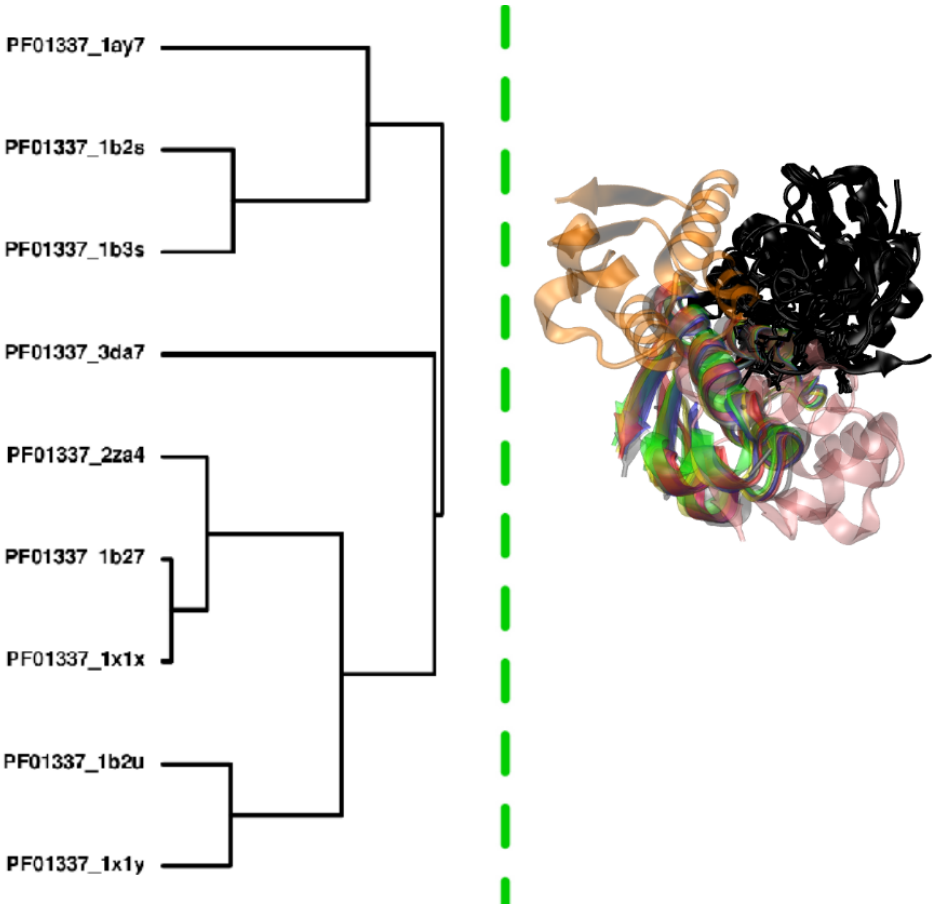
Defining Domain Family Binding Sites

- Spatial clustering of binding site direction vectors
- Ward's hierarchical clustering using Euclidean distance as metric



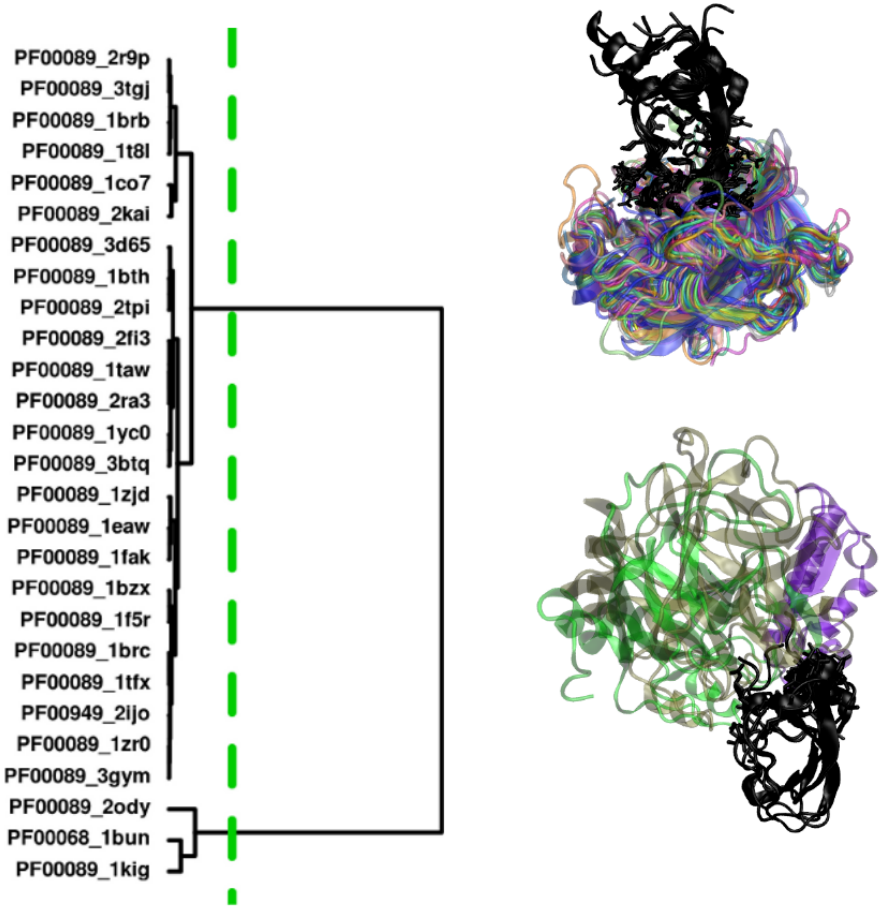
- Each cluster obtained defines a domain family binding site (DFBS)

