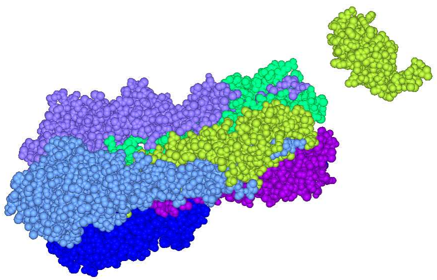
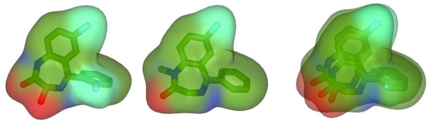
Protein Docking and Virtual Screening using Polar Fourier Correlations





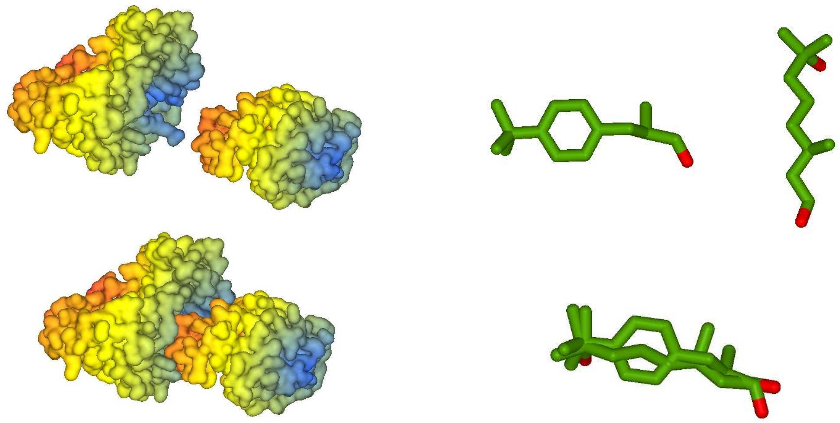
Dave Ritchie

Orpailleur Team

INRIA Nancy – Grand Est

Docking and Shape Matching are Both Recognition Problems

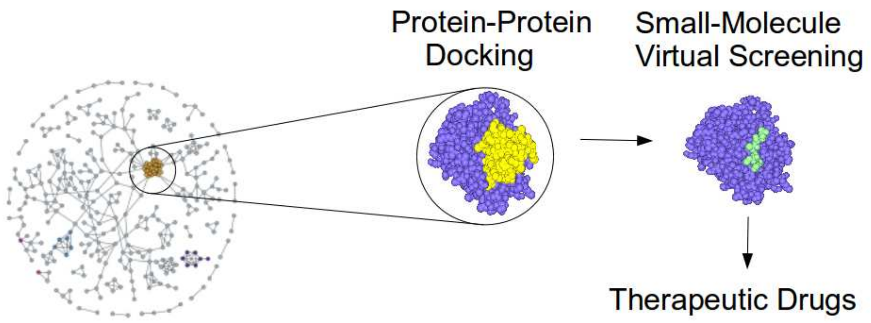
- Ignoring flexibility, docking and shape matching are both 6D search problems



- The challenge – find computationally efficient representations for:
  - protein docking ↔ translational + rotational search
  - ligand shape matching ↔ mainly rotational search

Protein-Protein Interactions and Therapeutic Drug Molecules

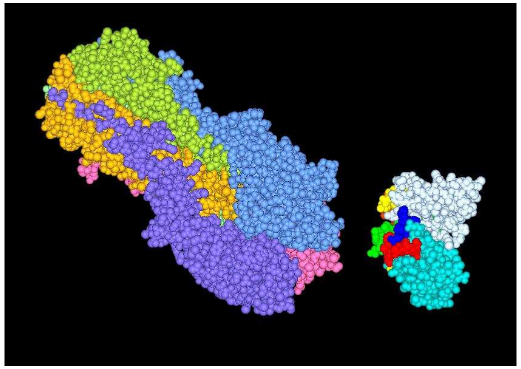
- Protein-protein interactions (PPIs) define the machinery of life
- Humans have about 30,000 proteins, each having about 5 PPIs



- Understanding PPIs could lead to immense scientific advances
- Small “drug” molecules often inhibit or interfere with PPIs

Why is Protein Docking Difficult ?

- Protein docking = predicting protein interactions at the molecular level

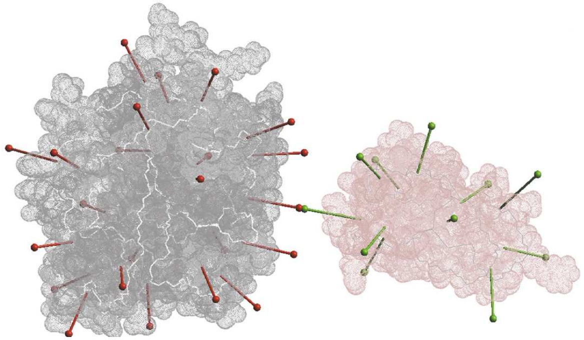


- If proteins are rigid => six-dimensional search space
- But proteins are flexible => multi-dimensional space!
- Current scoring functions cannot predict protein-protein binding affinity

