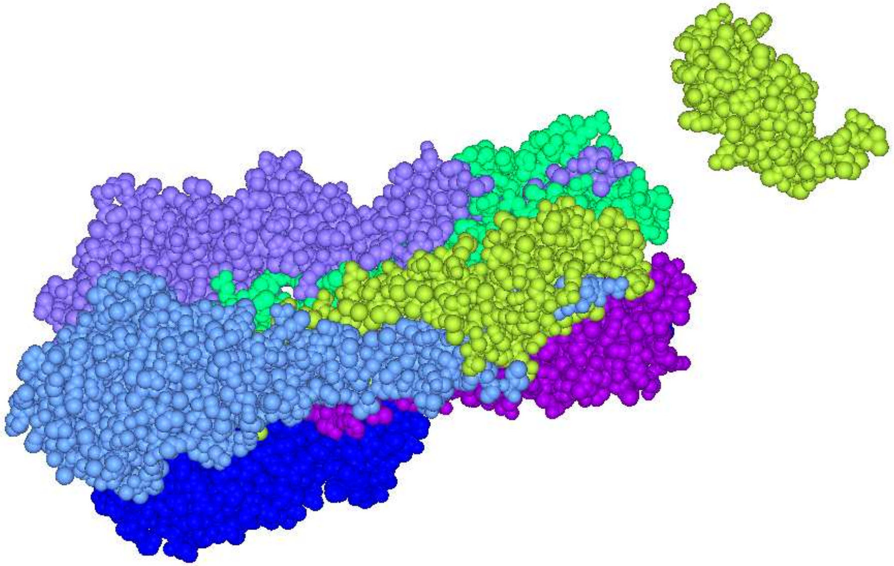
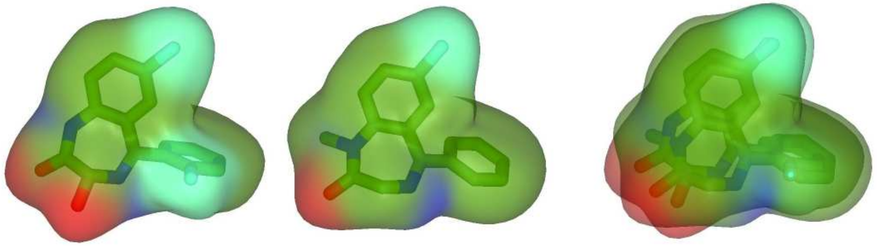
Protein Docking and 3D Ligand-Based Virtual Screening

Part 2





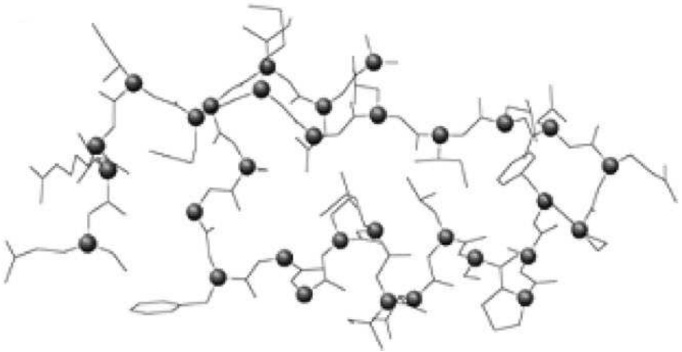
Dave Ritchie

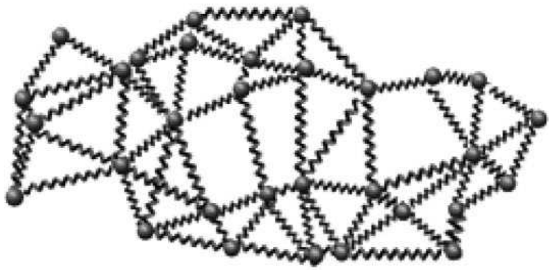
Orpailleur Team

INRIA Nancy – Grand Est

Modeling Protein Flexibility Using Elastic Network Models

- ENMs assume protein atoms (often just CAs) are coupled via a harmonic potential:





$$V = \sum_{i<j} C(d_{ij} - d^0_{ij})^2$$

$$H_{ij} = (\partial/\partial x_i)(\partial/\partial x_j)V$$

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

- C = constant, d_{ij} = distance, d^0_{ij} = reference distances, V = potential, $\underline{H}$ =Hessian

- $\underline{E}$ = matrix of eigenvectors $\underline{e}_i$ (normal mode “directions”), Λ_{ii} = eigenvalues (magnitudes)

- Then, sort by eigenvalues, and represent protein conformations as linear combinations

$$\underline{P}^{NEW} = \underline{P}^0 + \sum_{i=6}^{3N} w_i \underline{e}_i$$

- On-line examples: http://www.igs.cnrs-mrs.fr/elnemo/, and http://www.molmovdb.org/

- Problem #1: how to find weights w_i to give protein conformation $\underline{P}^{BOUND} = \underline{P}^{NEW}$?

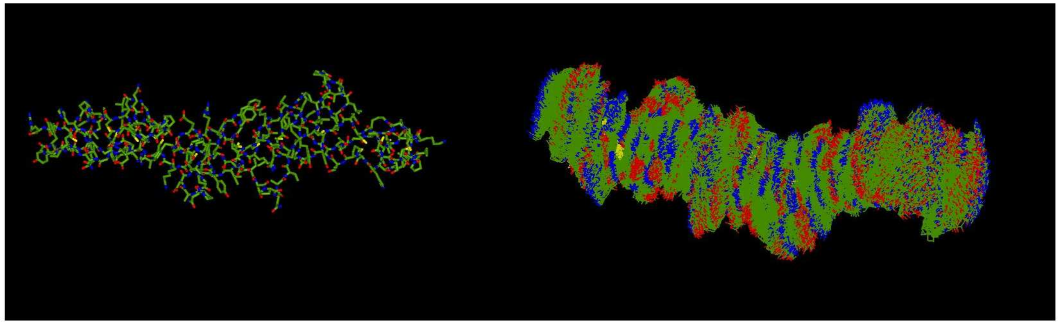
- Problem #2: How to sample and combine conformations for two proteins ?