Ribonuclease Family Has Only One Binding Site



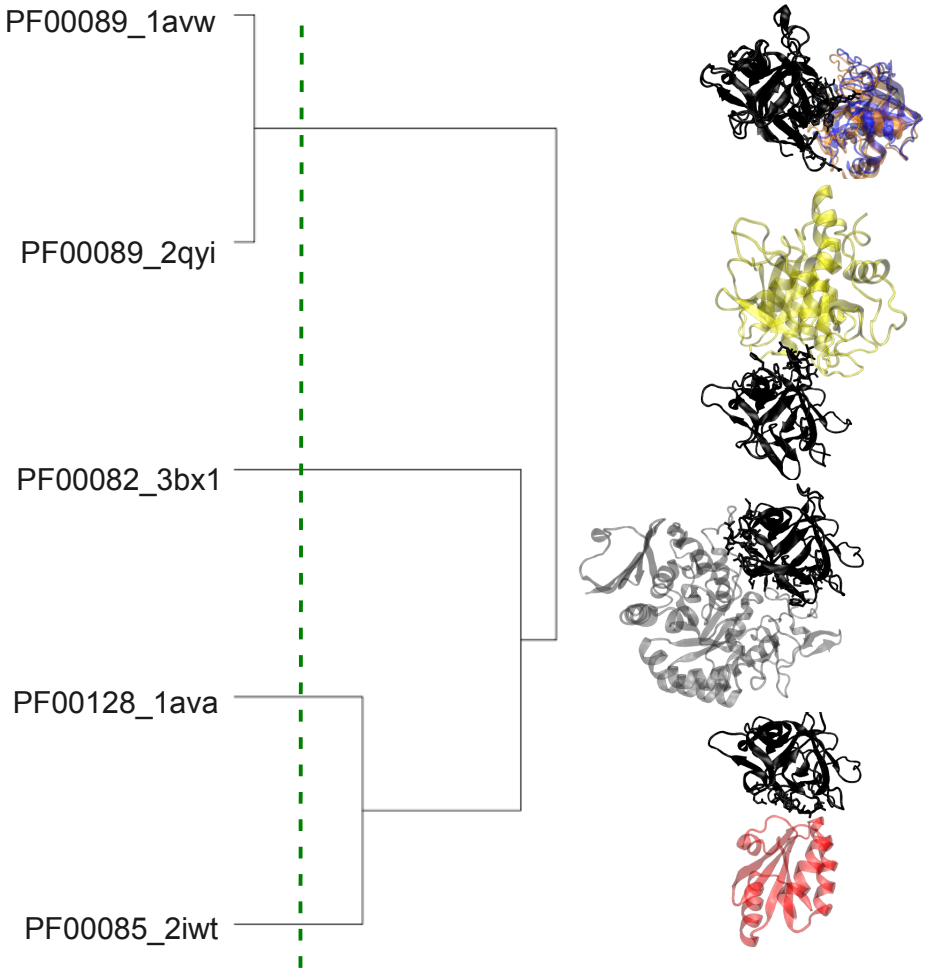
- 9 hetero DDIs involving one distinct Pfam partner

Kunitz BPTI Has Two Binding Sites



- 27 hetero DDIs involving 2 distinct Pfam partners

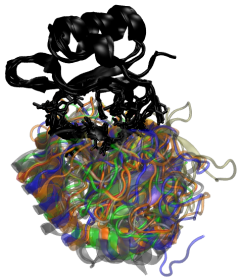
Kunitz Legume Family Has Four Binding Sites



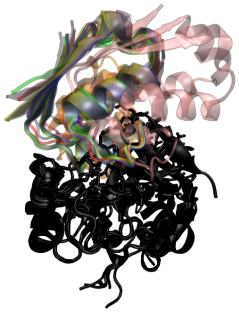
- 5 hetero DDIs involving 4 distinct Pfam partners

Calculated No. DFBSs for 10 Pfam Families

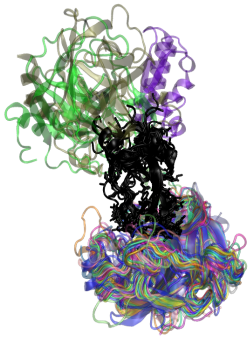
Potato inhibit (1)