ICM – Multi-Start Pseudo-Brownian Monte-Carlo Energy Minimisation

- Start by sticking "pins" in protein surfaces at 15Å intervals
- Find minimum energy for each pair of starting pins (6 rotations each):

$$E = E_{HVW} + E_{CVW} + 2.16E_{el} + 2.53E_{hb} + 4.35E_{hp} + 0.20E_{solv}$$

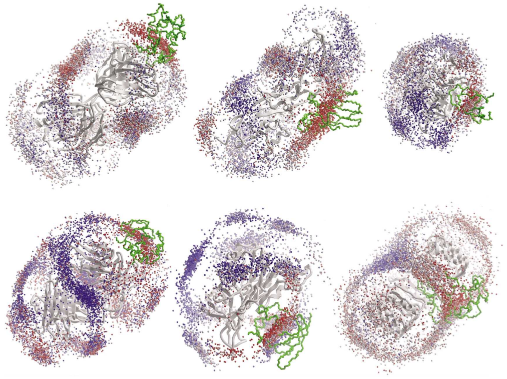


- Often gives good results, but is computationally expensive

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

Predicting Protein-Protein Binding Sites

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

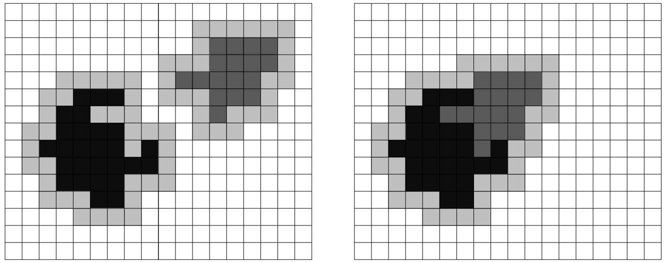


- Lensink & Wodak proposed that docking methods are the best predictors of binding sites

Fernández-Recio, Abagyan (2004), J Mol Biol, 335, 843–865
Lensink, Wodak (2010), Proteins, 78, 3085–3095

Protein Docking Using Fast Fourier Transforms

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



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

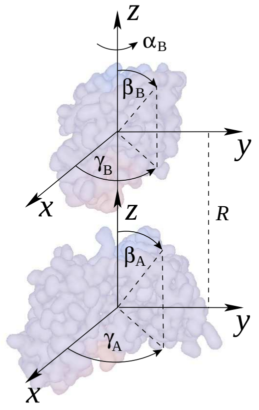
$$C[\Delta x, \Delta y, \Delta z] = \sum_{x,y,z} A[x,y,z] \times B[x+\Delta x, y+\Delta y, z+\Delta z]$$

- 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

Protein Docking Using Polar Fourier Correlations

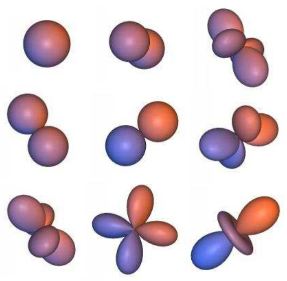
- Rigid body docking can be considered as a largely ROTATIONAL problem
- This means we should use ANGULAR coordinate systems



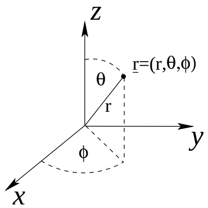
- With FIVE rotations, we should get a good speed-up?

Some Theory – The Spherical Harmonics

- The spherical harmonics (SHs) are examples of classical “special functions”



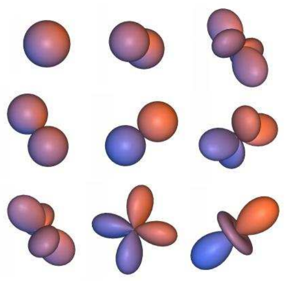
- Spherical polar coordinates: $\underline{r} = (r, \theta, \phi)$



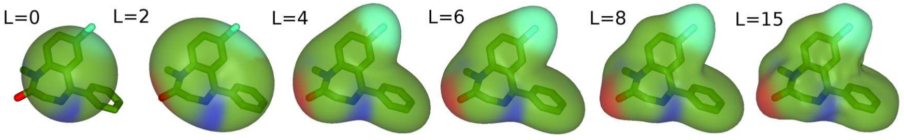
- The spherical harmonics are products of Legendre polynomials and circular functions:
  - Real SHs: $y_{lm}(\theta,\phi) = P_{lm}(\theta)\cos m\phi + P_{lm}(\theta)\sin m\phi$
  - Complex SHs: $Y_{lm}(\theta,\phi) = P_{lm}(\theta)e^{im\phi}$
  - Orthogonal: $\int y_{lm}y_{kj}\mathrm{d}\Omega = \int Y_{lm}Y_{kj}\mathrm{d}\Omega = \delta_{lk}\delta_{mj}$
  - Rotation: $y_{lm}(\theta',\phi') = \sum_j R^{(l)}_{jm}(\alpha,\beta,\gamma)y_{lj}(\theta,\phi)$

Spherical Harmonic Molecular Surfaces

- Use SHs as orthogonal shape “building blocks”:



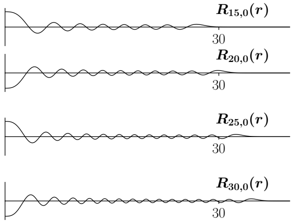
- Encode distance from origin as SH series to order L:
- $r(\theta,\phi) = \sum_{l=0}^{L}\sum_{m=-l}^{l} a_{lm}y_{lm}(\theta,\phi)$
- Reals SHs: $y_{lm}(\theta,\phi)$
- Coefficients: a_{lm}
- Solve the coefficients by numerical integration
- Normally, L=6 is sufficient for good overlays

