Protein-Protein Docking – Current Methods and New Challenges

Dave Ritchie

Team Orpailleur
Inria Nancy – Grand Est

Outline

Review of Selected CAPRI Targets

Some Algorithms Used in CAPRI

Assembling Symmetric Multimers

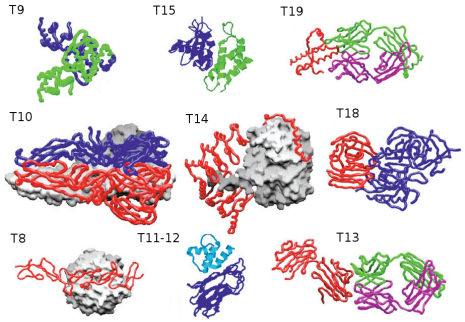
Hybrid Approaches – Knowledge-Based + MD

New Challenges – Structural Systems Biology

New Challenges – Modeling Large Molecular Machines

The CAPRI Blind Docking Experiment

- CAPRI = Critical Assessment of PRedicted Interactions
  - http://www.ebi.ac.uk/msd-srv/capri/
- Given the unbound structure, predict the unpublished 3D complex...

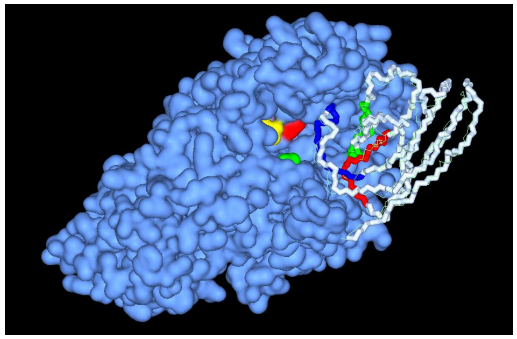


T8 = nidogen/laminin
T9 = LiCT dimer
T10 = TEV trimer
T11-12 = cohesin/dockerin
T13 = Fab/SAG1
T14 = PP1δ/MYPT1
T15 = colicin/ImmD
T18 = Xylanase/TAXI
T19 = Fab/bovine prion

- T11, T14, T19 involved homology model-building step...
- T15-T17 cancelled: solutions were on-line & found by Google !!

CAPRI Target T6 Was A Relatively Easy Target

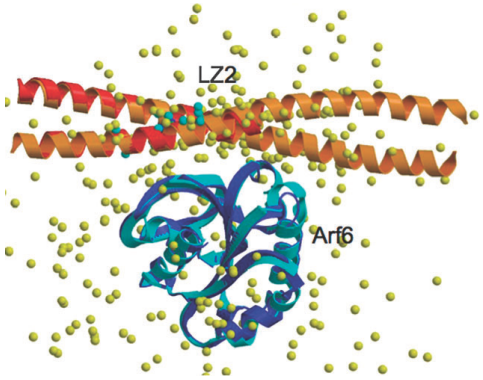
- AMD9 (camel antibody) / Amylase (pig)
- Little difference between unbound & bound conformations
- Classic binding mode: antibody loops blocking the enzyme active site



- Several CAPRI groups made "high accuracy" models (RMSD $\leq$ 1Å)

CAPRI Target T27 Was A Surprisingly Difficult Target

- Arf6 GTPase / LZ2 Leucine zipper was difficult for most predictors
  - http://www.ebi.ac.uk/msd-srv/capri/



- Circles show LZ2 centres:

blue = high quality

green = medium quality

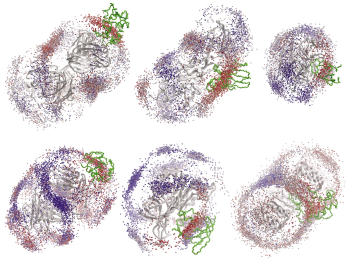
cyan = acceptable quality

yellow = wrong

Janin (2010) Molecular BioSystems, 6, 2362–2351

Predicting Protein-Protein Binding Sites

- Many algorithms/servers exist for predicting protein binding sites
  - For a review: Fernández-Recio (2011), WIREs Comp Mol Sci 1, 680–698
- Many docking algorithms show clusters of orientations – docking "funnels"