Andrusier et al. (2008), Proteins, 73, 271–289 (recent review on flexible docking)

Tirion (1996) Physical Review Letters, 77, 1905–1908 (original ENM article)

Simulating Flexibility During Docking using "Essential Dynamics"

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



- Covariance matrix, C:

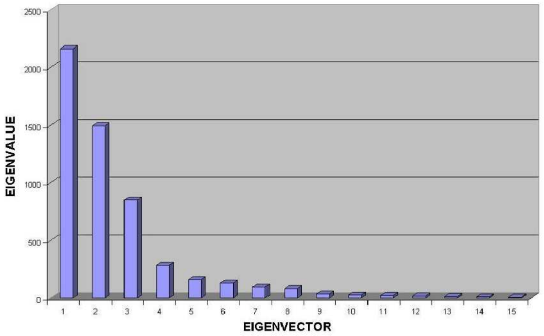
$$C_{ij} = < (x_i - \overline{x}_i)(x_j - \overline{x}_j) >$$

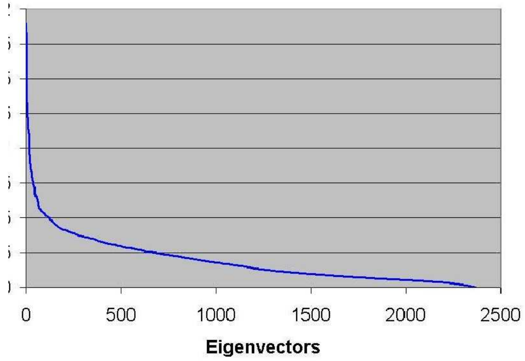
- Calculate eigenvectors, E:

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

- Estimate Unbound to Bound:

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





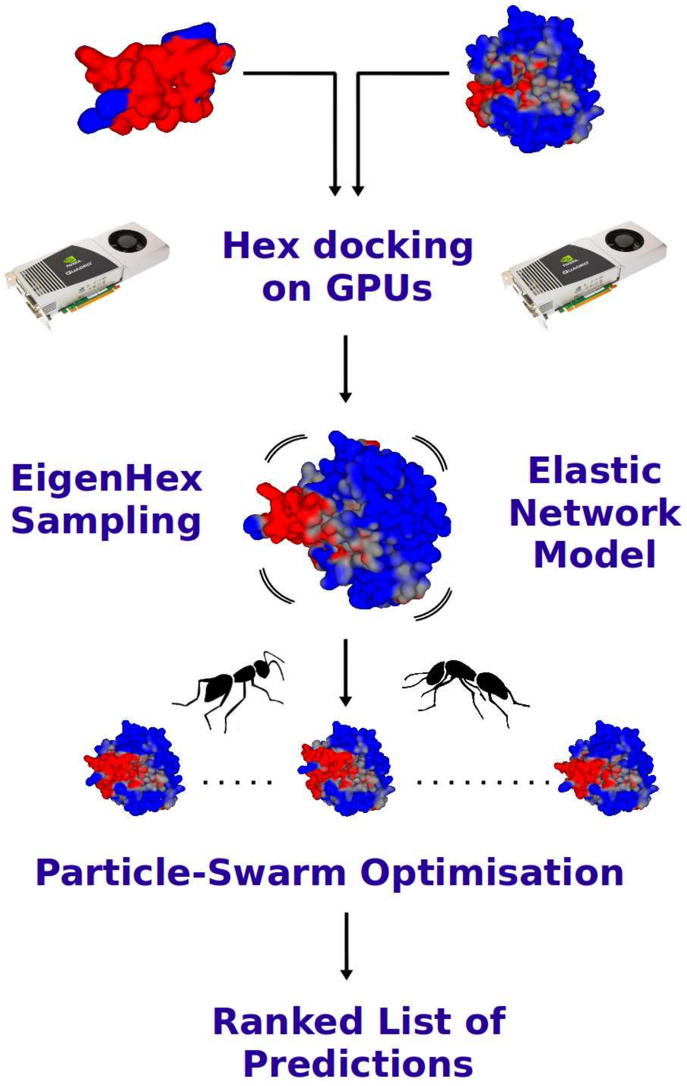
- The first few eigenvectors encode most of the internal fluctuations
- See also SwarmDock – http://bmm.cancerresearchuk.org/∼SwarmDock/