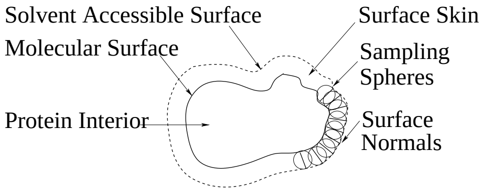


Ritchie and Kemp (1999) J. Comp. Chem. 20 383–395

Docking Needs a 3D “Spherical Polar Fourier” Representation

- Need to introduce special orthonormal Laguerre-Gaussian radial functions, $R_{nl}(r)$
- $R_{nl}(r) = N^{(q)}_{nl}e^{-\rho/2}\rho^{l/2}L^{(l+1/2)}_{n-l-1}(\rho); \qquad \rho = r^2/q, \quad q = 20.$





- Surface Skin: $\sigma(\underline{r}) = \begin{cases} 1; & \underline{r} \in \text{surface skin} \\ 0; & \text{otherwise}\end{cases}$ Interior: $\tau(\underline{r}) = \begin{cases} 1; & \underline{r} \in \text{protein at} \\ 0; & \text{otherwise}\end{cases}$
- Parametrise as: $\sigma(\underline{r}) = \sum_{n=1}^{N}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} a^{\sigma}_{nlm}R_{nl}(r)\,y_{lm}(\theta,\phi)$
- TRANSLATIONS: $a^{\sigma\prime\prime}_{nlm} = \sum_{n'l'}^{N} T^{(|m|)}_{nl,n'l'}(R)a^{\sigma}_{n'l'm}$

SPF Protein Shape-Density Reconstruction

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

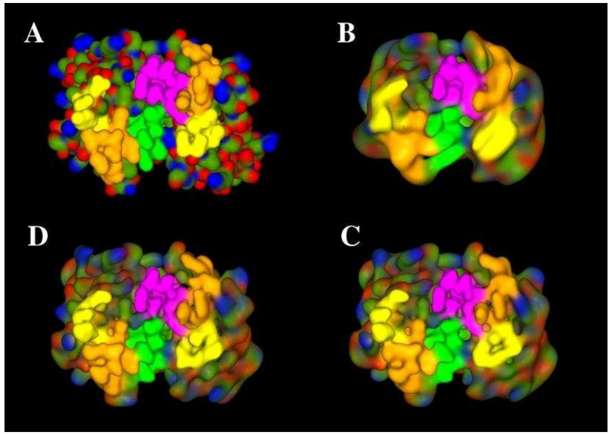
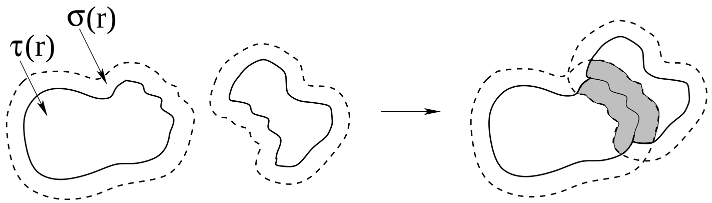


Image	Order	Coefficients
A	Gaussians	-
B	N = 16	1,496
C	N = 25	5,525
D	N = 30	9,455

Ritchie (2003) Proteins Struct. Funct. Bionf. 52 98–106

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

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

D.W. Ritchie and G.J.L. Kemp (2000) Proteins Struct. Funct. Bionf. 39 178–194

Hex Polar Fourier Correlation Example – 3D Rotational FFTs

- **Set up 3D rotational FFT as a series of matrix multiplications...**

Rotate: $a'_{nlm} = \sum_{t=-l}^{l} R^{(l)}_{mt}(0, \beta_A, \gamma_A)a_{lt}$

Translate: $a''_{nlm} = \sum_{kj}^{N} T^{(|m|)}_{nl,kj}(R)a'_{kjm}$

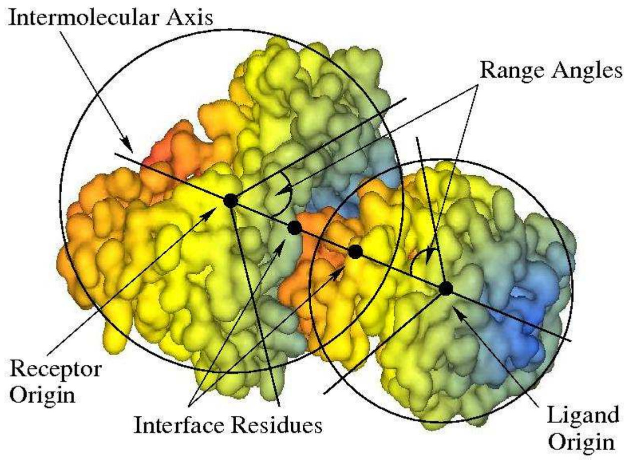
Real to complex: $A_{nlm} = \sum_{t} a''_{nlt}U^{(l)}_{tm}, \qquad B_{nlm} = \sum_{t} b_{nlt}U^{(l)}_{tm}$