- Lensink & Wodak: docking methods are best predictors of binding sites

Fernández-Recio, Abagyan (2004), J Molecular Biology, 335, 843–865

Lensink, Wodak (2010), Proteins, 78, 3085–3095

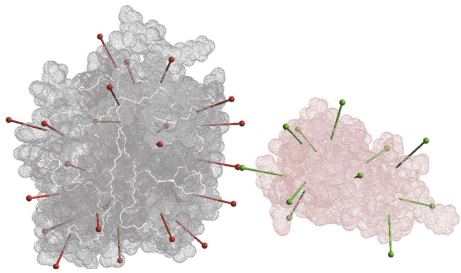
CAPRI Results: Targets 8 – 19

Software	T8	T9	T10	T11	T12	T13	T14	T18	T19
ICM	**		*	**	***	*	***	**	**
PatchDock	**	*	*	*	*	-	**	**	*
ZDOCK/RDOCK	**			*	***	***	***	**	**
FTDOCK	*		*	**	*		**	**	*
RosettaDock	-			**	***	**	***		***
SmoothDock	**				***	***	**	**	*
RosettaDock	***	-	-	**	***				**
Haddock	-	-	**	**		***	***		
ClusPro	**				***	*			*
3D-DOCK	**			*	*		**		*
MolFit	***			*	***		**		
Hex				**	***	*	*		
Zhou	-	-		-	***	**	*	*	
DOT					***	***	**		
ATTRACT	**		-	-	-	-	***		**
Valencia	*			*	*	-			-
GRAMM	-	-		-	-	-	**	**	
Umeyama				**	*				
Kaznessis	-	-			***				
Fano	-	-		*					

Mendez *et al.* (2005) Proteins Struct. Funct. Bionf. 60, 150-169

ICM Docking – Multi-Start Pseudo-Brownian Search

- Start by sticking pins in protein surfaces at 15Å intervals
- For each pair of pins, find minimum energy (6 rotations for each):
  - $E = E_{HVW} + E_{CVW} + 2.16E_{el} + 2.53E_{hb} + 4.35E_{hp} + 0.20E_{solv}$

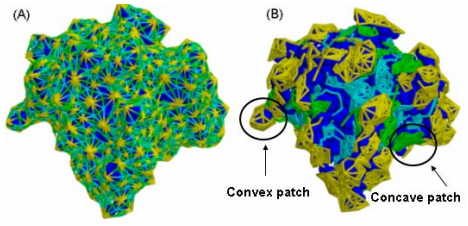


- Often gives good results, but is computationally expensive

Fernández-Recio, Abagyan (2004), J Mol Biol, 335, 843–865

PatchDock – Docking by Geometric Hashing

- Use "MS" program to calculate mesh surfaces for each protein
- Divide the mesh into convex "caps", concave "pits", and flat "belts"



- For docking, match pairs of concave/convex, and flat/any ...
- ... then test for steric clashes between rest of surfaces
- The method is fast (minutes/seconds), and gave good results in CAPRI

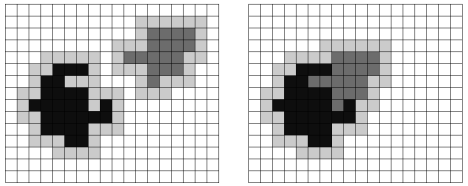
Duhovny et al. (2002), LNCS 2452, 185–200

Schneidman-Duhovny et al. (2005), NAR, 33, W363–W367

Connolly (1983), J Appl Cryst, 16, 548–558

Protein Docking Using Fast Fourier Transforms

- Conventional approaches digitise proteins into 3D Cartesian grids...