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

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

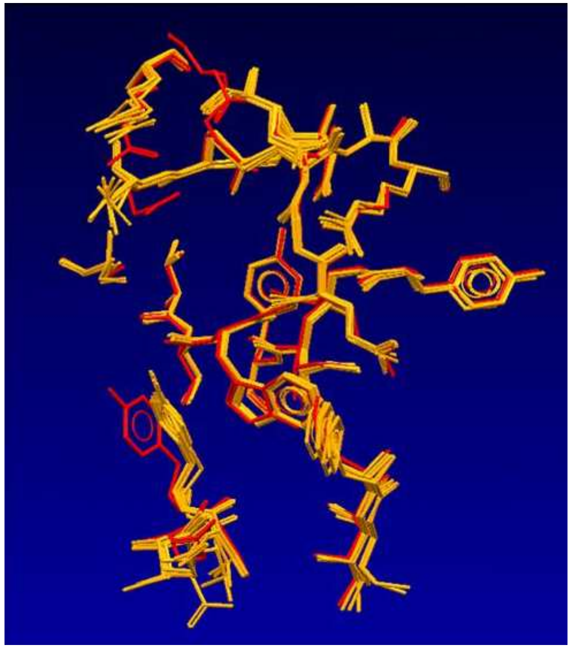
Results

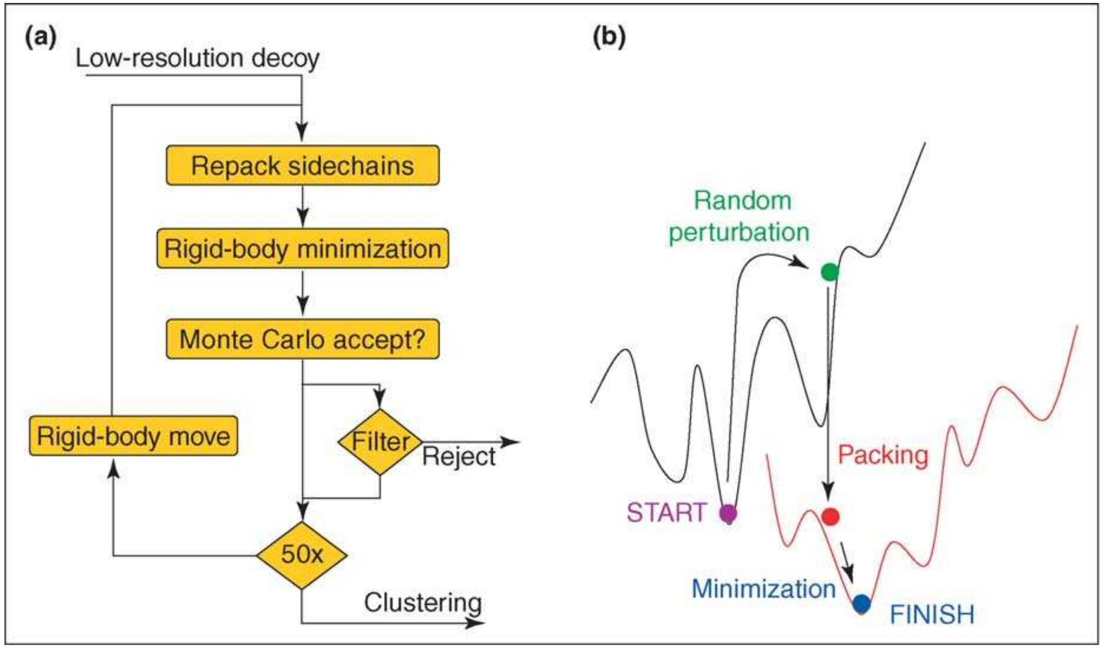
- DARS potential works well but...
- Still need a better scoring function
- Much effort – small improvement !!

Venkatraman, Ritchie (2012), Proteins – http://dx.doi.org/10.1002/prot.24115

RosettaDock – Flexible Refinement by Side Chain Re-Packing

- Given a rigid body starting pose, repeat 50 times:
  - REMOVE and RE-BUILD side chains; apply local rigid-body minimisation
  - apply Monte-Carlo accept/reject



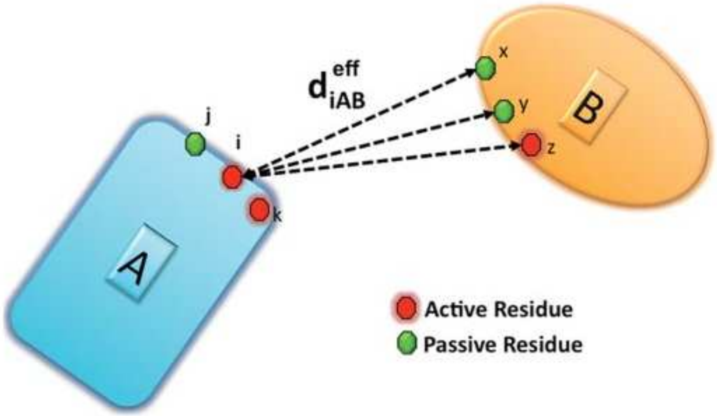


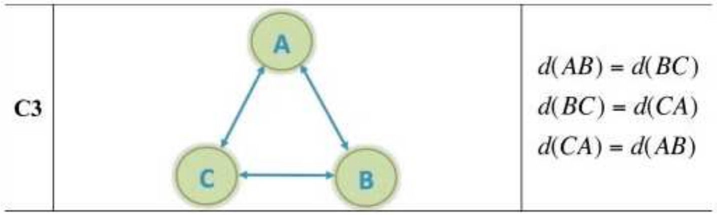
- Successful for several CAPRI targets; also works well for 50% of Docking Benchmark v2

Gray (2006) Current Opinion in Structural Biology, 16, 183–193

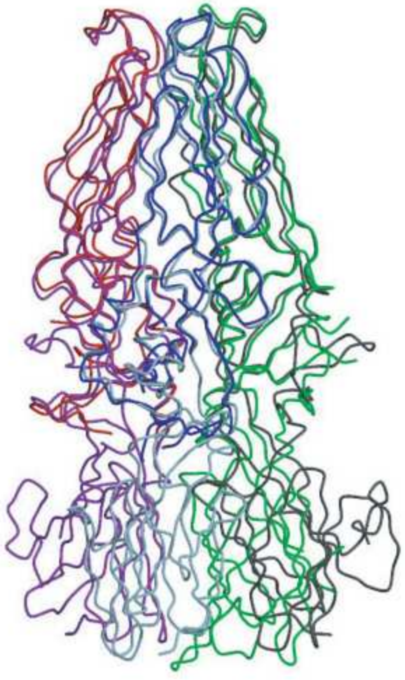
Haddock – "Highly Ambiguous Data-Driven Docking"

- 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: $d^{\text{eff}}_{iAB} = \left(\sum_{m_{iA}=1}^{N_{iA}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{kB}} \left(\frac{1}{d^6_{m_{iA},n_{kB}}}\right) \right)^{-1/6}$