Multiply: $C_{muv} = \sum_{nl} A^{*}_{nlm}B_{nlv}\Lambda^{um}_{lv}$

3D FFT: $S(\alpha_B, \beta_B, \gamma_B) = \sum_{muv} C_{muv}e^{-i(m\alpha_B + 2u\beta_B + v\gamma_B)}$

- **On one CPU, docking takes from 15 to 30 minutes**

Exploiting Proir Knowledge in SPF Docking



- **Knowledge of even only one key residue can reduce search space enormously...**
- **This accelerates the calculation and helps to reduce false-positive predictions**

The CAPRI Experiment (Critical Assessment of PRedicted Interactions)

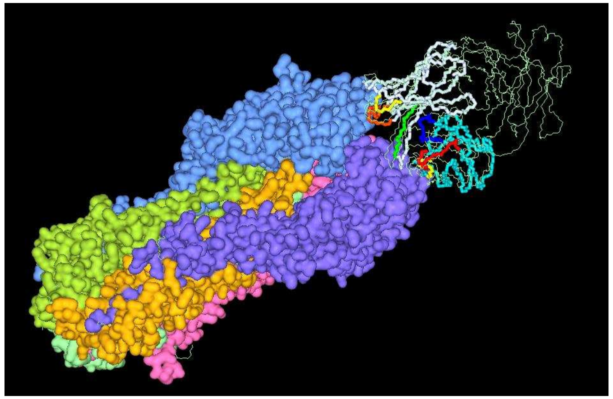
Predictor	Software	Algorithm	T1	T2	T3	T4	T5	T6	T7
Abagyan	ICM	FF			**			***	**
Camacho	CHARMM	FF	*					***	***
Eisenstein	MolFit	FFT	*	*					***
Sternberg	FTDOCK	FFT		*				**	*
Ten Eyck	DOT	FFT	*	*				**	
Gray		MC						**	***
Ritchie	Hex	SPF			**			***	
Weng	ZDOCK	FFT		**					**
Wolfson	BUDDA/PPD	GH	*						***
Bates	Guided Docking	FF	-	-	-				***
Palma	BIGGER	GF	-		-			**	*
Gardiner	GAPDOCK	GA	*	*	-	-	-	-	-
Olson	Surfdock	SH	*			-	-	-	-
Valencia		ANN	*	-	-	-	-	-	-
Vakser	GRAMM	FFT		*		-	-	-	-

$*$ **low,** $**$ **medium,** $* * *$ **high accuracy prediction;** — **no prediction**

Mendez et al. (2003) Proteins Struct. Funct. Bionf. 52 51–67

Hex Protein Docking Example – CAPRI Target 3

- Example: best prediction for CAPRI Target 3 – Hemagglutinin/HC63



Ritchie and Kemp (2000), Proteins Struct. Funct. Bionf. 39 178–194

Ritchie (2003), Proteins Struct. Funct. Genet. 52 98–106

CAPRI Results: Targets 8–19 (2003 – 2005)

Predictor	Software	T8	T9	T10	T11	T12	T13	T14	T15–T17	T18	T19
Abagyan	ICM	**		*	**	***	*	***		**	**
Wolfson	PatchDock	**	*	*	*	*	-	**		**	*
Weng	ZDOCK/RDOCK	**			*	***	***	***		**	**
Bates	FTDOCK	*		*	**	*		**		**	*
Baker	RosettaDock	-			**	***	**	***			***
Camacho	SmoothDock	**				***	***	**		**	*
Gray	RosettaDock	***	-	-	**	***					**
Bonvin	Haddock	-	-	**	**		***	***			
Comeau	ClusPro	**				***	*				*
Sternberg	3D-DOCK	**			*	*		**			*
Eisenstein	MolFit	***			*	***		**			
Ritchie	Hex				**	***	*	*			
Zhou		-	-		-	***	**	*		*	
Ten Eyck	DOT					***	***	**			
Zacharias	ATTRACT	**		-	-	-	-	***			**
Valencia		*			*	*	-				-
Vakser	GRAMM	-	-		-	-	-	**		**	
Homology	modelling				#			#			#
Cancelled									#		

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

High Order FFTs, Multi-Threading, and Graphics Processors

- Spherical polar coordinates give an analytic formula for 6D correlations:

In particular:

$$S_{AB} = \sum_{jsmlvrt} \Lambda^{rm}_{js} T^{(|m|)}_{js,lv}(R) \Lambda^{tm}_{lv} e^{-i(r\beta_A - s\gamma_A + m\alpha_B + t\beta_B + v\gamma_B)}$$

- This allows high order FFTs to be used – 1D, 3D, and 5D
- ... multiple FFTs can easily be executed in parallel
- ... also, it is relatively easy to implement on modern GPUs

- Up to 512 arithmetic "cores"
- Up to 6 Gb memory
- Easy API with C++ syntax
- Grid of threads model ("SIMT")

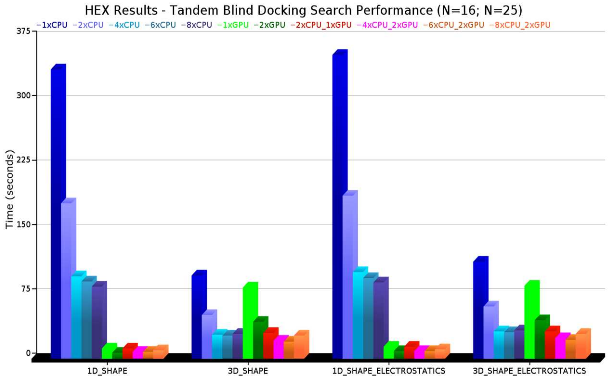
- Due to memory latency effects, 1D FFTs are MUCH FASTER than 3D FFTs ...