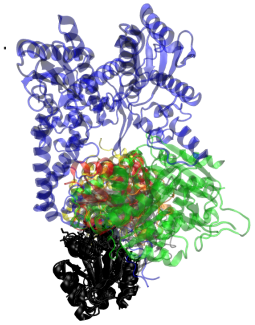
Ribonuclease (1)



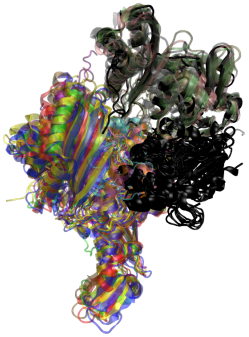
Kunitz BPTI (2)



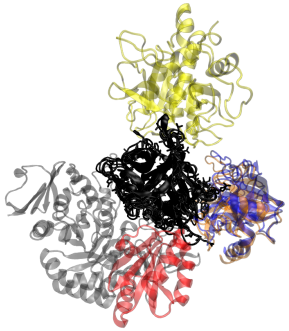
Thioredoxin (2)



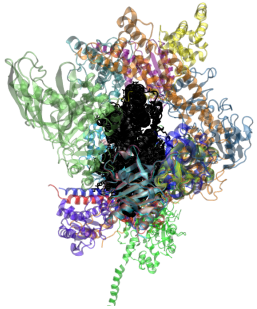
Fer2 (3)



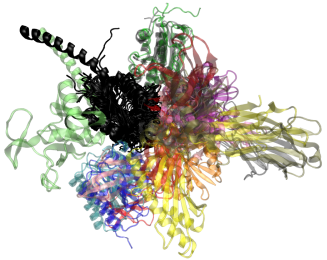
Kunitz legume (4)



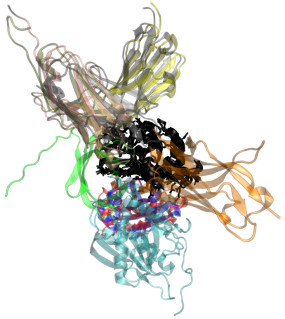
Actin (4)



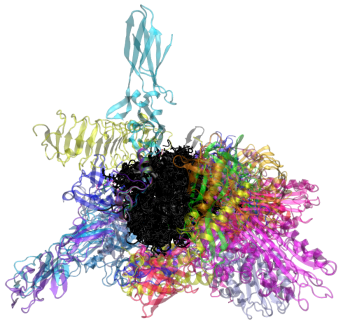
Lectin C (4)



Lys (5)

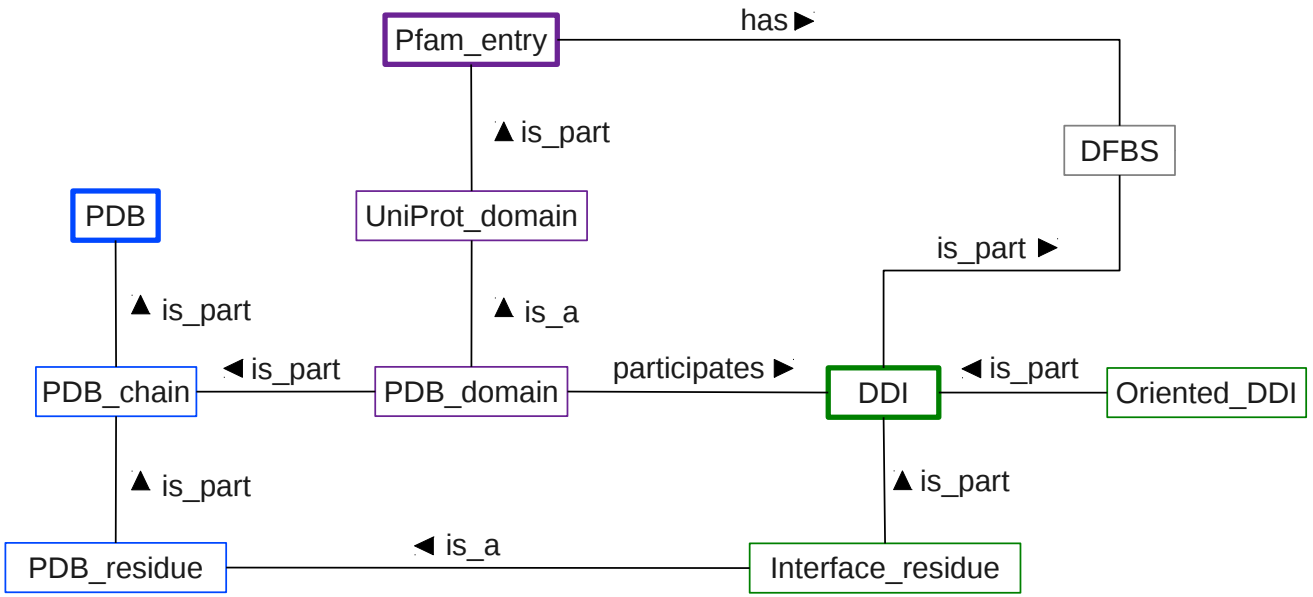


Trypsin (6)