$d(AB) = d(BC)$
$d(BC) = d(CA)$
$d(CA) = d(AB)$



T10 = TEV trimer

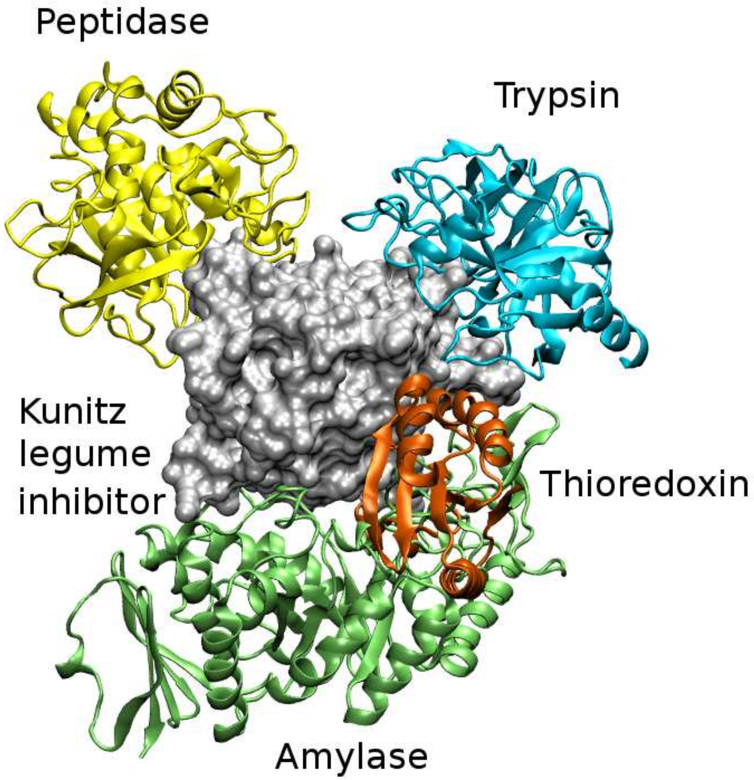
- V. good CAPRI results
- Restraints from (e.g.):
  SAXS
  mutagenesis
  mass spectroscopy
  NMR (RDC, CSP)

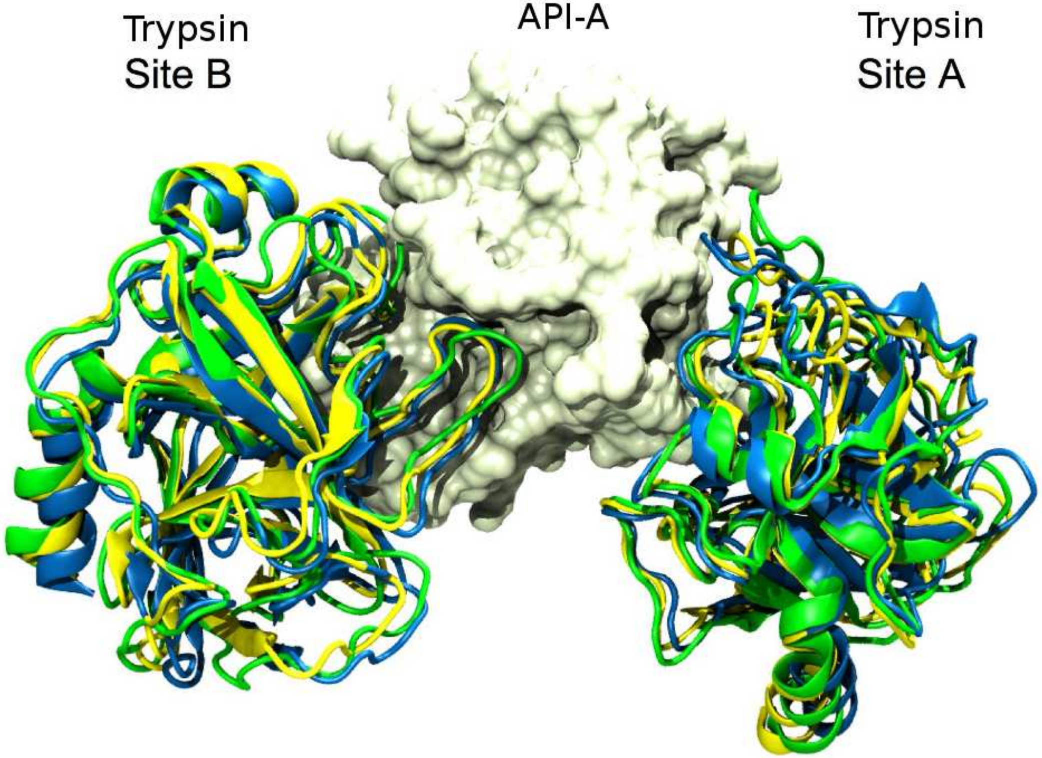
van Dijk et al. (2005) FEBS J, 272, 293–312

van Dijk et al. (2005) Proteins, 60, 232–238

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



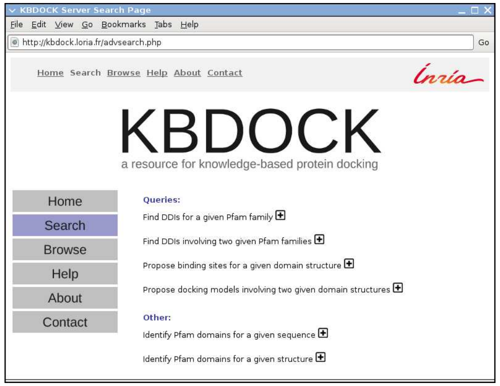


- Performing focused Hex + MD refinement gave a total of 9 “acceptable” solutions

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
http://kbdock.loria.fr/advsearch.php
Go
Home Search Browse Help About Contact
Inria
Home
Search
Browse
Help
About
Contact

KBDOCK

a resource for knowledge-based protein docking

Queries:

Find DDIs for a given Pfam family ⊞

Find DDIs involving two given Pfam families ⊞

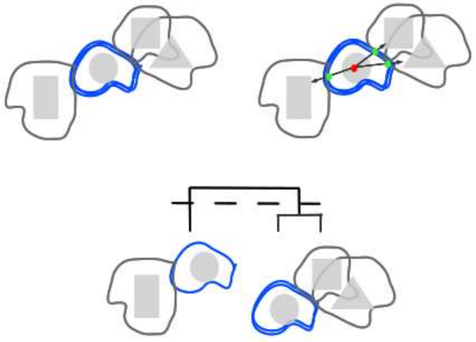
Propose binding sites for a given domain structure ⊞