- ...and use FFTs to calculated TRANSLATIONAL correlations:

$$C[\Delta x, \Delta y, \Delta z] = \sum_{x,y,z} A[x, y, z] \times B[x + \Delta x, y + \Delta y, z + \Delta z]$$

- BUT for docking, have to repeat for many rotations – expensive!
- Conventional grid-based FFT docking = SEVERAL CPU-HOURS

Katchalski-Katzir *et al.* (1992) PNAS, 89 2195–2199

Quick Summary of FFT Docking Methods

3D Cartesian FFT Methods

- DOT (shape + electro): http://www.sdsc.edu/CCMS/DOT/
- FTDOCK (shape + electro) http://www.sbg.bio.ic.ac.uk/docking/
- GRAMM (shape?) http://vakser.bioinformatics.ku.edu/main/resources_gramm.php
- ZDOCK (shape + "ACP") http://zdock.umassmed.edu/software/
- PIPER (shape + "DARS" potential): http://cluspro.bu.edu/
- MegaDock (shape only?): http://www.bi.cs.titech.ac.jp/megadock/

Polar Fourier FFT Methods

- Hex (shape + electro): http://hex.loria.fr/
- Frodock (shape only?): http://chaconlab.org/methods/docking/frodock/

Interactive FFT with 3D Graphics

- Hex!

Knowledge-Based Protein Docking Potentials

- Several groups have developed "statistical potentials"
- Example: DARS – "Decoys As Reference State" – http://structure.bu.edu/

- Define interaction energy ("inverse Boltzmann"):
  - $E_{IJ} = -RT \ln(P_{IJ}^{nat} / P_{IJ}^{ref})$
  - P_{IJ}^{nat} = prob. that atoms I and J are in contact in native complex
  - P_{IJ}^{ref} = reference state prob., calculated from 20,000 docking decoys

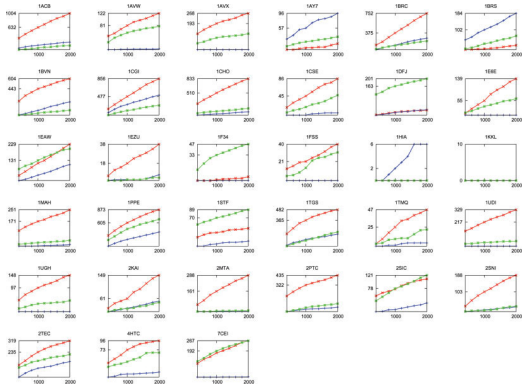
- This gives a matrix of 18 x 18 atom-type interaction energies
  - Clever trick: diagonalise matrix to get first 4 or 6 leading terms...
  - ... allows PIPER to use 4 or 6 FFTs instead of 18

- PIPER + DARS is one of the best approaches in CAPRI...

Kozakov et al. (2006) Proteins, 65, 392–406

DARS Finds More Hits Than ZDOCK or Shape-Only

- These plots compare "hits" versus "rank"

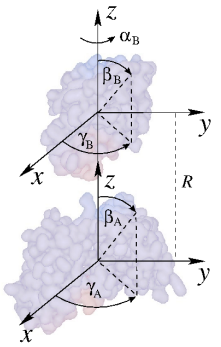


- DARS potential = red; ZDOCK (ACP) = green; shape-only = blue

Kozakov et al. (2006) Proteins, 65, 392–406

Consider Protein Docking in Polar Coordinates

- Rigid docking can be considered as a largely ROTATIONAL problem
- This means we should use ANGULAR coordinate systems



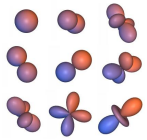
- With FIVE rotations, we should get a good speed-up?