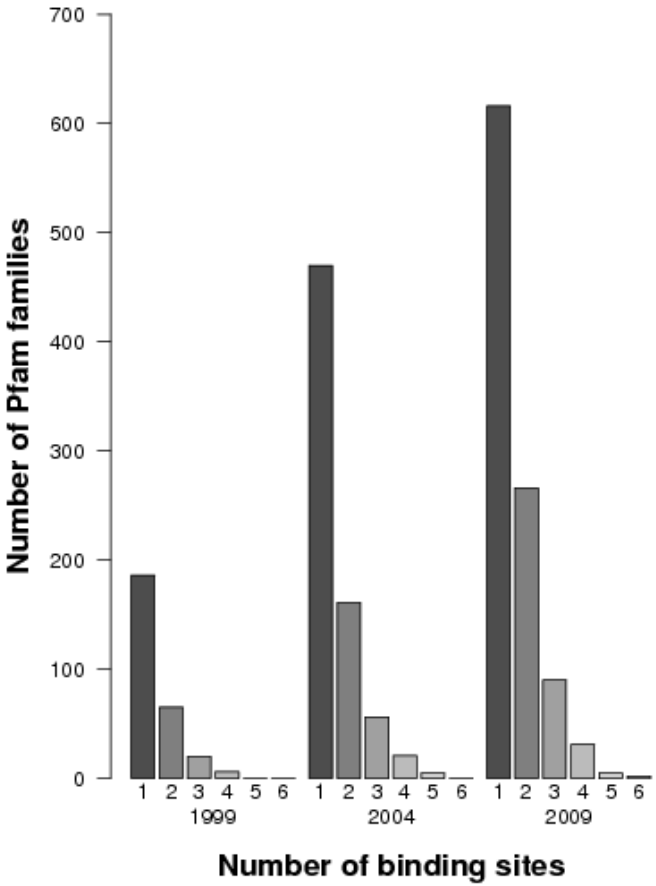
The KBDOCK Database

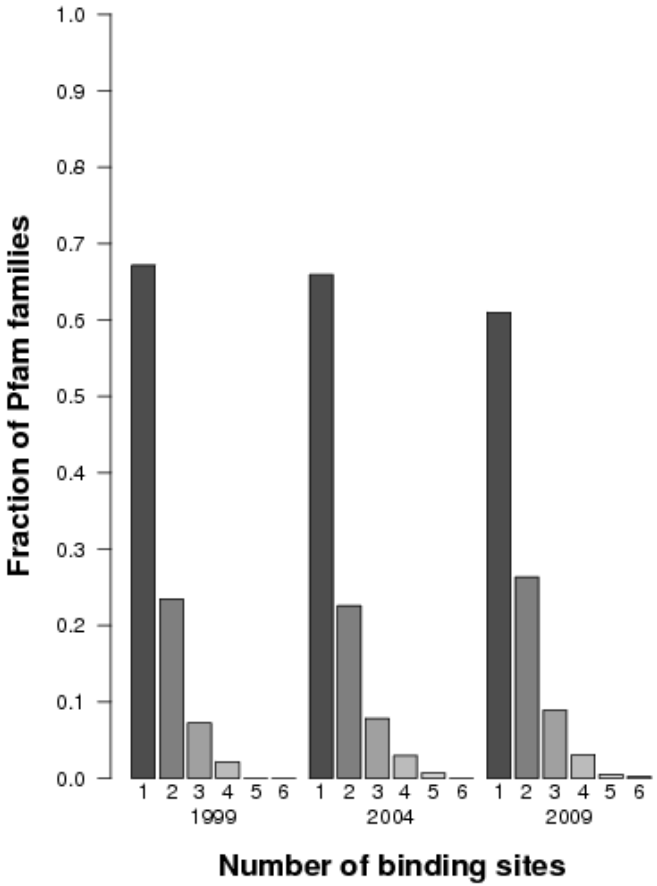
- Stores 3D DDIs by Pfam family in a MySQL database
- Statistics: 1,035 Pfam families, 2,721 NR hetero DDIs, 1,637 DFBSs



- Prolog engine for complex queries
- PHP-based web interface (http://kbdock.loria.fr)