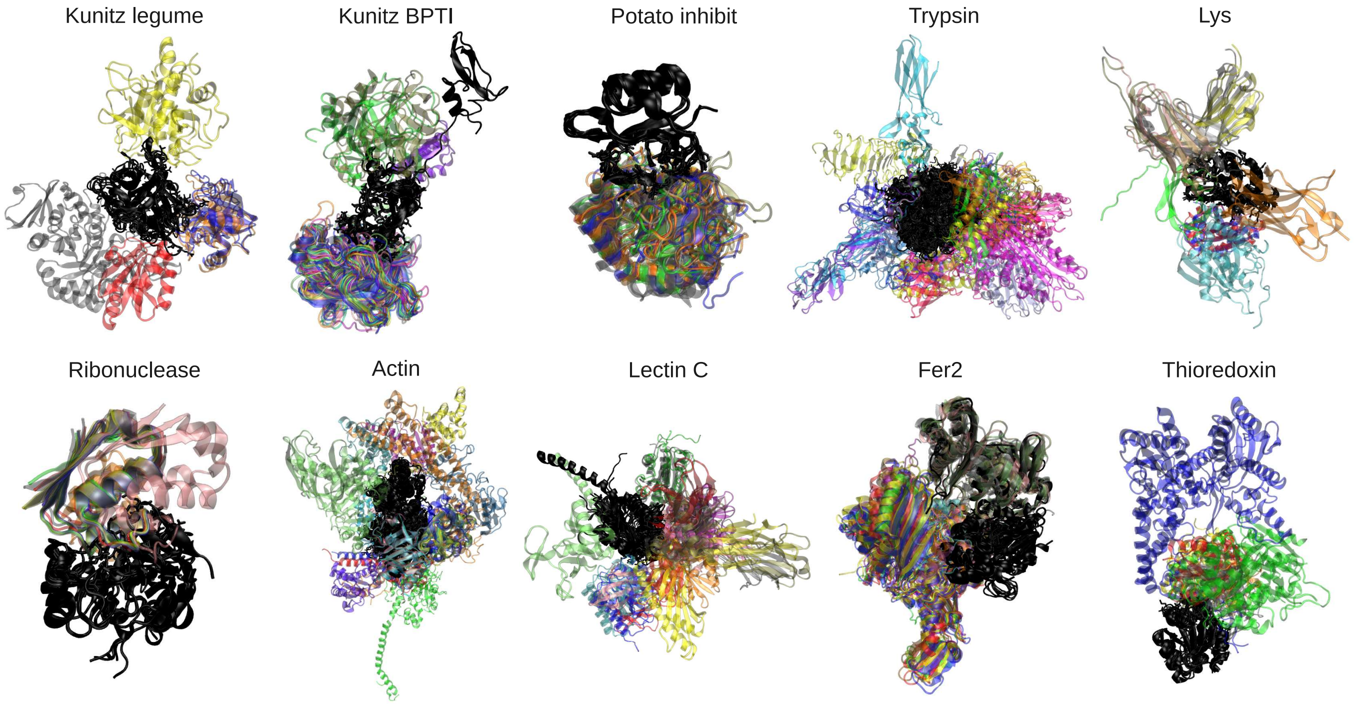
Propose docking models involving two given domain structures ⊞

Other:

