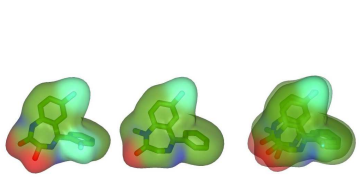
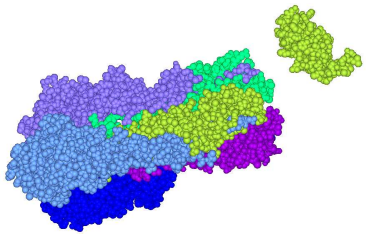
High Performance Algorithms for Molecular Shape Recognition

David Ritchie





Habilitation Defence – A08, LORIA – 14:00, 5th April 2011

Rapporteurs	Gilles Bernot, professeur, Université Nice Sophia Antipolis
	Frederic Cazals, DR, INRIA Sophia Antipolis – Méditerranée
	Alexandre Varnek, professeur, Université de Strasbourg
Examinateurs	Bernard Girau, professeur, Université Henri Poincaré
	Bruno Lévy, DR, INRIA Nancy – Grand Est
	Paul Zimmermann, DR, INRIA Nancy – Grand Est

Acknowledgments – PhD Students and Postdocs

Aberdeen

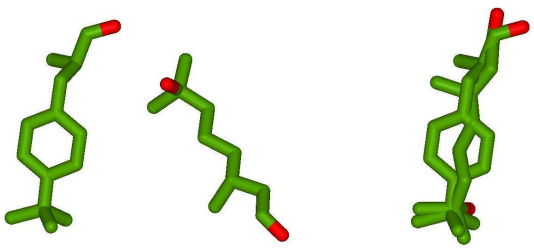
Diana Mustard
Alessandra Fano
Lazaros Mavridis
Violeta Pérez-Nueno
Antonis Koussounadis

Nancy

Anisah Ghoorah
Lazaros Mavridis
Violeta Pérez-Nueno
Vishwesh Venkatraman
Thomas Bourquard

The Problem of Molecular Shape Recognition

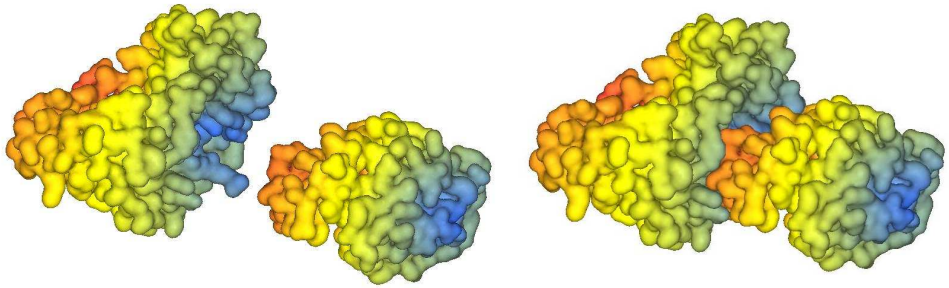
- Are these molecules similar?



- First, superpose them.
- Next, apply a similarity scoring function... Aah – Muguet!
- But how to superpose molecules and calculate similarity automatically?

Protein Docking – Another Molecular Recognition Problem

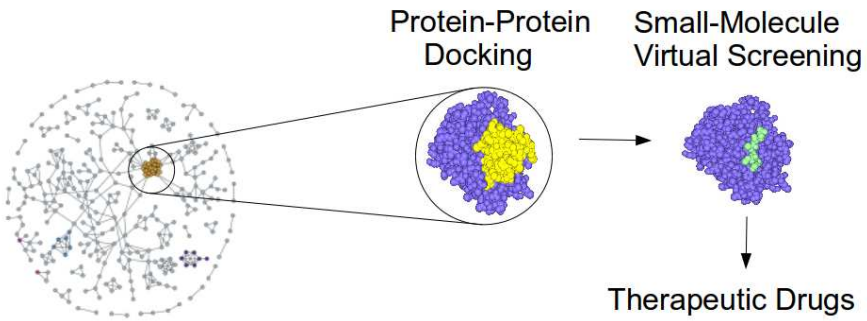
- A six-dimensional puzzle – do these proteins fit together?



- Yes, they fit!
- It is mostly a rotational problem: ONE translation plus FIVE rotations...
- But proteins are flexible => multi-dimensional space!
- So, how to calculate whether two proteins recognise each other?

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



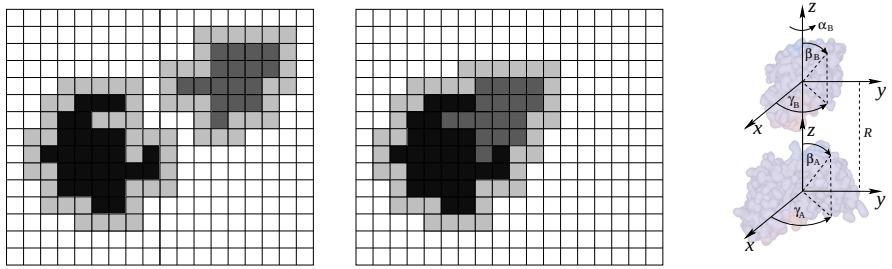
- Understanding PPIs could lead to immense scientific advances
- Small "drug" molecules often inhibit or interfere with PPIs

What Are High Performance Algorithms?

- Fast Fourier Transforms (FFTs) ...
- Principle Component Analyses (PCAs) ...
- So what's new ?
  - Treat docking and shape matching as rotational problems
  - Spherical Polar Fourier (SPF) correlations
  - SPF approach leads to high order 5D FFTs
  - Mapping docking calculations to GPUs
  - Coupling SPF and Knowledge-Based techniques

Protein Docking Using FFTs (The Old Way!)