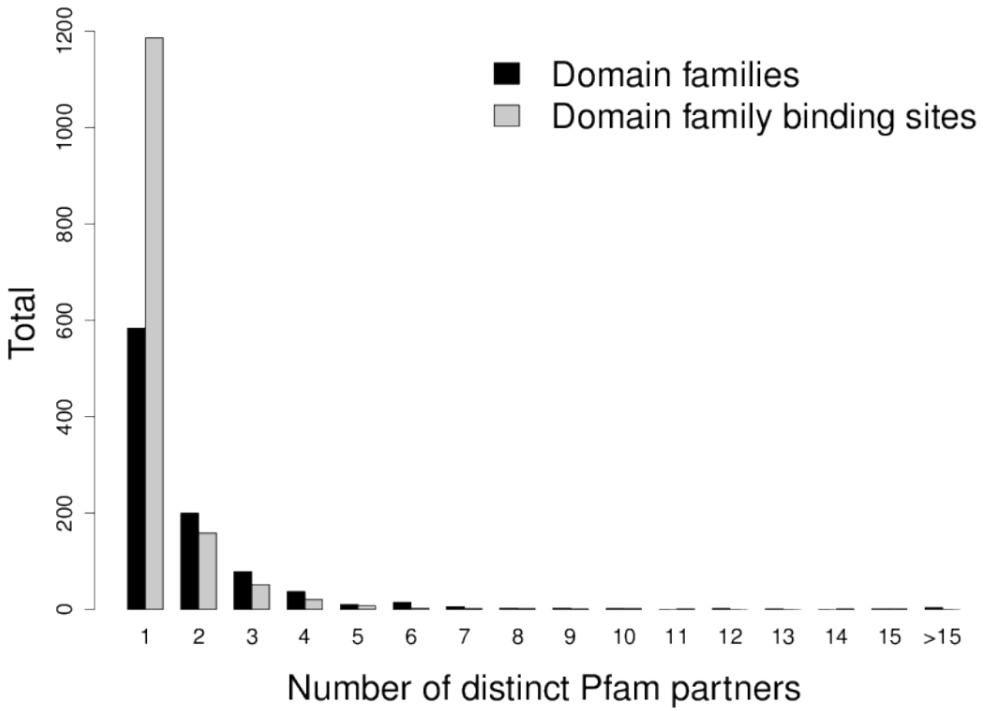
Number of DFBSs per Domain Family





- Nearly 70% of protein domain families have just one binding site
- Number of DFBS remains constant despite the growth of Pfam families

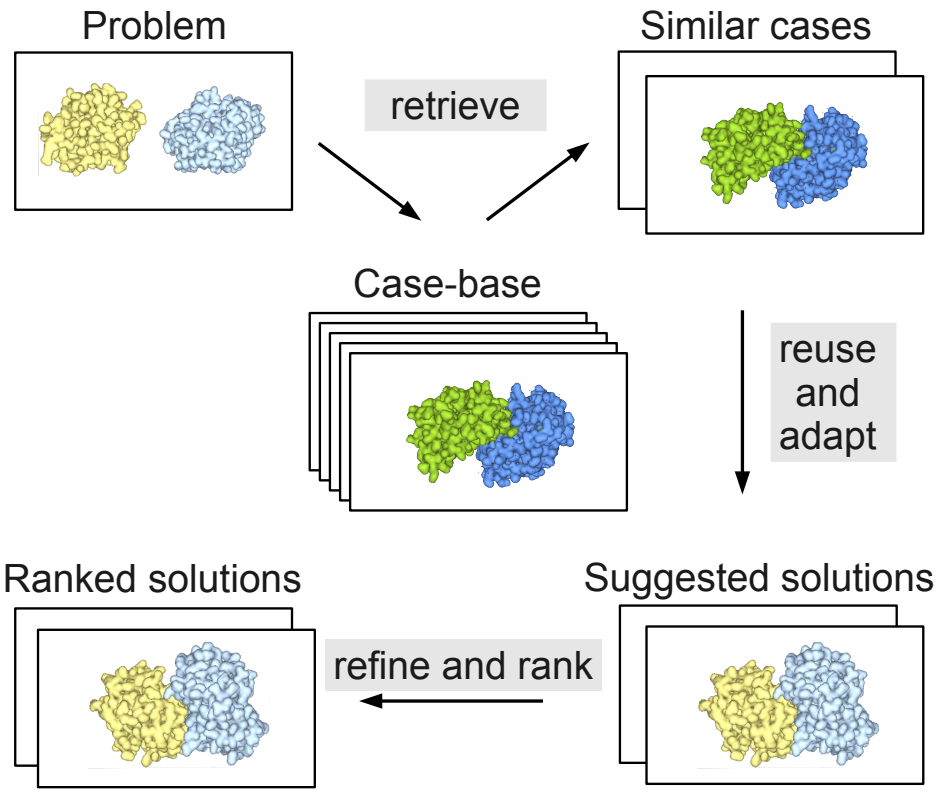
Number of Pfam Partners per DFBS



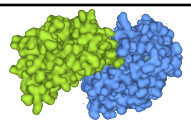
- Some 60% of Pfam families interact with just one Pfam family
- Over 80% of DFBSs interact with just one Pfam family
- 1,009 DFBS-DFBS interactions or domain-family interactions (DFIs)

Using KBDOCK for Case-Based Docking

- To use knowledge of related PPIs to help predict unknown PPIs



KBDOCK Case Representation

PDB	*1avw*	Structure	
Deposition date	*27-Sep-97*		
Expt. Technique	*X-ray diffraction*	Resolution	*1.75 Å*
Chain_1	*A*	Chain_2	*B*
Sequence_1	*IVGGYTCAANSI...*	Sequence_2	*DFVLDNEGNPL...*
PfamID_1	*Trypsin*	PfamID_2	*Kunitz_legume*
PfamAC_1	***PF00089***	**PfamAC_2**	***PF00197***
Region_1	*16-238*	Region_2	*502-675*
BindingSite_1	***2***	**BindingSite_2**	***1***
BS_res_1	*{Phe-502, ...}*	BS_res_2	*{His-57, ...}*
BS_centre_res_1	*Ser-195*	BS_centre_res_2	*Ser-560*
BS_centre_xyz_1	*(x, y, z)*	BS_centre_xyz_2	*(x, y, z)*