Spherical Polar Fourier Representations

- Represent protein shape as a 3D shape-density function...

$$\tau(\underline{r}) = \sum_{nlm}^{N} a^{\tau}_{nlm} R_{nl}(r)\, y_{lm}(\theta, \phi)$$

- ...using spherical harmonic, $y_{lm}(\theta, \phi)$, and radial, $R_{nl}(r)$, basis functions



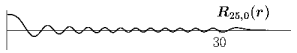
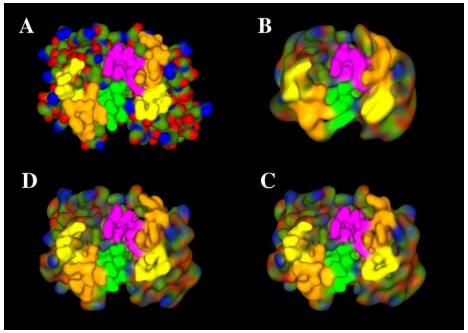
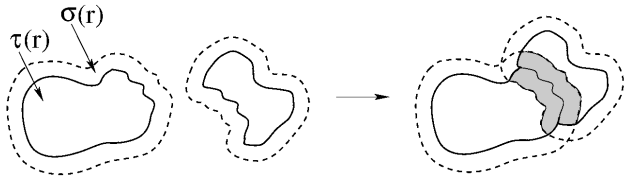


Image	Order	Coefficients
A	Gaussians	-
B	N = 16	1,496
C	N = 25	5,525
D	N = 30	9,455



Protein Docking Using SPF Density Functions



Favourable: $\int (\sigma_A(\underline{r}_A)\tau_B(\underline{r}_B) + \tau_A(\underline{r}_A)\sigma_B(\underline{r}_B))\mathrm{d}V$

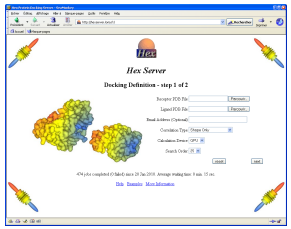
Unfavourable: $\int \tau_A(\underline{r}_A)\tau_B(\underline{r}_B)\mathrm{d}V$

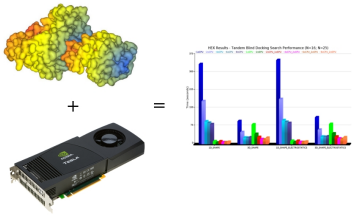
Score: $S_{AB} = \int (\sigma_A\tau_B + \tau_A\sigma_B - Q\tau_A\tau_B)\mathrm{d}V$, Penalty Factor: $Q = 11$

Orthogonality: $S_{AB} = \sum_{nlm} \left(a^{\sigma}_{nlm} b^{\tau}_{nlm} + a^{\tau}_{nlm}\left(b^{\sigma}_{nlm} - Q b^{\tau}_{nlm}\right)\right)$

Search: 6D space = 1 distance + 5 Euler rotations: $(R, \beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B)$

HexServer – GPU-Accelerated Web Server



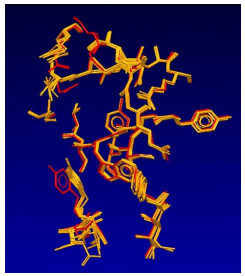


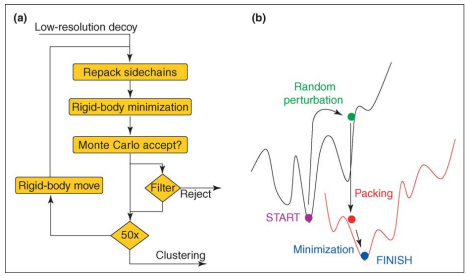
- Very fast – can cover 6D search space using 1D, 3D, or 5D FFTs...
- "Easy" to accelerate the 1D FFTs on highly parallel GPUs ...
- Widely used around the world – 33,000 downloads...

http://www.loria.fr/hex/ and http://www.loria.fr/hexserver/

RosettaDock – Flexible Side Chain Re-Packing

- Given a rigid body starting pose, repeat 50 times:
  - REMOVE and RE-BUILD side chains
  - Minimise as rigid-body with Monte-Carlo accept/reject