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



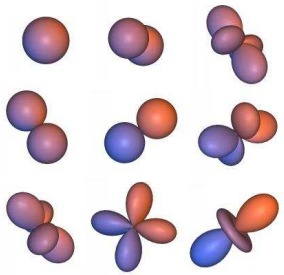
- ...and use FFTs to calculate translational correlations:

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

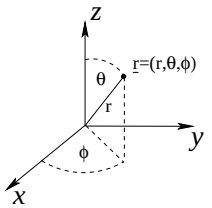
- BUT – have to rotate one protein and repeat, which is expensive!
- POLAR coords allow the rotational nature of problem to be exploited

The Spherical Harmonics

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



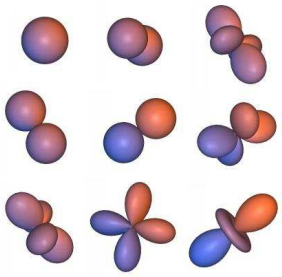
- Spherical polar coordinates: $\underline{r} = (r, \theta, \phi)$



- 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: $\int y_{lm}y_{kj}\mathrm{d}\Omega = \int Y_{lm}Y_{kj}\mathrm{d}\Omega = \delta_{lk}\delta_{mj}$
  - Complex↔Real: $e^{im\phi} = \cos m\phi + i\sin m\phi$

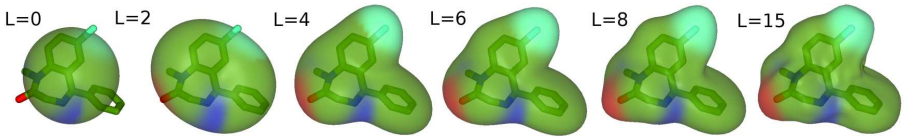
Spherical Harmonic Molecular Surfaces

- 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

Some Theory – Addition Theorems and Rotations

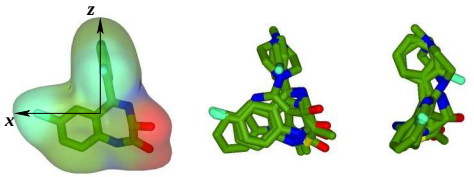
- An addition theorem is a relation between $f(a+b)$ and $f(a)$ and $f(b)$ ...
  - Example: $e^{i(\alpha+\phi)} = e^{i\alpha} \times e^{i\phi}$
- Addition theorems are useful for shifting coordinate systems:
  - e.g. z-rotation: $Y_{lm}(\theta,\phi+\alpha) = e^{-im\alpha} Y_{lm}(\theta,\phi)$
- Calculating a general 3D rotation (3 Euler angles) is thanks to Wigner
  - Rotated SHs: $Y_{lm}(\theta',\phi') = \sum_{m'} D^{(l)}_{m'm}(\alpha,\beta,\gamma) Y_{lm'}(\theta,\phi)$
- Here, we wish to fix the coordinate system and rotate the objects (molecules)
  - "Object": $r(\theta,\phi) = \sum_{lm} A_{lm} Y_{lm}(\theta,\phi)$
  - Rotated "Object": $r(\theta,\phi)' = \sum_{lm}[\sum_{m'} D^{(l)}_{mm'}(\alpha,\beta,\gamma) A_{lm'}] Y_{lm}(\theta,\phi)$

FFT-Based Surface Shape Matching

- For multiple rotational samples: $e^{i\alpha} \implies FFT(\alpha)$
- 3D FFTs are possible: $D^{(l)}_{mm'}(\alpha,\beta,\gamma) = \sum_t \Gamma^{(l)}_{mtm'} \times e^{-im\alpha} e^{-it\beta} e^{-im'\gamma}$
- Vector Interpretation: $\{a_{lm}\ ;\ |m| \le l \le L\} \rightarrow \underline{a}$
- Distance Interpretation: $D = \int \left(r_A(\theta,\phi) - r_B(\theta,\phi)\right)^2 \mathrm{d}\Omega = |\underline{a}|^2 + |\underline{b}|^2 - 2\underline{a}.\underline{b}$
- Overlap Interpretation: $S = \int r_A(\theta,\phi) r_B(\theta,\phi) \mathrm{d}\Omega = \underline{a}.\underline{b}$
- Carbo Similarity: $S = \underline{a}.\underline{b}/(|\underline{a}|.|\underline{b}|)$
- Use icosahedral sampling and 1D or 3D FFTs for very fast rotational superpositions:

Can We Avoid Performing Rotational Comparisons?

- Rotation-invariant descriptors:
  - RI coefficients: $A_l = \sqrt{\sum_m a_{lm}^2}$ and $A_L = \sqrt{\sum_l A_l^2}$
  - RI "distance": $D_{RI} = A_L^2 + B_L^2 - 2\sum_{l=0}^{L} A_l B_l$
- Canonical orientations:
  - First, align principal radii to the axes using L=6
  - Then, compare using Carbo: $S = \underline{a}.\underline{b}/(|\underline{a}|.|\underline{b}|)$

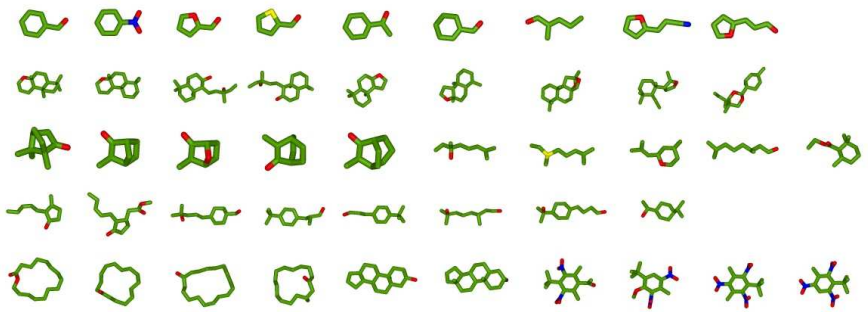


- We find that canonical shape comparison is much better than rotation-invariant
  - So, for a large database, store all molecules in canonical orientations...