- $c(d1/b1, d2/b2)$ represents a DDI instance in the case base
  - $d1/b1$ means "DFBS $b1$ on domain family $d1$"
- c('PF00089'/'2', 'PF00197'/'1') identifies the above case

KBDOCK Case Retrieval

- Prolog notation
  - Lowercase for instantiated terms ('atoms')
  - Uppercase for uninstantiated terms ('variables')
- $q(d1/B1, d2/B2)$ denotes a new problem (query)
  - $d1$ and $d2$ are always instantiated
- Case unification

If $c(d1/B1, d2/B2) \in CB$
Then q is a full-homology case (FH)

Else if $c(d1/B1, D2/B2) \in CB$ and $c(D1/B1, d2/B2) \in CB$ where $D1 \neq d1$, $D2 \neq d2$
Then q is a semi-homology-two case (SH-two)

Else if $c(d1/B1, D2/B2) \in CB$ or $c(D1/B1, d2/B2) \in CB$ where $D1 \neq d1$, $D2 \neq d2$
Then q is a semi-homology-one case (SH-one)

Else q does not unify with any cases – no homology

Retrieval of FH, SH-two and SH-one Cases

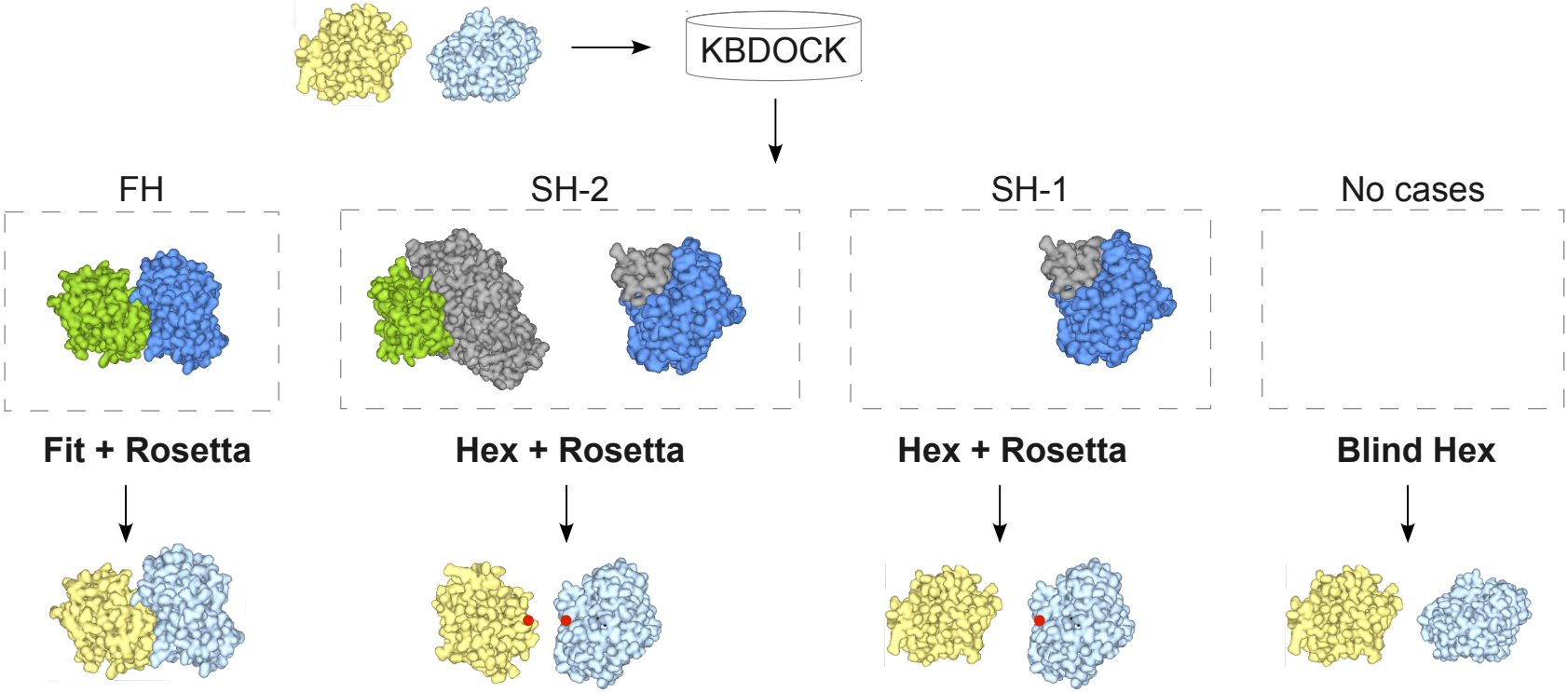
- 73 single-domain targets from Protein Docking Benchmark 4.0
- Excluding structures which have been solved after the target