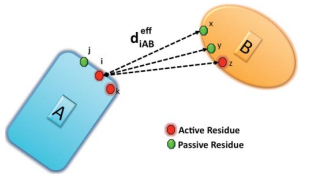


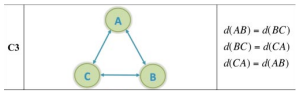


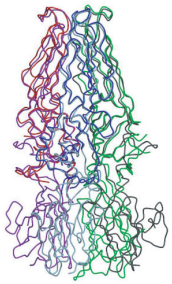
- Successful on several CAPRI targets and 50% of Docking Benchmark v2

Haddock – "Highly Ambiguous Data-Driven Docking"

- Flexible refinement using CNS with ambiguous interaction restraints (AIRs)
- Use of "active" and "passive" residues ensures active residues at interface
- E.g. residue *i* of protein A: $d_{iAB}^{\text{eff}} = \left(\sum_{m_{iA}=1}^{N_{iA}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{kB}} \left(\frac{1}{d_{m_{iA},n_{kB}}^{6}} \right) \right)^{-1/6}$







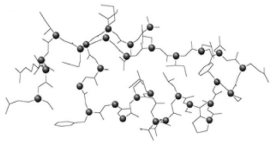
T10 = TEV trimer

- Restraints from:
  - SAXS
  - mutagenesis
  - mass spec
  - NMR

van Dijk et al. (2005) FEBS J, 272, 293–312
van Dijk et al. (2005) Proteins, 60, 232–238

Modeling Protein Flexibility Using Elastic Network Models

- ENMs assume protein C_α atoms are coupled via a harmonic potential ..
  - V=potential, d_{ij}=distance, d_{ij}^0=ref distances, $\underline{H}$=Hessian, C=const
  - $\underline{E}$=eigenvector matrix, $\underline{e}_i$=normal modes, Λ_{ii}=magnitudes



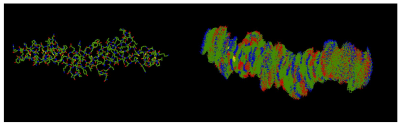
$$V = \sum_{i<j} C(d_{ij} - d_{ij}^0)^2$$
$$H_{ij} = (\partial/\partial x_i)(\partial/\partial x_j)V$$
$$\underline{H} = \underline{E}^T.\underline{\Lambda}.\underline{E}$$

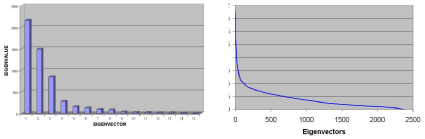
- Then, represent protein as a linear combination of first eigenvectors:
  - $\underline{P}^{NEW} = \underline{P}^0 + \sum_{i=6}^{3N} w_i \underline{e}_i$
- On-line examples:
  - *ElNémo* web-server: http://www.igs.cnrs-mrs.fr/elnemo/
  - Macromolecular Movements: http://www.molmovdb.org/

Tirion (1996), Physical Review Letters, 77, 1905–1908 (first paper)
Andrusier et al. (2008), Proteins, 73, 271–289 (review

Simulating Flexibility Using "Essential Dynamics"

- Generate distance-constrained samples in CONCOORD, then apply PCA





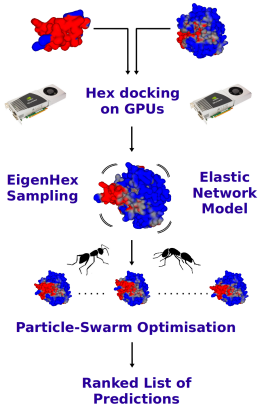
- Covariance matrix, C:
  $C_{ij} = < (x_i - \overline{x}_i)(x_j - \overline{x}_j) >$
- Eigenvectors, E:
  $\underline{C} = \underline{E}.\Lambda.\underline{E}^T$
- Conformations, P:
  $\underline{P}^{NEW} \simeq \underline{P}^0 + \sum_{k=1}^{n} \alpha_k \underline{e}_k$

- First eigenvectors encode most of RMSD between bound and unbound
- See also SwarmDock – http://bmm.cancerresearchuk.org/~SwarmDock/