Mavridis, Hudson, Ritchie (2007), J Chem Inf Model 45(5) 1787-1796

Clustering the Odour Dataset

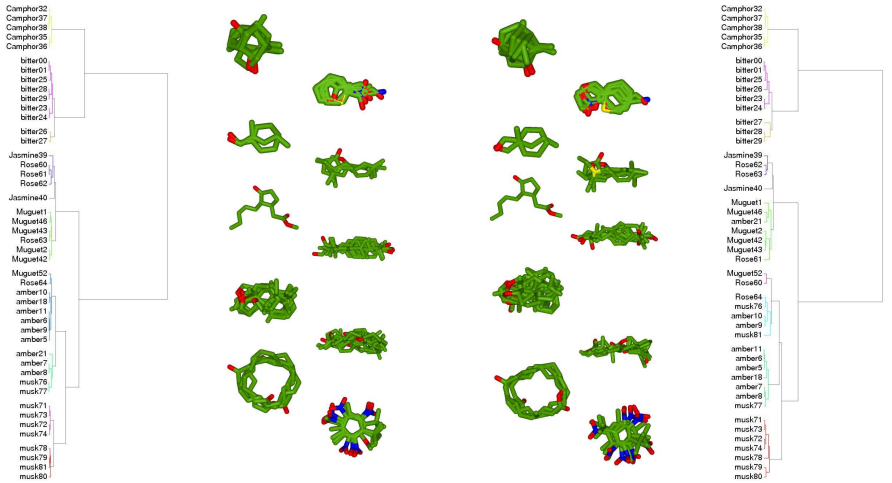
- 7 classes: bitter, ambergris, camphoraceous, rose, jasmine, muguet, musk
  - Takane et al. (2004) Org Biomol Chem 2 3250–3255



- Following Takene et al., the 46 molecules were clustered into 10 groups...
  - (Takene et al. originally clustered them on quantum mechanics vibrational frequencies)

Odour Dataset Clustering Results

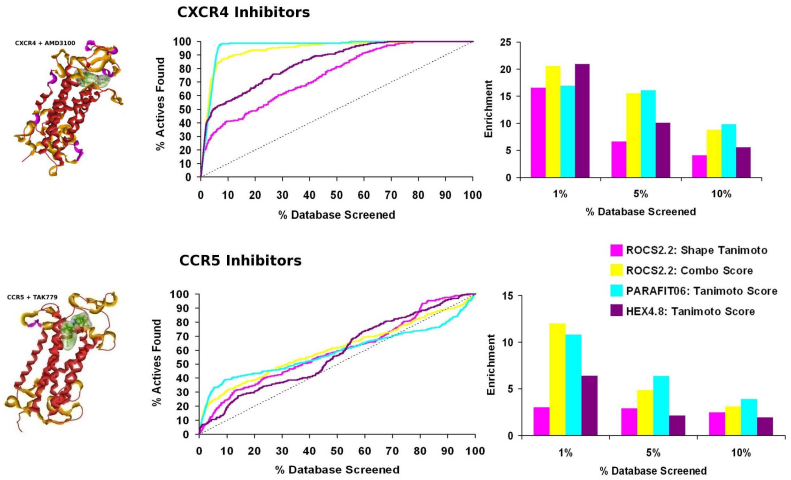
Clustering Superposed Pairs | **Clustering Canonical Orientations**



Mavridis, Hudson, Ritchie (2007), J Chem Inf Model 45(5) 1787-1796

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...

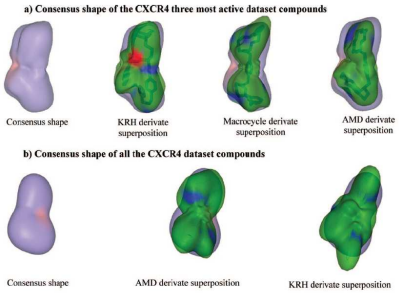


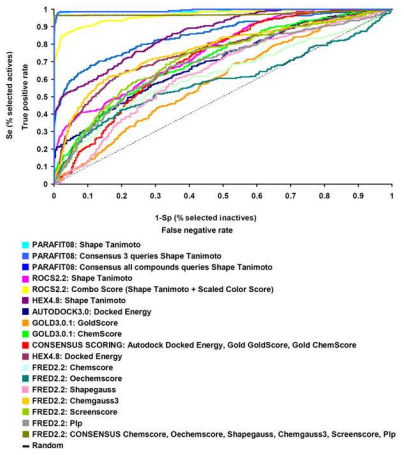
Pérez-Nueno et al. (2008) J Chem Inf Model 48, 509–533.

SH Consensus Shapes Can Improve VS Screening Performance

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

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

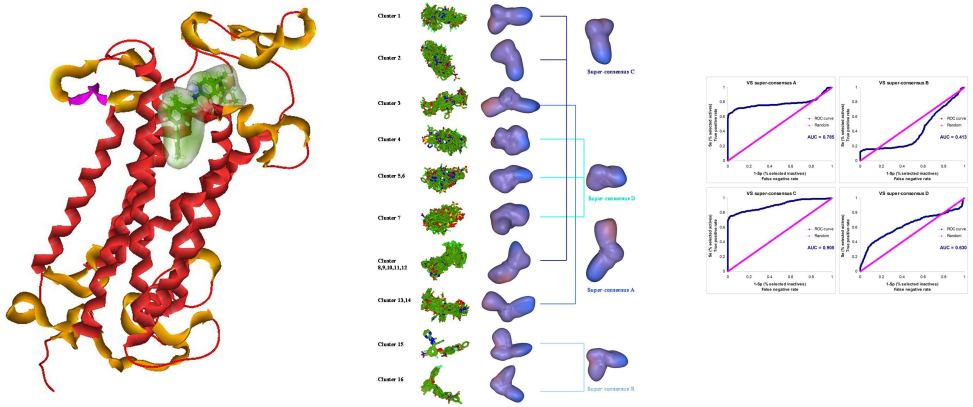




Pérez-Nueno et al. (2008) J Chem Inf Model 48, 509–533.