Target class	Total targets	FH targets	SH-two targets	SH-one targets	No templates
No date filtering					
Enzyme	36	24 / 24	(3 + 1) / 5	3 / 5	2
Other	37	21 / 21	(0 + 0) / 3	5 / 11	2
Total	**73**	**45 / 45**	**(3 + 1) / 8**	**8 / 16**	**4**
With date filtering					
Enzyme	36	13 / 13	(2 + 1) / 5	7 / 11	7
Other	37	13 / 13	(0 + 0) / 1	8 / 15	8
Total	**73**	**26 / 26**	**(2 + 1) / 6**	**15 / 26**	**15**

Ghoorah *et al.* (2011), Bioinformatics, 27, 2820–2827

The KBDOCK Docking Pipeline

- Hex for focused rigid-body docking (keep top 200)
- RosettaDock for side-chain re-packing (refine 200x100)



Docking Benchmark Results – Full-Homology Cases

- Data set: 54 single-domain benchmark targets
- 24 FH, 4 SH-two and 26 SH-one cases

Target PDB	Type	KBDOCK only		KBDOCK+Rosetta		Blind Hex	
1cgi	E	1	5.8	1	6.5	9.1	2
1n8o	E	1	8.5	2	9.2	–	–
2sni	E	1	5.7	1	4.5	–	–
1gpw	O	1	3.8	1	5.6	8.8	5
1grn	O	1	6.4	2	2.0	–	–
3cph	O	1	9.4	1	8.9	–	–

Results summary for 24 FH cases

	KBDOCK only	KBDOCK+Rosetta	Blind Hex
Total	23/24	21/24	6/24
Avg. RMSD	8.7	5.7	8.2
Avg. Rank	1	2	4
Avg. Time (min)	2	50	3

Docking Benchmark Results – Semi-Homology Cases

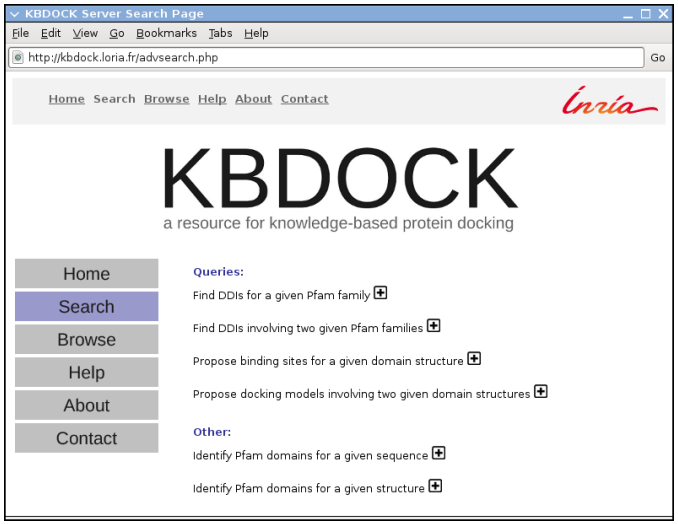
- 26 SH-one and 4 SH-two cases

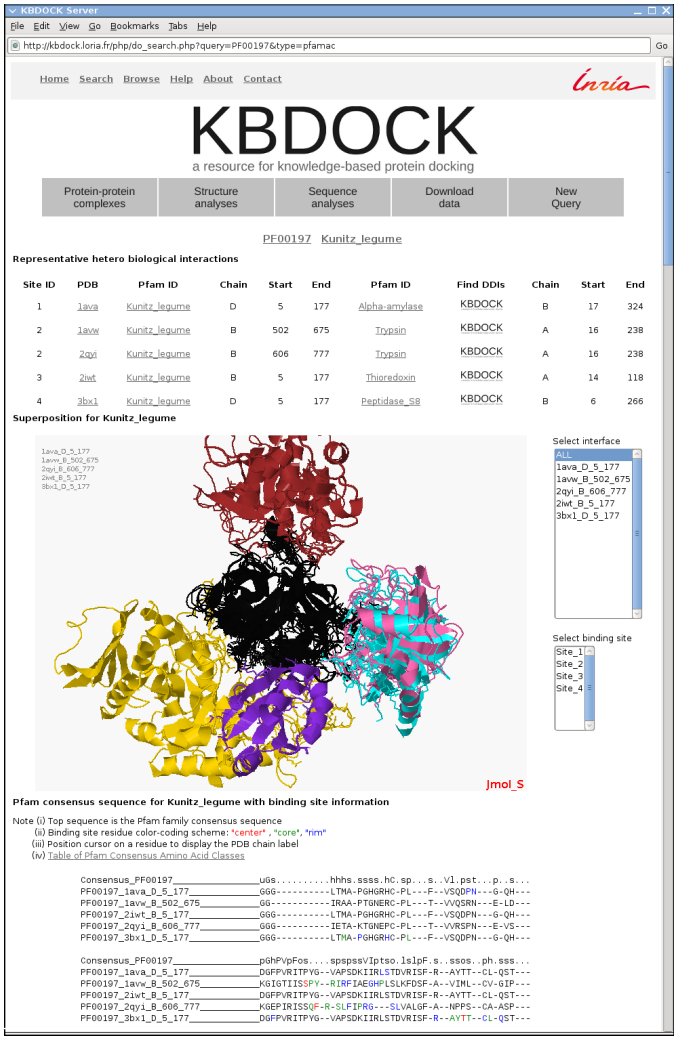
Target PDB	Type	Focused Hex		Hex+Rosetta		Blind Hex	
SH-one targets							
1fle	E	1	7.0	1	5.6	–	–
1gl1	E	8	8.1	82	7.9	6	7.3
1ppe	E	1	3.5	1	3.7	1	3.3
1lfd	O	48	8.6	–	–	–	–
SH-two targets							
1r0r	E	2	7.3	6	4.8	61	9.9
1acb	E	1	8.6	8	7.6	–	–