Identify Pfam domains for a given sequence ⊞

Identify Pfam domains for a given structure ⊞

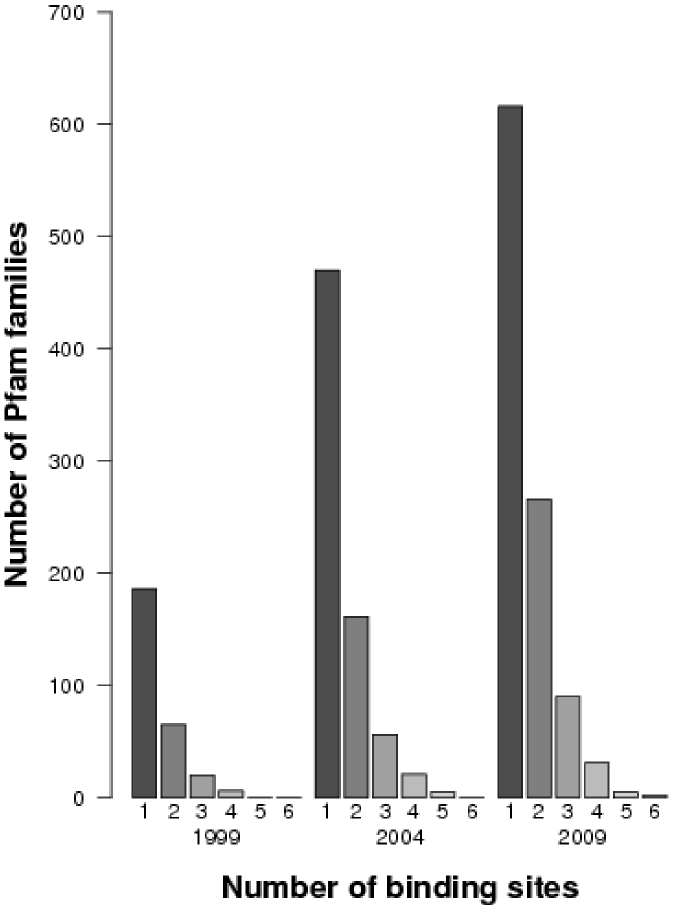


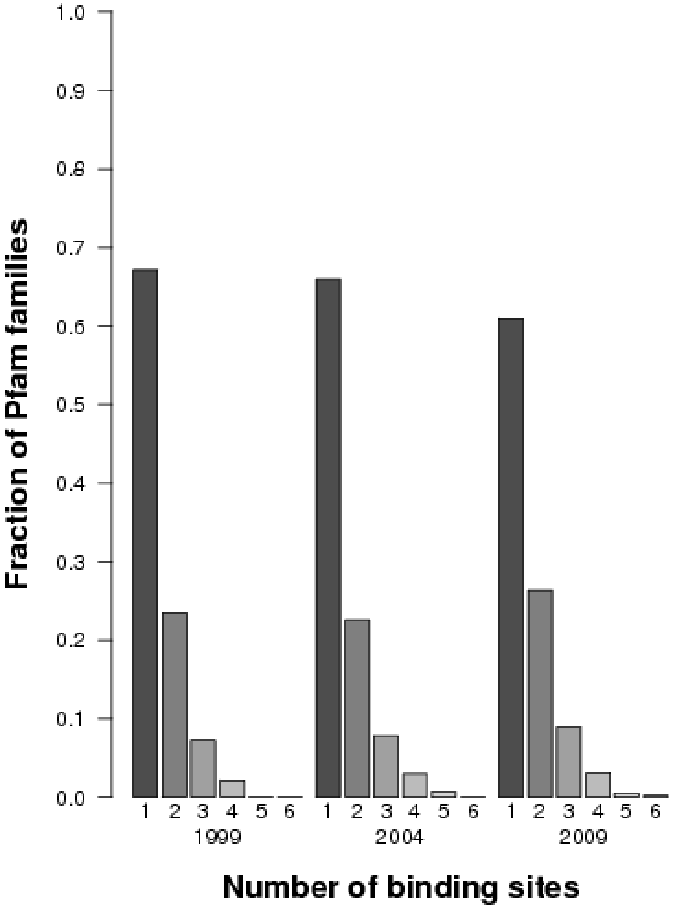


- Aim: to provide PFAM family-level structural templates for knowledge-based docking

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

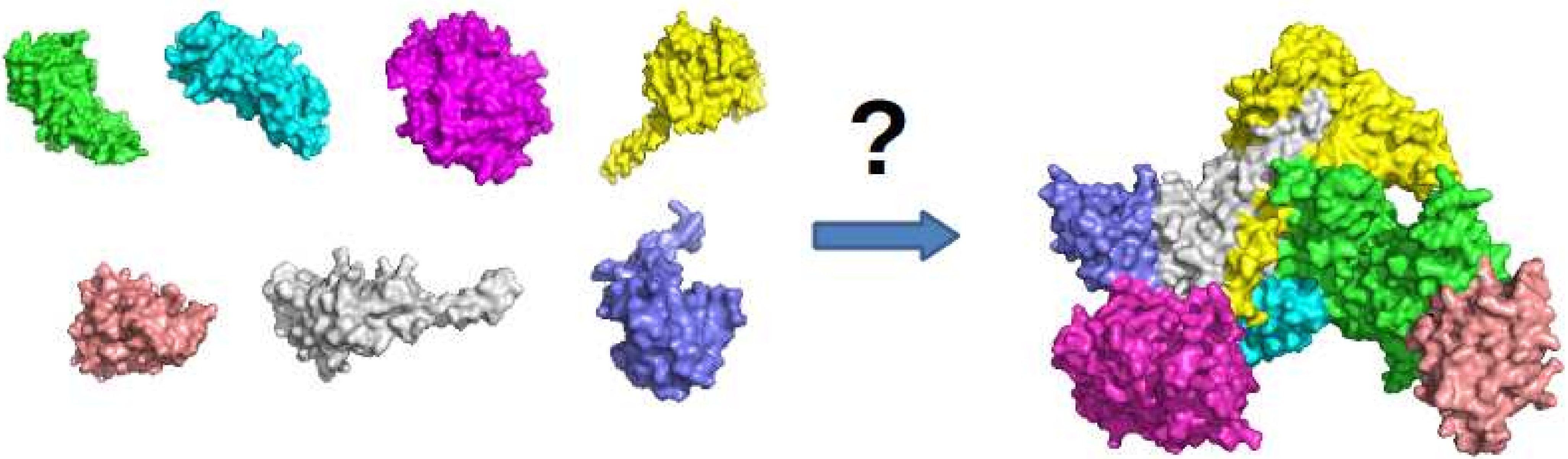
Target class	Total targets	FH templates	Two SH templates	One SH template	Zero templates
Without date filtering					
Enzyme	36	24 / 24	(3 + 1) / 5	3 / 5	2
Other	37	21 / 21	(0 + 0) / 3	5 / 11	2
With date filtering					
Enzyme	36	13 / 13	(2 + 1) / 5	7 / 11	7
Other	37	13 / 13	(0 + 0) / 1	8 / 15	8

- 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

Assembling Multi-Component Protein Complexes

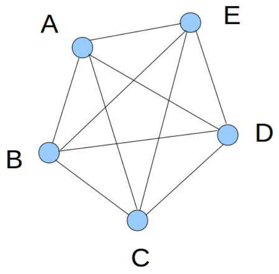
- Multi-component assembly is a highly combinatorial problem
- How to generate and score candidate orientations efficiently?



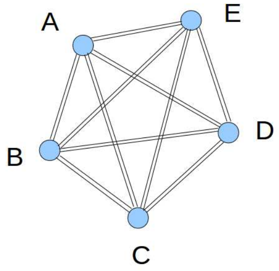
- Here, we use Minimum Weight Spanning Trees (MSTs), (Inbar et al., 2003)
- ... with an ant colony particle swarm optimisation (PSO) search algorithm

Inbar et al. (2003), Bioinformatics, 2003, i158–i168

Minimum Energy Spanning Trees

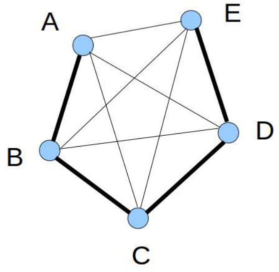


- **Here, we have N = 5 proteins and K = N(N-1)/2 = 10 “edges”**



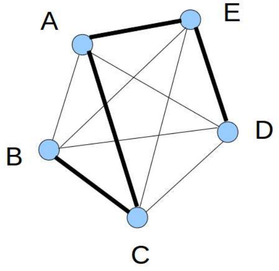
- **Each edge should consider many (e.g. P = 100) docking solutions**