Clustering and Classifiying Diverse HIV Entry Inhibitors

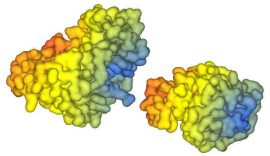
- We clustered the 354 known inhibitors for CCR5



- 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

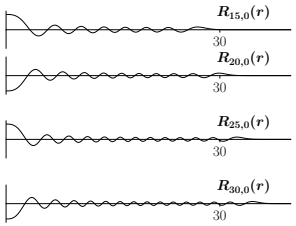
Pérez-Nueno, Ritchie, et al., (2008) J Chem Inf Model 48(11) 2146-2165

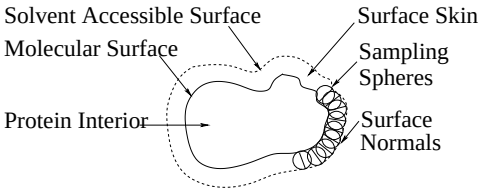
But What About the Docking Problem?



Docking Needs a 3D "Spherical Polar Fourier" Representation

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

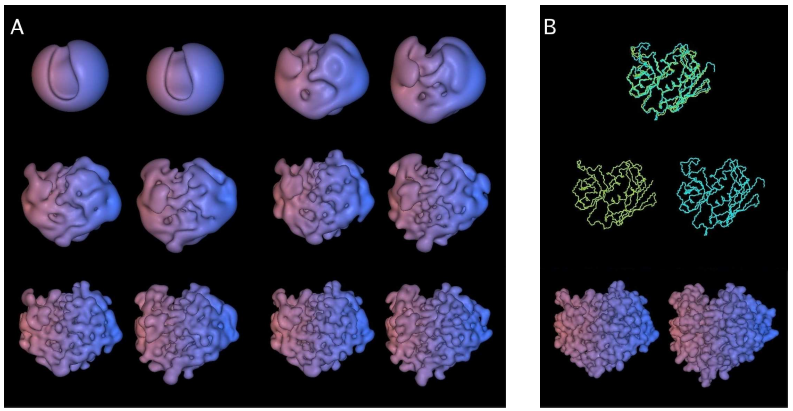




- Surface Skin: $\sigma(\underline{r}) = \begin{cases} 1; & \underline{r} \in \text{surface skin} \\ 0; & \text{otherwise} \end{cases}$ Interior: $\tau(\underline{r}) = \begin{cases} 1; & \underline{r} \in \text{protein at} \\ 0; & \text{otherwise} \end{cases}$
- Parametrise as: $\sigma(\underline{r}) = \sum_{n=1}^{N} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} a_{nlm}^{\sigma} R_{nl}(r)\, y_{lm}(\theta, \phi)$
- Translations: $a_{nlm}^{\sigma}(R) = \sum_{n'l'}^{N} T_{nl,n'l'}^{(|m|)}(R) a_{n'l'm}^{\sigma}$

SPF Protein Shape-Density Reconstruction and Superposition

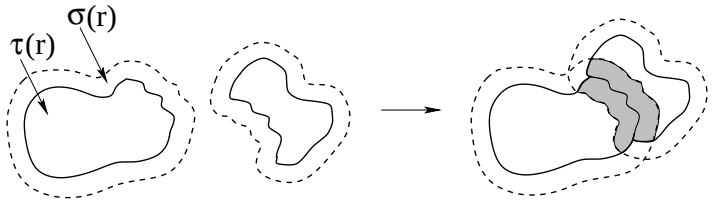
Shape-density: $\tau(\underline{r}) = \sum_{nlm}^{N} a_{nlm}^{\tau} R_{nl}(r) y_{lm}(\theta, \phi)$



- Similar proteins may be superposed using only low resolution expansions (N=6), top left

Ritchie (2003) Proteins, 52 98–106

Protein Docking Using SPF Density Functions (The New Way!)



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

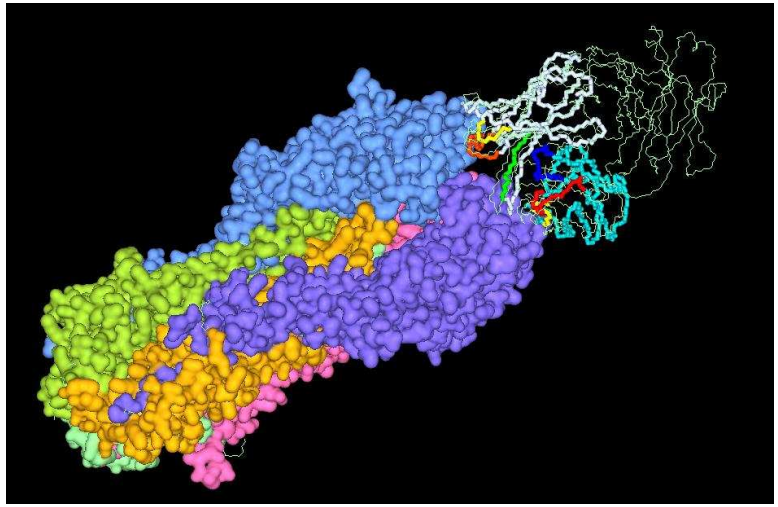
Score: $S_{AB} = \int (\sigma_A\tau_B + \tau_A\sigma_B - Q\tau_A\tau_B)\mathrm{d}V$ **Penalty Factor:** $Q = 11$