Ritchie, Kozakov, Vajda (2008), Bioinformatics 24 1865–1873

Ritchie, Venkatraman (2010), Bioinformatics, 26, 2398–2405

Protein Docking Speed-Up using Multiple GPUs and CPUs

- With multi-threading, we can use as many GPUs and CPUs as are available



- For best performance: use 2 GPUs alone, or 6 CPUs plus 2 GPUs
- With 2 GPUs, docking takes about 10 seconds – very important for large-scale!

Speed Comparison with ZDOCK and PIPER

- Hex: 52000 x 812 rotations, 50 translations (0.8Å steps)
- ZDOCK: 54000 x 6 deg rotations, 92Å 3D grid (1.2Å cells)
- PIPER: 54000 x 6 deg rotations, 128Å 3D grid (1.0Å cells)
- Hardware: GTX 285 (240 cores, 1.48 GHz)

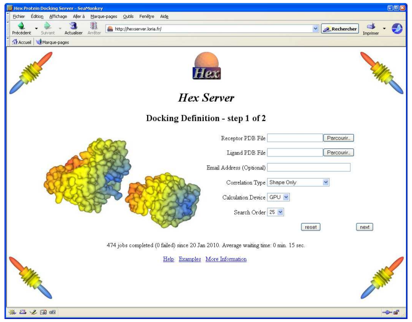
	Kallikrein A / BPTI (233 / 58 residues)#					
	ZDOCK	PIPER†	PIPER†	Hex	Hex	Hex‡
FFT	1xCPU	1xCPU	1xGPU	1xCPU	4xCPU	1xGPU
3D	7,172	468,625	26,372	224	60	84
(3D)*	(1,195)	(42,602)	(2,398)	224	60	84
1D	–	–	–	676	243	15

# execution times in seconds
* (times scaled to two-term potential, as in Hex)

- What's next?
  - Better energy functions & constraints...
  - Modeling flexibility...
  - Using homology templates...
  - Multi-component complexes...

"Hex" and "HexServer"

- Multi-threaded Hex: first (only) docking program to get full benefit of GPUs


Hex Protein Docking Server - SeaMonkey
Fichier Édition Affichage Aller à Marque-pages Outils Fenêtre Aide
Précédent Suivant Actualiser Arrêter
http://hexserver.loria.fr/
Rechercher
Imprimer
Accueil Marque-pages
Hex

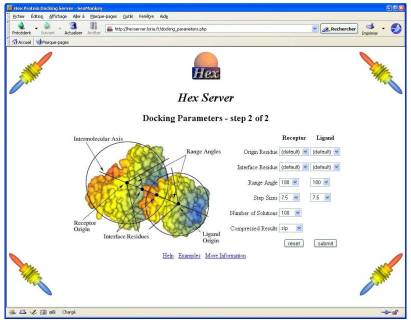
Hex Server

Docking Definition - step 1 of 2

Receptor PDB File [Parcourir..]
Ligand PDB File [Parcourir..]
Email Address (Optional)
Correlation Type: Shape Only
Calculation Device: GPU
Search Order: 25
reset next

474 jobs completed (0 failed) since 20 Jan 2010. Average waiting time: 0 min. 15 sec.

Help Examples More Information


Hex Protein Docking Server - SeaMonkey
Fichier Édition Affichage Aller à Marque-pages Outils Fenêtre Aide
Précédent Suivant Actualiser Arrêter
http://hexserver.loria.fr/docking_parameters.php
Rechercher
Imprimer
Accueil Marque-pages
Hex
Intermolecular Axis
Range Angles
Receptor Origin
Interface Residues
Ligand Origin
Chargé

Hex Server

Docking Parameters - step 2 of 2

	Receptor	Ligand
Origin Residue	(default)	(default)
Interface Residue	(default)	(default)
Range Angle	180	180
Step Sizes	7.5	7.5
Number of Solutions	100	
Compressed Results	zip	

reset submit

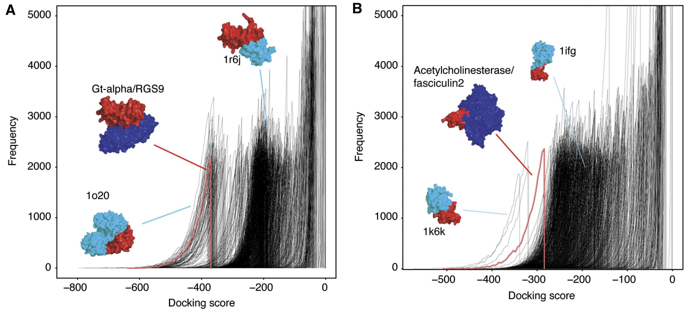
Help Examples More Information

- Hex: Over 25,000 down-loads...
- HexServer: About 1,000 docking jobs per month...

Ritchie and Kemp (2000) Proteins, 39, 178–194
...
Ritchie and Venkatraman (2010) Bioinformatics, 26, 2398–2405
Macindoe et al. (2010), Nucleic Acids Research, 38, W445–W449

Can Cross-Docking Distinguish The Correct PPI Partners?