Mustard, Ritchie (2005), Proteins 60, 269–274 (first NMA protein docking?)
Moal, Bates (2010) Int J Molecular Sciences, 11, 3623–3648 (SwarmDock)

EigenHex – Flexible Docking Using Pose-Dependent ENM

- Apply fresh eigenvector analysis to the top 1,000 Hex orientations



Overall approach:

- C_α elastic network model (ENM)
- Use up to 20 eivenvectors
- Search using PSO
- Score using DARS potential

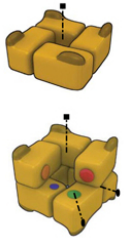
Results:

- DARS works well but...
- Still need better scoring function
- Much effort – small improvement!!

Venkatraman, Ritchie (2012), Proteins, 80, 2262–2274

Docking Symmetric Structures

Several groups have developed symmetry docking algorithms



- Molfit (D_2): Berchanski et al. (2003), Proteins, 53, 817–829
- M-ZDOCK (C_n): Pierce et al. (2005), Bioinformatics, 21, 1472–1478
- SymmDock (C_n): Schneidman et al. (2005), Proteins, 60, 224–231
- Cluspro (C_n,D_2, D_3): Comeau et al. (2005), JSB, 150, 233-244

(these algorithms "post-filter" blind docking searches)

Symmetric complexes are remarkably common in the PDB

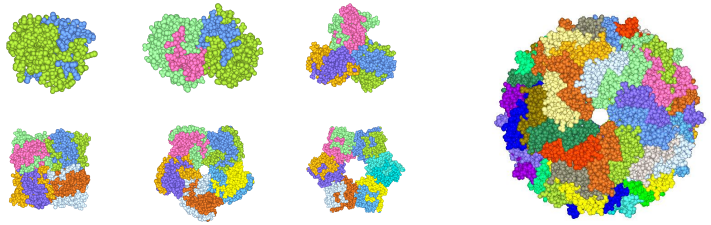
n	2	3	4	5	6	7	8
Cn	8740	992	223	107	76	29	5
Dn	2111	585	173	46	20	23	6

(data from: http://www.3dcomplex.org)

Coming Soon: "SAM" – Symmetry Assembler

Uses multiple 1D Polar Fourier FFT searches

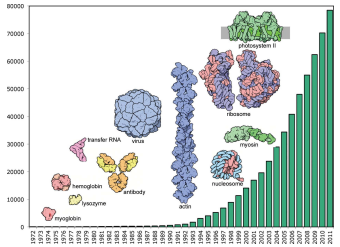
- Implemented for all point group symmetries: C_n, D_n, T, O, I
- Works well for small protein domains...



- Need to develop coarse-grained scoring for large proteins
- Need to extend to symmetric cryo-EM density fitting...

Systems Biology View of Protein-Protein Interactions

Protein interactions are central to many biological systems

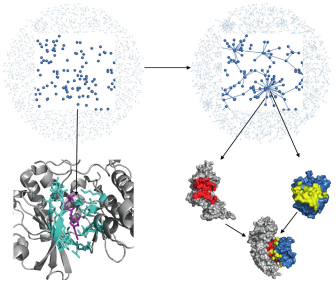


Each protein is part of a large network of interactions

- To understand how proteins really work, we need to know their three-dimensional structures... But solving structures is difficult!
- We need to exploit knowledge of known structures and interactions...

Protein-Protein Interaction Challenges

- Can we predict all interactions within a proteome – the interactome?

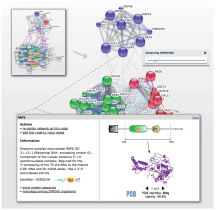


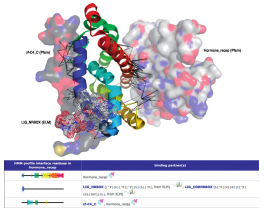
- For each interaction, can we predict the interface and 3D complex?
- For each protein can we predict its ligand binding sites?