- **Naive enumeration would give $P^{N(N-1)/2}$ possible combinations**



- **A spanning tree visits each node just once...**

- **... there are only $P^{N-1}N^{N-2}$ distinct spanning trees**

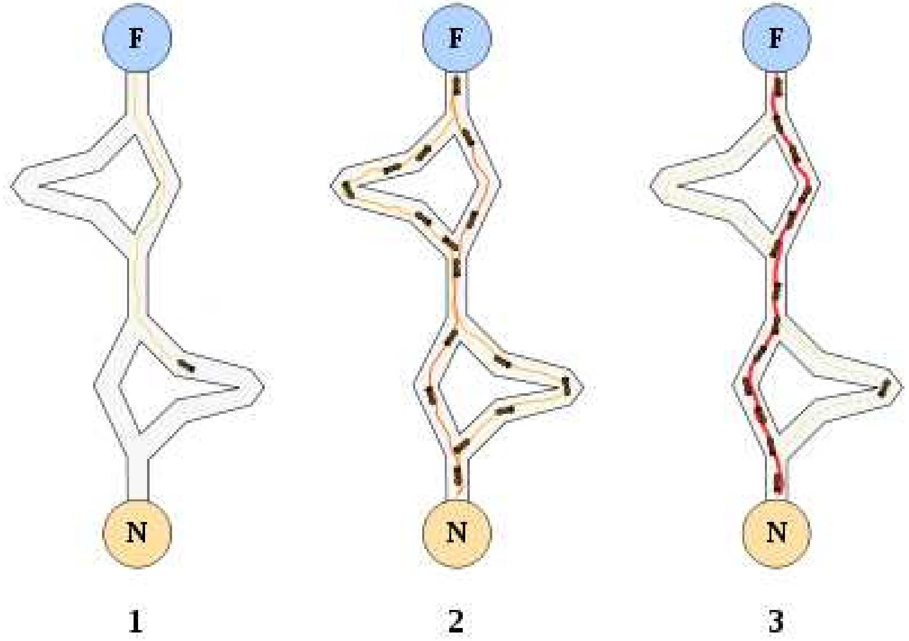


- **... and when $N < P$, we get $P^{N-1}N^{N-2} << P^{N(N-1)/2}$**

- **Strategy: search for the minimum energy spanning tree ...**

- **Getting technical: this is an “edge-weighted K-cardinality” problem...**

Multi-Component Docking using Ant-Colony Optimisation



Ant colonly optimisation is based on the behaviour of real ants

When an ant finds food, it leaves a trail of pheromones

Other ants follow strong pheromones trails to reach the food quickly

- Here, we use 10 ants in parallel for 1,000 iterations...
- Each ant is asigned to a randomly generated spanning tree
- It must detect and score steric clashes, and update its trail
- It then makes a new spanning tree using the latest pheromone trails...

MDOCK – Multi-Component Docking Results

- There are not many multi-component examples in the PDB

- Therefore, several 'targets" were made from the same complex...

- 1VCB = von Hippel-Lindau ElonginC-ElonginB tumor suppressor protein
- 1IKN = Transcription factor I-kappa-B-alpha / NF-kappa-B
- 1K8K = Bovine actin polymerisation initiation complex Arp2 / Arp3

Target	Chains	Time (min)	Rank	RMSD (Å)	Best RMSD (Å)
1VCB	A,B,C	43.8	1	0.58	0.58
1IKN	A,C,D	77.3	1	9.17	0.88
1K8K	A,B,D,E	123.5	1	4.96	2.19
1K8K	A,B,D,E,F	168.6	2	9.48	2.99
1K8K	A,B,D,E,F,G	194.1	15	4.63	3.53
1K8K	A,B,C,D,E,F,G	366.9	–	–	10.21

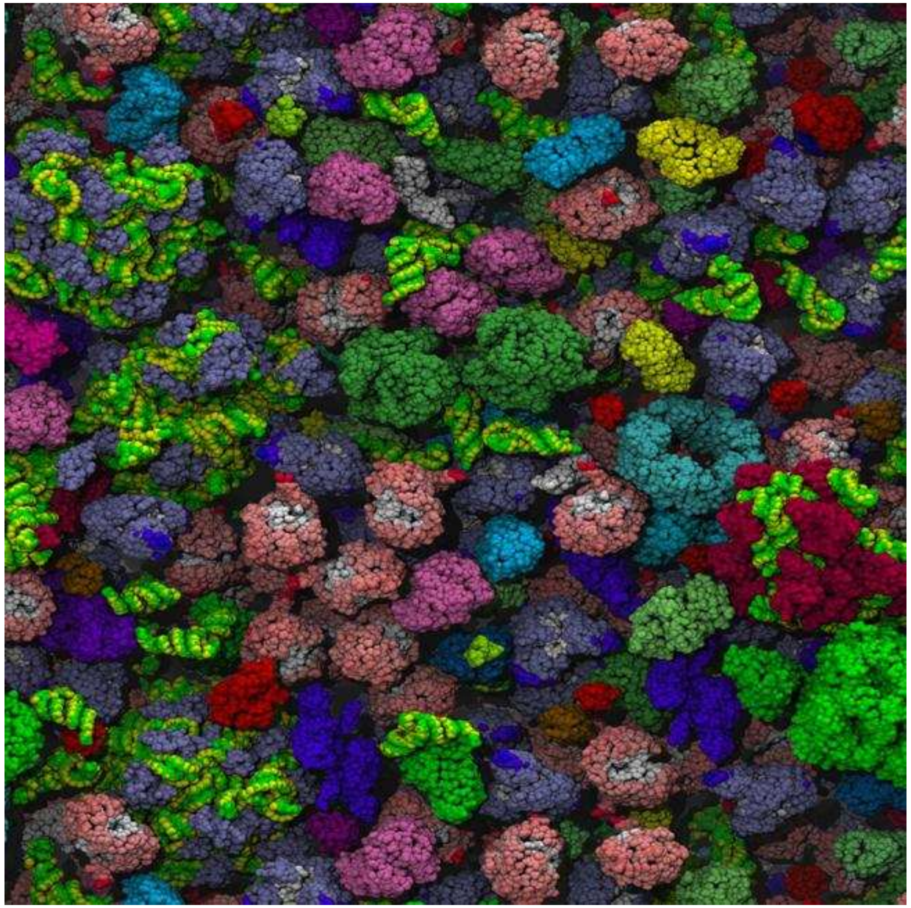
- Mostly good results, but why did we miss one?