- Wass et al. used Hex to cross-dock 56 true protein pairs with 922 non-redundant "decoys"
  - For each pair, they plotted the profile of the best 20,000 docking scores...



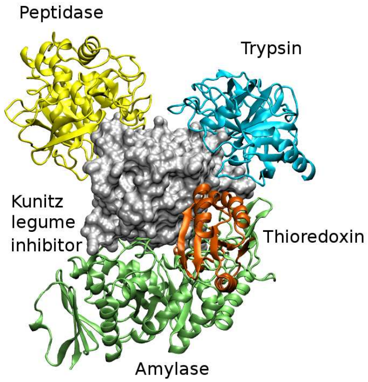
(negative scores are good; red/blue = correct PPI; red/cyan = incorrect interactions)

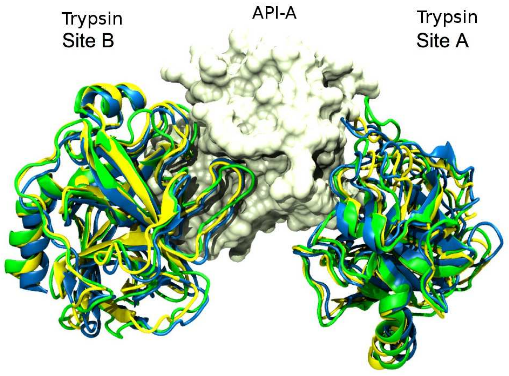
- 48/56 true PPIs have significantly (statistically) higher energies than background false pairs
- Only 8/56 true PPIs have indistinguishable profiles to the non-binders
- NB. this experiment is detecting energy funnels, not necessarily the correct docking pose

Wass et al. (2011) Mol Sys Biol 7, article 469

Knowledge-Based Protein Docking: CAPRI Target 40 (2009) – API-A/Trypsin

- We searched SCOPPI and 3DID for similar domain interactions to the target
- This helped to identify two key inhibitory loops on API-A around L87 and K145



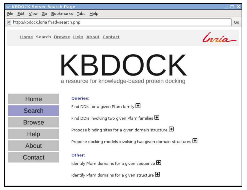


- Performing focused Hex + MD refinement gave a total of 9 "acceptable" solutions

The KBDOCK Database and Web Server

- Content: 2,721 non-redundant hetero DDIs involving 1,029 PFAM domain families
- For each PFAM family, all DDIs are superposed and spatially clustered

http://kbdock.loria.fr/



KBDOCK Server Search Page

File Edit View Go Bookmarks Tabs Help
Go
Home Search Browse Help About Contact
Inria
Home
Search
Browse
Help
About
Contact

http://kbdock.loria.fr/advsearch.php

KBDOCK

a resource for knowledge-based protein docking

Queries:

Find DDIs for a given Pfam family ⊞

Find DDIs involving two given Pfam families ⊞

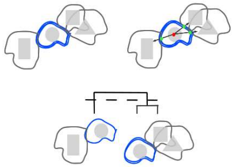
Propose binding sites for a given domain structure ⊞

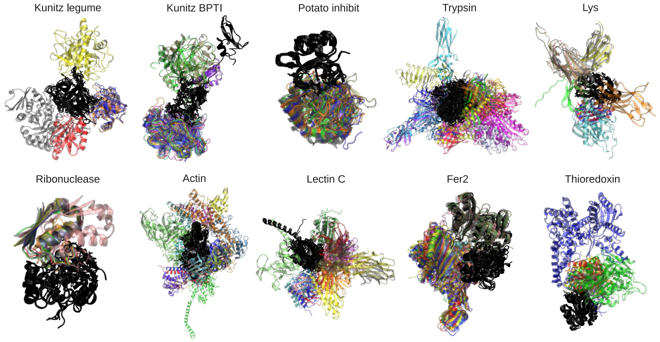
Propose docking models involving two given domain structures ⊞

Other:

Identify Pfam domains for a given sequence ⊞

Identify Pfam domains for a given structure ⊞

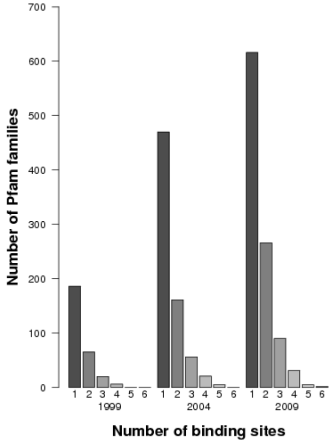


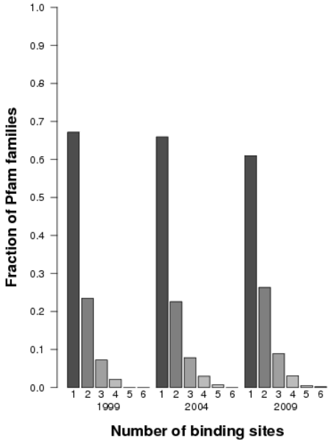


- Aim: to provide PFAM family-level structural templates for knowledge-based docking

KBDOCK – Analysis of PFAM Domain Family Binding Sites

- Nearly 70% of PFAM domain families have just one binding site
- Very few domains have more than two or three binding sites





- This supports the notion that protein binding sites are often re-used...