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: $(R, \beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B)$

Ritchie and Kemp (2000) Proteins, 39, 178–194

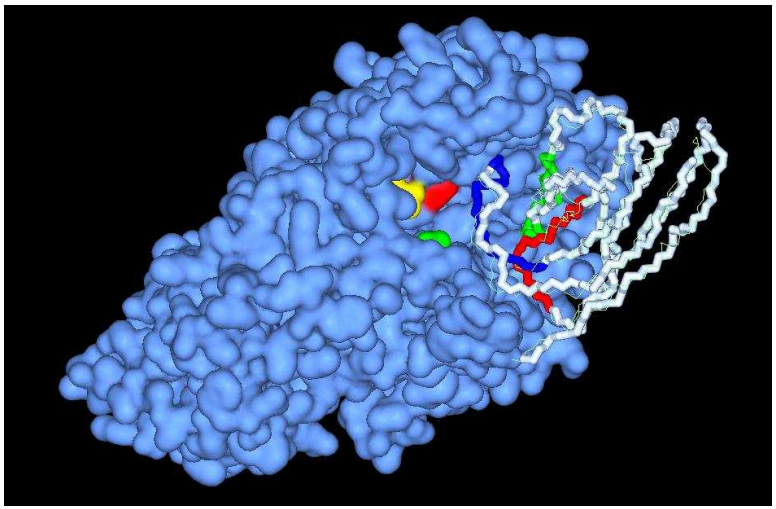
Docked Orientation for CAPRI Target 3 – Hemagglutinin/HC63



- CAPRI "medium accuracy" (1Å $\leq$ Ligand RMSD $\leq$ 5Å)

Ritchie (2003) Proteins, 52, 98–106.

Docked Orientation for CAPRI Target 6 – Amylase/AMD9

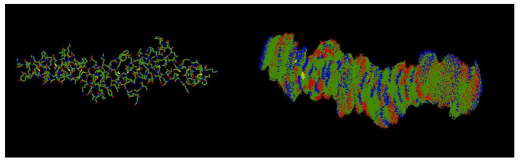


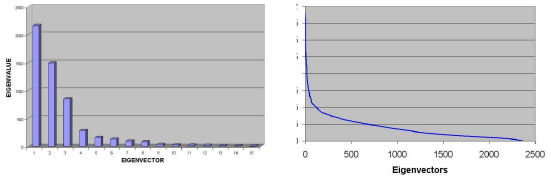
- CAPRI "high accuracy" (Ligand RMSD $\leq$ 1Å)

Ritchie (2003) Proteins, 52, 98–106.

Simulating Flexibility During Docking using "Essential Dynamics"

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





- Covariance matrix, C:

$$C_{ij} = < (x_i - \overline{x}_i)(x_j - \overline{x}_j) >$$

- Calculate eigenvectors, E:

$$\underline{C} = \underline{E}.\underline{\Lambda}.\underline{E}^T$$

- Estimate Unbound to Bound:

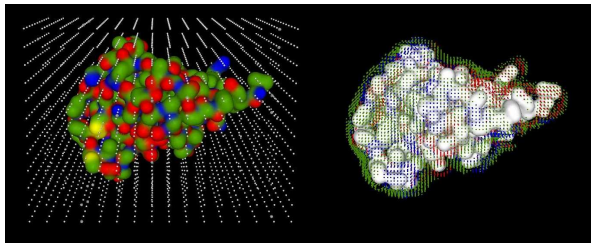
$$\underline{B} \simeq \underline{U} + \sum_{k=1}^{n} \alpha_k \underline{e}_k$$

- The first few eigenvectors encode most of the internal fluctuations
- We were the first to show that this could improve rigid body docking...

Mustard and Ritchie (2005), Proteins 60, 269–274

Using PCA to Predict Chemical Complementarity

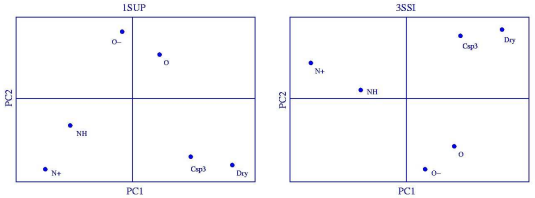
- We used "GRID" to calculate chemical potentials around proteins



Chemical probes
O, O–,
N, NH, N+,
Csp3, Dry

Colour codes
R (+), G (hyd), B (-)

- We then applied PCA to the potential grids

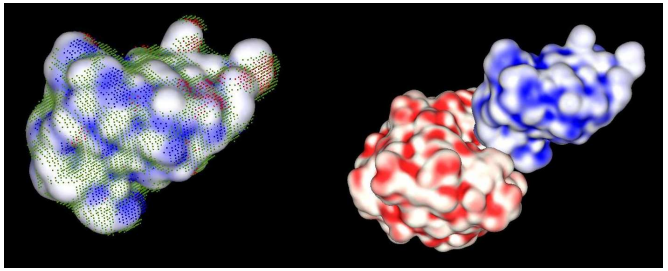


- This showed that N+, O–, and "Dry" explained 70–75% of the variance...

Fano et al. (2006) J Chem Inf Model 46, 1223–1235.

Protein Docking Using GRID Probe Potentials

- Docking the subtilisin/SSI-inhibitor using GRID probe potentials:



N+ = blue

O– = red

Dry = green

- We developed a probe-shape energy correlation:

$$E = \frac{1}{2}\int \left[\left(\phi_A^{\mathrm{N}+} + \phi_A^{\mathrm{O}-} + \phi_A^{\mathrm{Dry}}\right) * \tau_B + \left(\phi_B^{\mathrm{N}+} + \phi_B^{\mathrm{O}-} + \phi_B^{\mathrm{Dry}}\right) * \tau_A\right] \mathrm{d}V$$

- This gave better prediction (rank 5) than shape+elec (10) or shape (13)
- Promising, but not enough time to automate it all... To be revisited!