Results summary for 30 SH cases

	Focused Hex	Hex+Rosetta	Blind Hex
Total	8/30	5/30	6/30
Avg. RMSD	6.8	5.0	6.8
Avg. Rank	2	3	3
Avg. Time (hour)	0	175	0

KBDOCK Web Server – http://kbdock.loria.fr





Ghoorah *et al.* (2014), Nucleic Acids Research, 42, D389–D395

Conclusions – Docking by Case Based Reasoning

- KBDOCK introduces the notion of **domain family binding sites**
- KBDOCK describes 213,954 Pfam domains using 6,516 domain families
- All 3D hetero PPIs (june 2013) can be described by 4,001 DFBSs and 2,517 DFIs
- All 3D homo PPIs can be described by 12,498 DFBSs and 4,443 DFIs
- DFBSs provide a direct and easy way to do docking by homology
- FH templates usually give very high quality models
- SH templates can provide useful information for docking
- RosettaDock refinement can improve RMSD, but is very expensive

- We need a new benchmark set for docking by homology?

Acknowledgments

Anisah Ghoorah

Agence Nationale de la Recherche (ANR)

Programs and papers:

http://hex.loria.fr/
http://kbdock.loria.fr/

KBDOCK Demo – Basic Operations

- KBDOCK web site: http://kbdock.loria.fr/

- Browsing domain-domain interactions
- Viewing DDI networks
- Structural superpositions in Jmol (or JsMol)
- Structure-based sequence alignments
- Looking at structural neighbours
- Downloading structural templates
- ...
- Using Kpax (again) to superpose targets onto templates
- ...
- Ask me!

Practical Activities – 1

Finding domain interactions involving API-A

- Download the API-A data from: http://hex.loria.fr/emmsb/t40.tgz
  - t40_c.fasta (sequence), t40_c.pdb (structure)
- Use KBDOCK to find the Pfam domain for this sequence
  - Tip: the Search page allows pasting a sequence or uploading a structure
- View some representative inter-chain hetero interactions
- Can you identify LEU-87 and LYS-145 on API-A?
  - Tip: In Jmol right-mouse for Menu; then: Set Picking → Identity

Downloading the template structures

- Download and uncompress all hetero interaction partners
  - (this should give a folder called PF00197)
- Delete the 3e8l structure – this is the solution structure!

Practical Activities – 2

Modeling API-A/Trypsin by structural homology

- Use Kpax to superpose one of the Trypsin complexes onto t40_c.pdb
  - Tip: in Kpax, the "query" never moves; only the "target(s)" move(s)
- (this should give a very good template for one of the binding modes)

- For the other mode, superpose Peptidase S8 complex (1bx1) onto t40_c.pdb
- Use Hex to identify the API-A loop that interacts with Peptidase S8
- Use Hex again to place a Trypsin active site around this loop...

- Refine your proposed docking pose using a focused docking search in Hex

Practical Activities – 3

Modeling a bi-enzyme complex using structural neighbours

In our paper (Ghoorah et al. 2014, NAR, 42, D389–D395), we proposed that a GATase/ImGP cyclase complex (PDB code 1GPW, Pfam codes PF00117, PF00977) could be modeled using a complex from structural neigbours found by Kpax (PDB code 2NV2, Pfam codes PF01174, PF01680). Here, your task is to verify that the proposed model is structurally reasonable.

- Download the data provided: http://hex.loria.fr/emmsb/gatase.tgz
- Look at Figure 3 of the paper (PDF file)
- Use KBDOCK to search for structural neighbours of 1GPW
- Verify that a proposed complex is indeed 2NV2

- Using the given PDBs, use Kpax to superpose one complex onto the other
- Tip: Kpax can move two structures with one superposition using:
  - kpax -ligand query.pdb target.pdb ligand.pdb
- Tip: You can put two structures into one file using a shell command:
  - cat file_a.pdb file_b.pdb >file_ab.pdb