KBDOCK – Template-Based Protein Docking Results

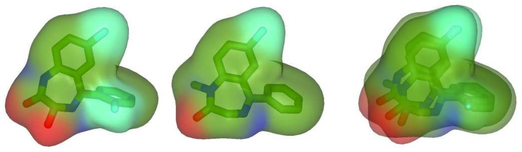
- The Protein Docking Benchmark 4.0 contains 176 protein-protein complexes
- We selected 73 single-domain complexes
- A "Full-Homology" (FH) template matches both target domains
- A "Semi-Homology" (SH) template matches just one target domain

Target class	Total targets	FH templates	Two SH templates	One SH template	Zero templates
Without date filtering					
Enzyme	36	24 / 24	(3 + 1) / 5	3 / 5	2
Other	37	21 / 21	(0 + 0) / 3	5 / 11	2
With date filtering					
Enzyme	36	13 / 13	(2 + 1) / 5	7 / 11	7
Other	37	13 / 13	(0 + 0) / 1	8 / 15	8

- If a FH template exists, it is almost always correct
- Even if there is no FH template, SH templates can still provide useful information

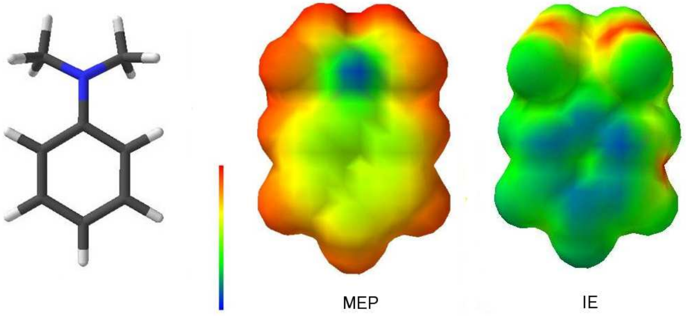
Ghoorah et al. (2011), Bioinformatics, 27, 2820–2827

But What About the Virtual Screening ?



ParaSurf – SH Surfaces & Properties from Semi-Empirical QM

- From MOPAC or VAMP calculate:
  - Density contours of $2 \times 10^{-4}\mathrm{e}/\text{Å}^3$ ($\sim$ SAS)
  - Key local properties: MEP, $\mathrm{IE_L}$, $\mathrm{EA_L}$, α_L
- Encode as SH expansions to L=15: $f(\theta,\phi) = \sum_{l=0}^{L}\sum_{m=-l}^{l} f_{lm} y_{lm}(\theta,\phi)$



Lin & Clark (2005) J Chem Inf Model, 45, 1010–1016; Clark (2004) J Mol Graph 22 519–525

ParaFit – High Throughput SH Surface & Property Matching

Distance: $D = \int (r_A(\theta,\phi) - r_B(\theta,\phi)')^2 \mathrm{d}\Omega$ (in units of area)

Orthogonality: $D = |\underline{a}|^2 + |\underline{b}|^2 - 2\underline{a}.\underline{b}'$

Rotation: $b'_{lm} = \sum_{m'} R^{(l)}_{mm'}(\alpha,\beta,\gamma) b_{lm'}$

Hodgkin: $S = 2\underline{a}.\underline{b}'/(|\underline{a}|^2 + |\underline{b}|^2)$

Carbo: $S = \underline{a}.\underline{b}'/(|\underline{a}|.|\underline{b}|)$

Tanimoto: $S = \underline{a}.\underline{b}'/(|\underline{a}|^2 + |\underline{b}|^2 - \underline{a}.\underline{b}')$

Multi-property: $S = pS^{\mathrm{shape}} + qS^{\mathrm{MEP}} + rS^{\mathrm{IE_L}} + sS^{\mathrm{EA_L}} + tS^{\alpha_\mathrm{L}}$

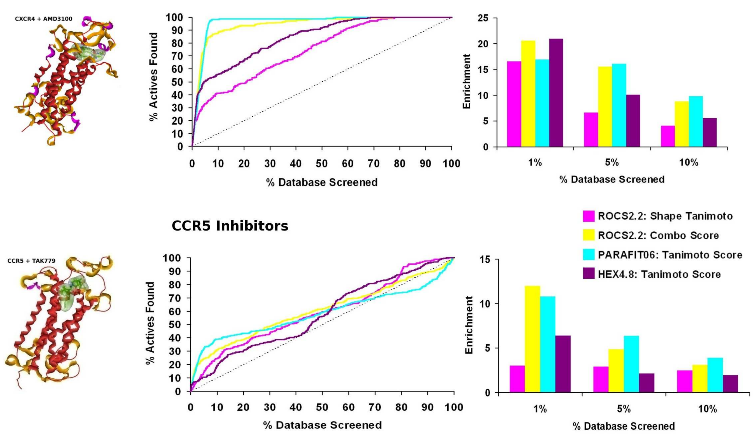
Perez-Nueno et al. (2010), Mol Inf, 30, 151–159

SH-Based Virtual Screening of HIV Entry Inhibitors

- Database of 248 CXCR4 and 354 CCR5 inhibitors + 4696 decoys
- Performed SH-based VS to distinguish actives from decoys...