Fano et al. (2006) J Chem Inf Model 46, 1223–1235.

5D FFT Correlations from Complex Overlap Expressions

Complex SHs, Y_{lm}:
$$y_{lm}(\theta,\phi) = \sum_t U_{mt}^{(l)} Y_{lt}(\theta,\phi)$$

Complex coefficients:
$$A_{nlm} = \sum_t a_{nlt} U_{tm}^{(l)}$$

Complex overlap:
$$E = \sum_{kjsmnlv} D_{ms}^{(j)*}(0,\beta_A,\gamma_A) A_{kjs}^{*} T_{kj,nl}^{(|m|)}(R) D_{mv}^{(l)}(\alpha_B,\beta_B,\gamma_B) B_{nlv}$$

Collect coefficients:
$$S_{js,lv}^{(|m|)}(R) = \sum_{kn} A_{kjs}^{*} T_{kj,nl}^{(|m|)}(R) B_{nlv}$$

To give:
$$E = \sum_{jsmlv} D_{ms}^{(j)*}(0,\beta_A,\gamma_A) S_{js,lv}^{(|m|)}(R) D_{mv}^{(l)}(\alpha_B,\beta_B,\gamma_B)$$

And finally:
$$E = \sum_{jsmlvrt} \Gamma_{js}^{rm} S_{js,lv}^{(|m|)}(R) \Gamma_{lv}^{tm} e^{-i(r\beta_A - s\gamma_A + m\alpha_B + t\beta_B + v\gamma_B)}$$

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

nVidia Graphics Processors (GPUs)

- Modern GPUs have very high ($\sim$ teraflop) compute performance
- SIMT architecture = simultaneous instructions, multiple threads

- nVidia GPUs:
- Grid of threads model
- Uniform architecture/interface – "CUDA"
- 16–32 multi-processors
- 240–512 arithmetic "cores"
- 4–6 Gb main memory
- ONLY $\sim$ 16 Kb memory per multi-processor

- Need to aim for "high arithmetic intensity" on each multi-processor...
- Thankfully, matrix multiplications etc. fit these constraints perfectly

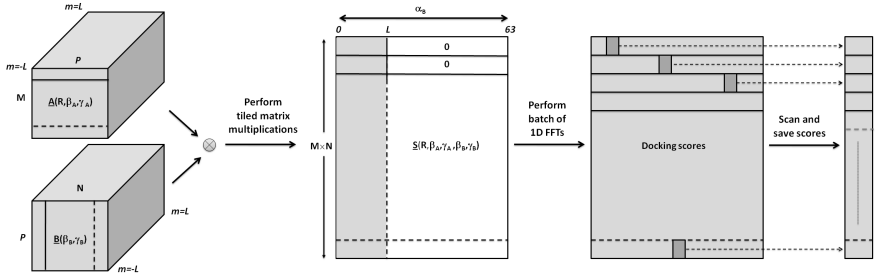
GPU Implementation – Perform Multiple FFTs

- Next, calculate multiple 1D FFTs of the form:

$$S_{AB}(\alpha_B) = \sum_m e^{-im\alpha_B} \sum_{nl} A^{\sigma}_{nlm}(R, \beta_A, \gamma_A) \times B^{\tau}_{nlm}(\beta_B, \gamma_B)$$

4. On GPU, cross-multiply transformed A with rotated B coefficients (as above)

5. On GPU, perform batch of 1D FFTs using cuFFT and save best orientations

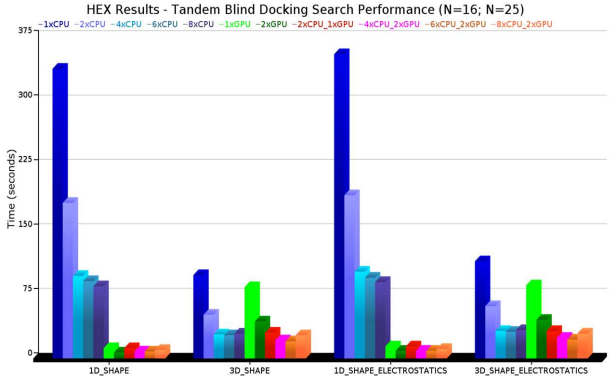


- 3D FFTs in $(\alpha_B, \beta_B, \gamma_B)$ can be calculated in a similar way...

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

Protein Docking on GPUs

- 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 only about 15 seconds – very important for large-scale!
- Overall, including set-up, Hex 1D FFT is about 45x faster on FX-5800 than on iCore7

Protein Docking – 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)

Protein Docking – 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)

- Next mission? – give Hex a better potential function!

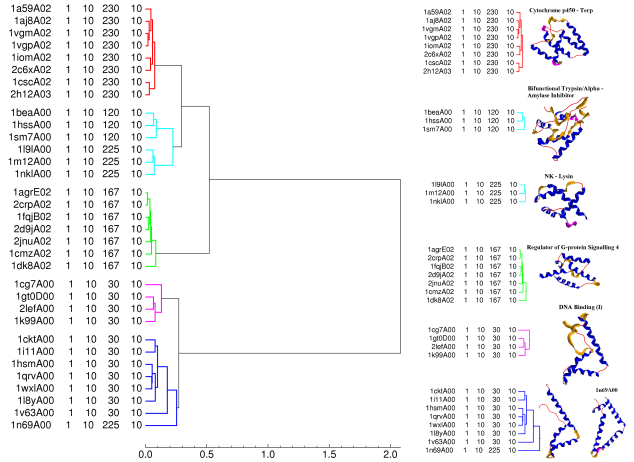
Current Work

and

Future Perspectives

Clustering CATH Protein Structure Superfamilies

- "CATH" is a "gold standard" classification of protein structures
  - Auto/expert curated: $\sim$ 12,000 structures, $\sim$ 1,200 folds
- Our first test – can we cluster the members of five selected families?