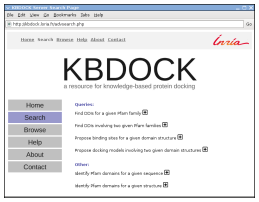
Wass, David, Sternberg (2011) Current Opinion in Structural Biology, 21, 382–390

Protein-Protein Interaction Resources

- STRING – Search Tool for Retrieval of Interacting Genes
  - 12 million known PPIs; 44 million predicted – http://string.embl.de/
- 3DID – 160,000 DDIs – http://3did.irbbarcelona.org/
- KBDOCK – Knowledge-Based Docking ("Domain Family Binding Sites")
  - 280,000 DDIs + 4,000 DFBIs – http://kbdock.loria.fr/



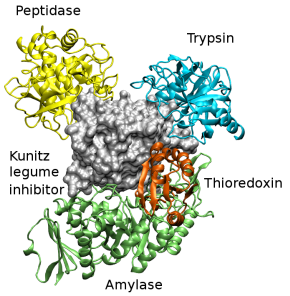


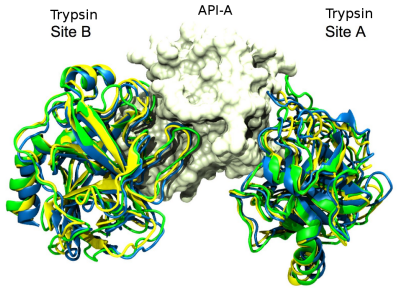


Szklarzyk et al. (2011), Nucleic Acids Research, 39, D561–D568
Stein et al. (2010), Nucleic Acids Research, 33, D413–D417
Ghoorah et al. (2014), Nucleic Acids Research, 42, D389–D395

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

- It was given that there were TWO different binding sites
- We searched SCOPPI and 3DID for similar 3D interactions
- This helped to identify two inhibitory loops on API-A



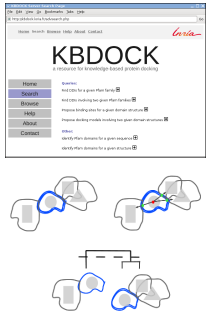


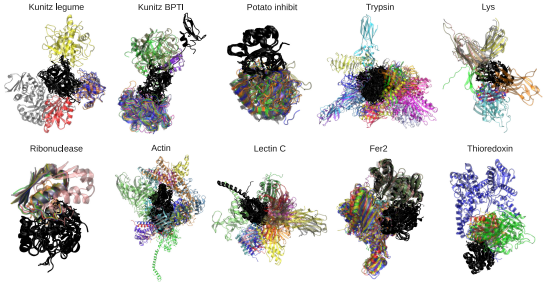
- Using Hex + MD refinement gave NINE "acceptable" solutions

The KBDOCK Database and Web Server

- Domains are superposed and clustered by PFAM family
- $\sim$ 8,000 non-redundant domain family binding sites (DFBSs)
- $\sim$ 20,000 domain family interactions (DFIs)

http://kbdock.loria.fr/

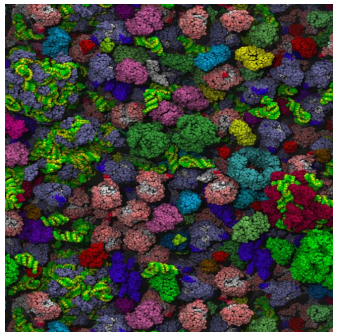




Ghoorah et al. (2014) NAR, 42, D389-D395

The Inside of a Cell is Highly Crowded

- This image shows a model of the cytoplasm in *E. Coli*

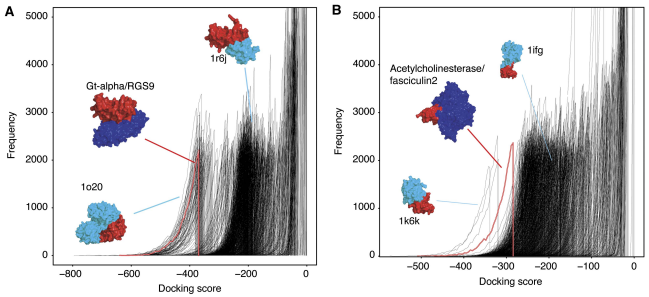


- Can we use docking algorithms to predict the protein-protein interactions ?