CXCR4 Inhibitors

CCR5 Inhibitors



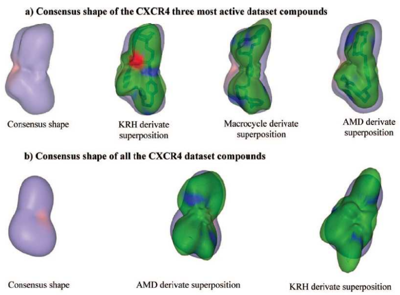
(for CXCR4, query = AMD3100; for CCR5, query = TAK779)

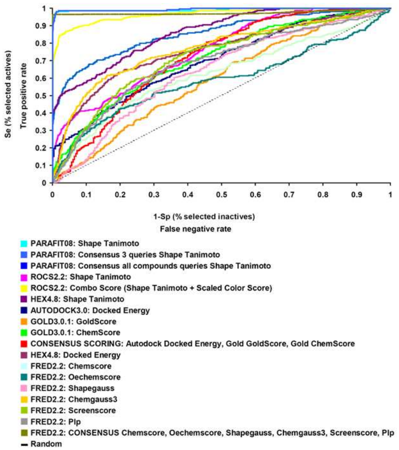
Pérez-Nueno et al. (2008) J Chem Inf Model 48, 509–533

SH Consensus Shapes Can Improve VS Screening Performance

- The Consensus shape is the "average" of a group of shapes...

$$\tilde{r}(\theta,\phi) = \frac{1}{N}\sum_{k=1}^{N}\sum_{lm} a^k_{lm} y_{lm}(\theta,\phi)$$


a) Consensus shape of the CXCR4 three most active dataset compounds
Consensus shape
KRH derivate superposition
Macrocycle derivate superposition
AMD derivate superposition
b) Consensus shape of all the CXCR4 dataset compounds
Consensus shape
AMD derivate superposition
KRH derivate superposition

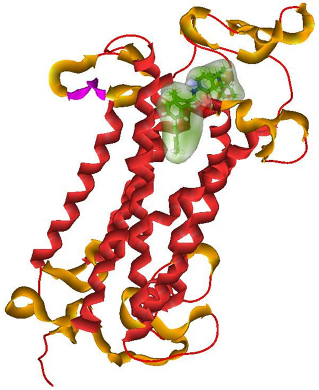


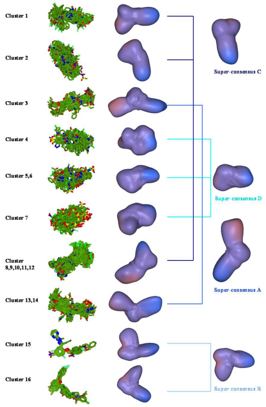
- For CXCR4, using the consensus of top 3 actives gives best overall VS performance

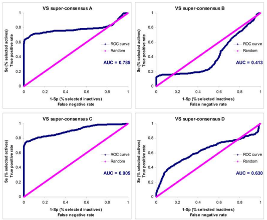
Pérez-Nueno et al. (2008) J Chem Inf Model 48, 509–533

Clustering and Classifiying Diverse HIV Entry Inhibitors

- We clustered the 354 known inhibitors for CCR5





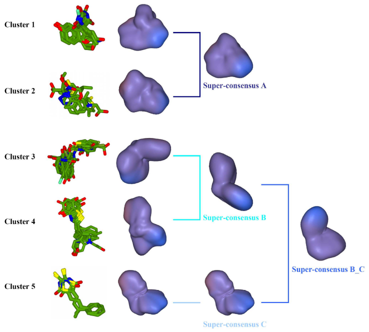


- We classified the inhibitors into four main clusters; merging clusters worsens the AUCs
- Therefore, the CCR5 ligands form no less than FOUR main groups
- Docking with Hex indicates these groups bind within THREE sub-sites in the CCR5 pocket

Pérez-Nueno, Ritchie, et al., (2008) J Chem Inf Model 48(11) 2146-2165

Promiscuous Protein Targets Seem to be Rather Common

- Example: ALR2 is know to bind at least 5 different ligand scaffold families...



- Several other promiscuous targets in the literature:
  - the $\alpha 1\beta 1$ and $\alpha 2\beta 1$ integrins,
  - factor H, LRP6, PPAR-γ, LXR-β,
  - ACHE, P38, FXA, VEGFR2, PXR,
  - β-secretase, thrombin, CDK2,
  - LAIR-1, LAIR-2, LTBLP-2, NS2B-NS3.

- For ligand-based virtual screening, these examples suggest:
  - cluster the 3D shapes of any known ligands before performing VS ...
  - compare shape-based VS performance with and without clustering ...
  - ... any large differences could suggest a promiscuous (multi-site?) substrate.

Pérez-Nueno, Ritchie (2011). Expert Opinion on Drug Discovery, 7, 1–17.

Conclusions and Future Prospects

- Polar Fourier representations are useful for protein docking and VS
- Rigid-body protein docking on a GPU now takes only a few seconds
- Knowledge-based protein docking is becoming increasingly useful
- Most Pfam families have just one binding site – often re-used
- Several proteins bind multiple ligand families – promiscuous targets
- SH consensus shape queries can improve and explain VS performance
- GPU-based correlation techniques could open several possibilities:
  - All-vs-all protein docking and ligand shape-matching ?

Acknowledgments

Anisah Ghoorah
Matthieu Chavent
Lazaros Mavridis
Violeta Pérez-Nueno
Vishwesh Venkatraman

BBSRC, EPSRC, ANR

Software & Papers: http://hex.loria.fr/

HexServer: http://hexserver.loria.fr/