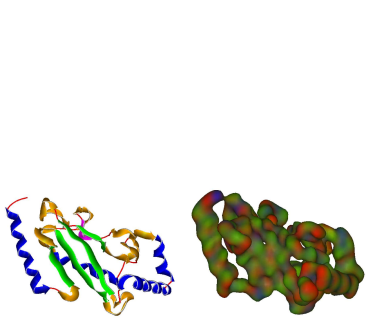


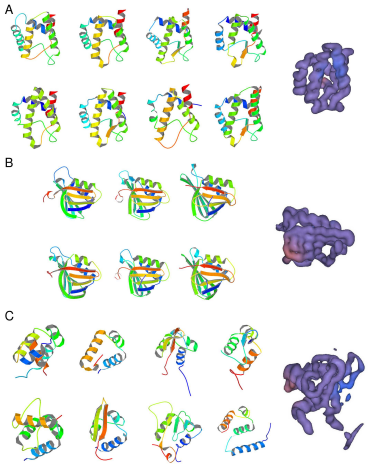
- Most structures are correctly grouped
- Global shape-density matching does not always agree with the expert "topology"
- We should consider shape-density as a database "view"?

Mavridis and Ritchie (2010), Pacific Symposium Biocomputing, 281–292.

3D-Blast – Comparing Protein Fold Family Consensus Shapes

- We can now also work with consensus protein backbone shapes:



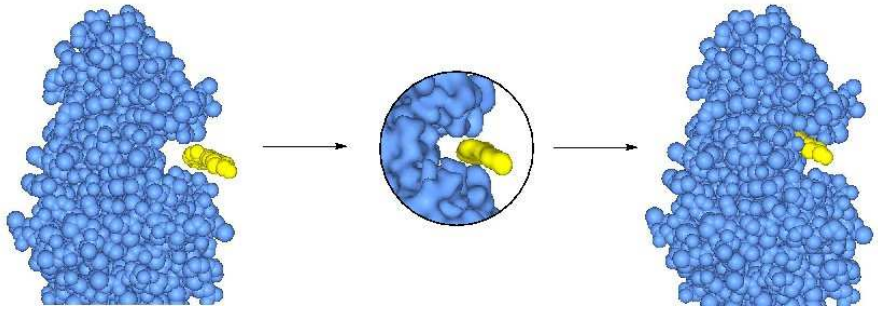


- This could provide a new way to index and search 3D structural databases...

Mavridis et al. (2011), manuscript submitted.

3D-Snap – Fast and Faithful 3D Virtual Drug Screening

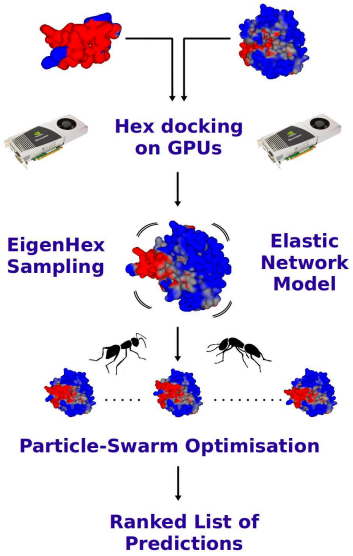
- 3D functions should give better VS performance than 2D SH surfaces...



- Ligand-ligand, ligand-pocket, pocket-pocket will all be possible...
- Consensus 3D shapes should work well too...
- I also want to explore new basis functions:
  - e.g. Gegenbauer polynomials (best for rotation + translation?)

EigenHex – Flexible Protein Docking

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



Overall approach

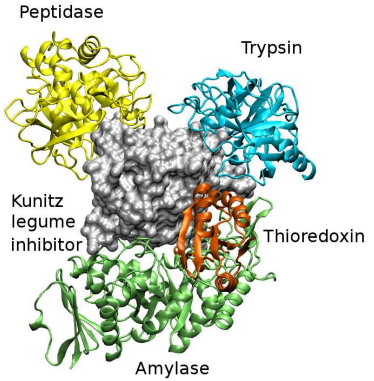
- Cα elastic network model (ENM)
- Use up to 20 eivenvectors
- Search using PSO
- Score using "DARS" potential

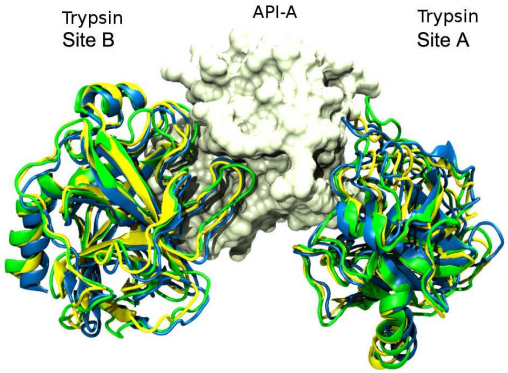
Results so far

- DARS works very well...
- Still need a better scoring function

Knowledge-Based Docking: CAPRI Target 40 – 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