McGuffee, Elcock (2009), PLoS Comp Biol, 6, e1000694

Large-Scale Cross-Docking Using Hex

- Wass et al. cross-docked 56 true pairs with 922 non-redundant "decoys"
- For each pair, they plotted the profile of the best 20,000 docking scores...
- (-ve scores are good; red/blue = correct PPI; red/cyan = incorrect interactions)

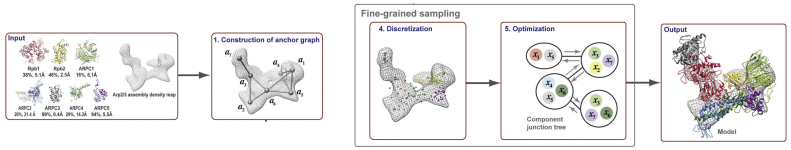


- 48/56 true PPIs have significantly higher energies than false pairs
- Only 8/56 true PPIs have indistinguishable profiles to the non-binders

Wass et al. (2011) Molecular Systems Biology, 7, article 469

IMP – Integrative Modeling Platform

- Python system for multi-component modeling – http://salilab.org/imp/
- Combines data from: cryoEM (mainly), X-Ray, NMR, SAXS, Modeller, ...
- ... with with interaction data from BioGRID – http://thebiogrid.org/



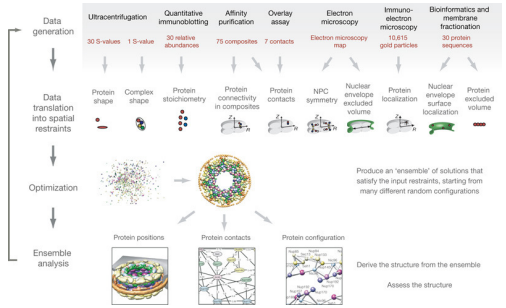
- Minimise multi-term objective function:
  - $F = \sum_i \alpha_i + \sum_{i<j} \beta_{ij}$
  - $\alpha_i$are single-body terms (e.g. density fitting score, protrusion penalty)
  - β_{ij} are two-body terms (e.g. docking scores)
- But it is a **highly combinatorial** search space, with missing/incomplete data...

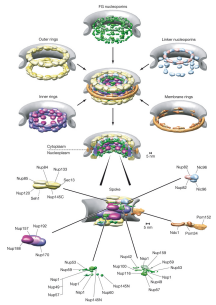
Russel et al. (2012) PLoS Biology, 10, e1001244

Lasker et al. (2009) J Molecular Biology, 388, 180–194

Putting The Pieces Together – The Nuclear Pore Complex

- The NPC has some 650 components – raw data at http://salilab.org/npc/





- It required an immense multi-disciplinary effort to build this model ...
- See Dreyfuss et al. for an interesting computational validation of the model

Alber *et al.* Nature (2007) 450, 683–694 and 695–701

Dreyfuss *et al.* Proteins (2012) 80, 2125–2136

Conclusions

- (+) Better potentials are helping to improve pair-wise docking
- (+) Cross-docking can detect true partners remarkably often
- (+) General symmetry assembly is "coming soon"...
- (−) Modeling protein flexibility during docking is still difficult
- (+) Knowledge-based protein docking is becoming very useful
  - Most Pfam families have just one binding site – often re-used
- (+) Current strategy: "data-driven" or "knowledge-based" docking
- (?) The next challenge – modeling "the structural interactome"
  - All-vs-all docking ?
  - Electron-microscopy density fitting ?
  - Assembling multi-component machines ?

Thank You!

Acknowledgments

Anisah Ghoorah
Matthieu Chavent
Diana Mustard
Vishwesh Venkatraman
Lazaros Mavridis

BBSRC, EPSRC, ANR

Hex program and papers:

http://hex.loria.fr/