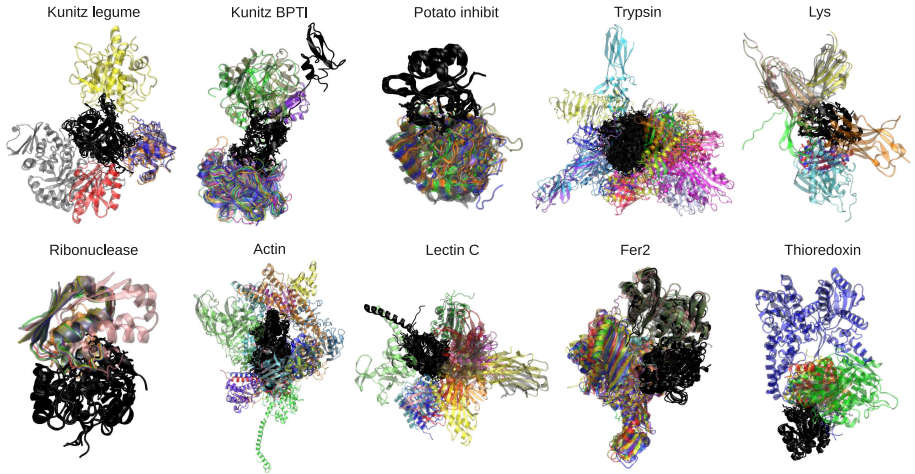




- Focused Hex docking + MD refinement gave NINE "acceptable" solutions in CAPRI

Using Known Protein Interfaces to Predict Unknown Interactions

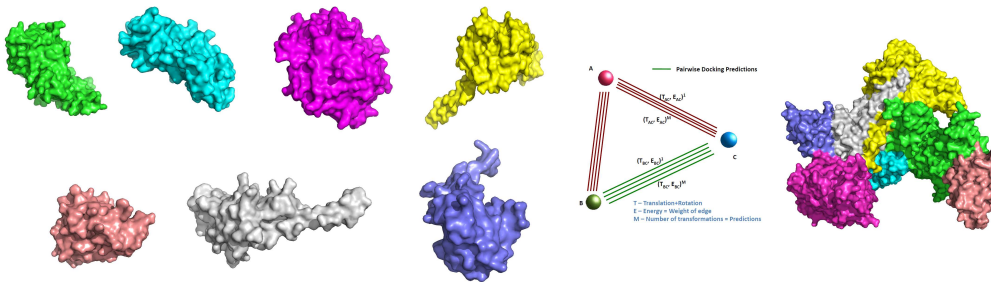
- KBDOCK – A PPI Database for Knowledge-Based Docking



Ghoorah et al. (2011), manuscript in preparation.

Assembling Multi-Component Protein Complexes

- Multi-component assembly is a highly combinatorial problem
- First, generate multiple pair-wise predictions
- Next, perform breadth-first search using a particle-swarm approach

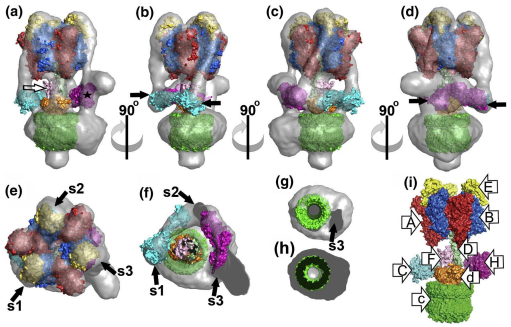


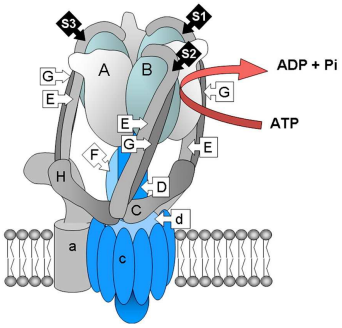
- The challenge – how to score the trial orientations efficiently?

Venkatraman and Ritchie (2011), manuscript submitted.

Assembling Molecular Machines?

- A recent example – the ATPase motor



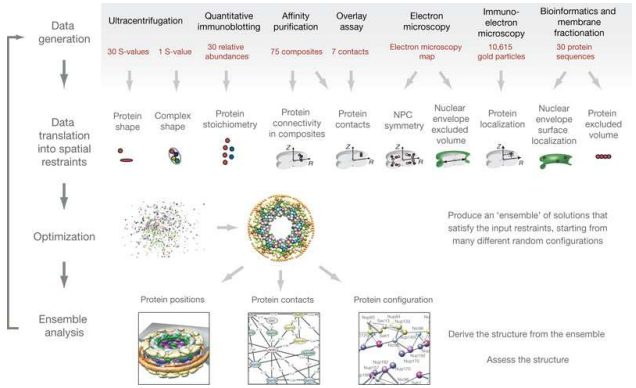


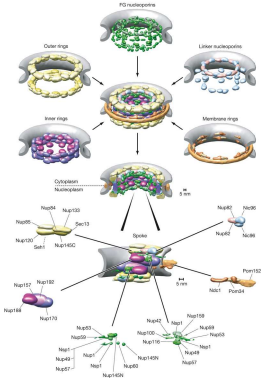
- There are hundreds (perhaps thousands?) more such machines!

Figure from Muench et al. (2009) J. Mol Biol 386 989–999

Putting It All Together?

- The Nuclear Pore Complex has some 650 protein components...





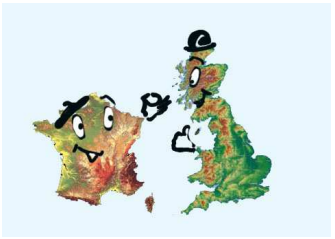
- It required an immense multi-disciplinary effort to build this model
- The challenge – can we do this automatically?

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

Conclusions

- Molecular shape recognition is an important aspect in:
  - Virtual drug screening
  - Protein-ligand interactions
  - Macromolecular assembly
- SPFs provide a novel and useful techinique for shape recognition
- Shape-based techniques will be increasingly useful in many areas:
  - Computational chemistry
  - Structural biology
  - ... and beyond!

And Finally – Special Thanks for the French Translation!



Anisah Ghoorah
Matthieu Chavent
Malika Smaïl-Tabbone
Yasmine Asses
Bernard & Françoise Maigret

French + English documents: http://www.loria.fr/∼ritchied/hdr/