- However, it would be very expensive to apply this algorithm to blind docking ...

Venkatraman, Ritchie (2012), in press.

The Inside of a Cell is Highly Crowded

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

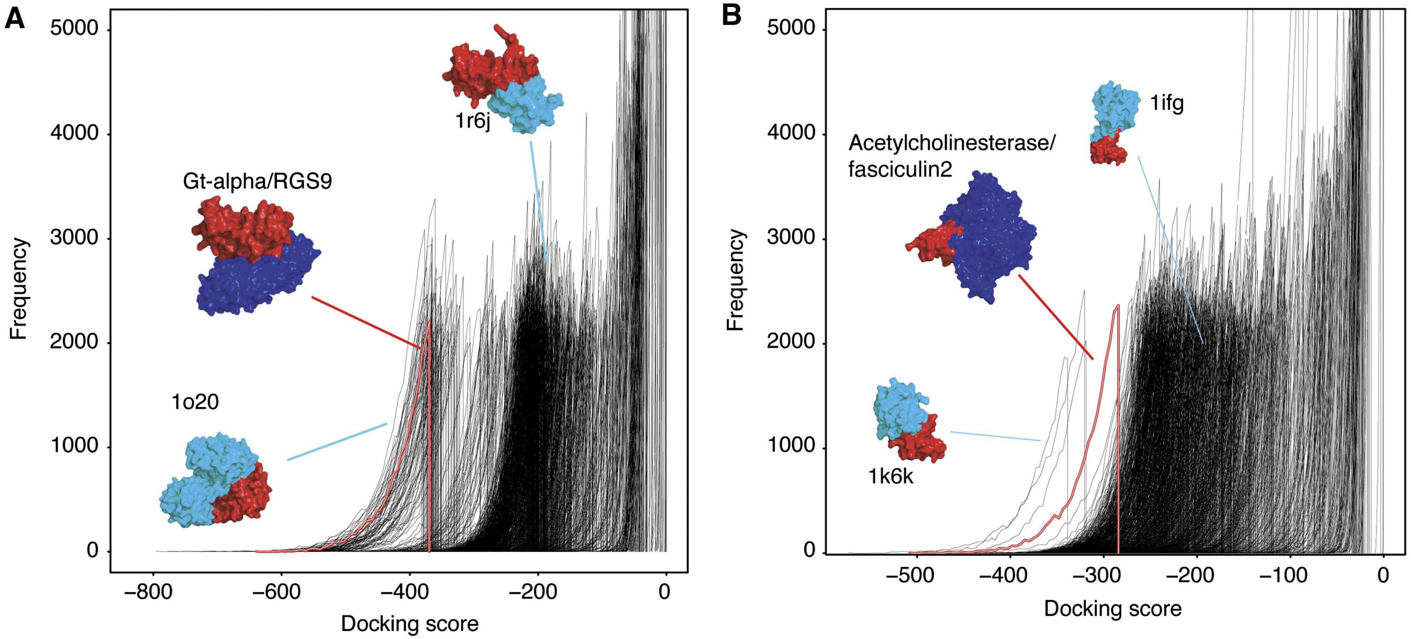


- Can we use docking algorithms to predict the protein-protein interactions ?

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

Large-Scale Cross-Docking Has Only Recently Become Feasible

- 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

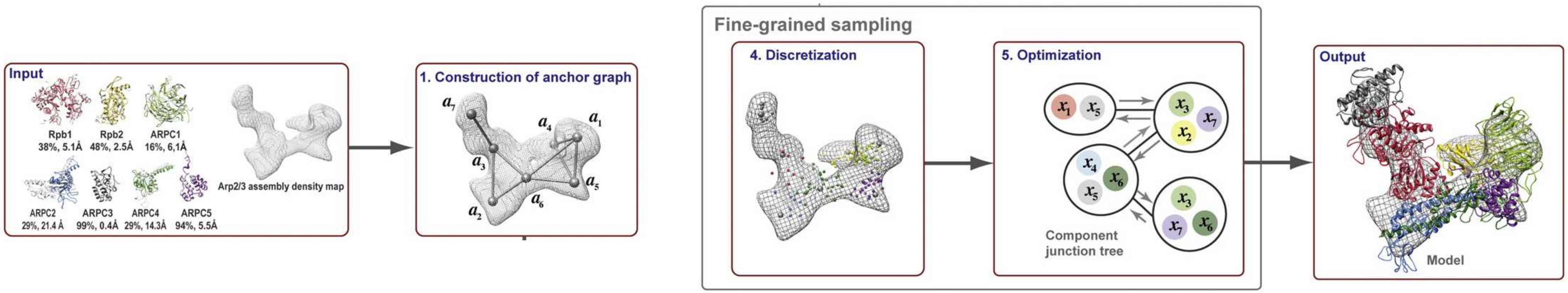
- Only 8/56 true PPIs have indistinguishable profiles to the non-binders

- NB. this experiment is detecting energy funnels, not necessarily the correct docking pose

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

IMP – Integrative Modeling Platform

- Python-based system for integrative multi-component modeling – http://salilab.org/imp/



- Combines structural data from: cryoEM (mainly), X-Ray, NMR, SAXS, Modeller, ...

  ... with interaction data from BioGRID – http://thebiogrid.org/

- The overall approach is to maximise a multi-term objective function:

  $$F = \sum_i \alpha_i + \sum_{i<j} \beta_{ij}$$

  α_i are single-body terms (e.g. goodness of fit in a density map, protrusion penalty)

  β_{ij} are two-body terms (e.g. the docking score for two proteins in contact)

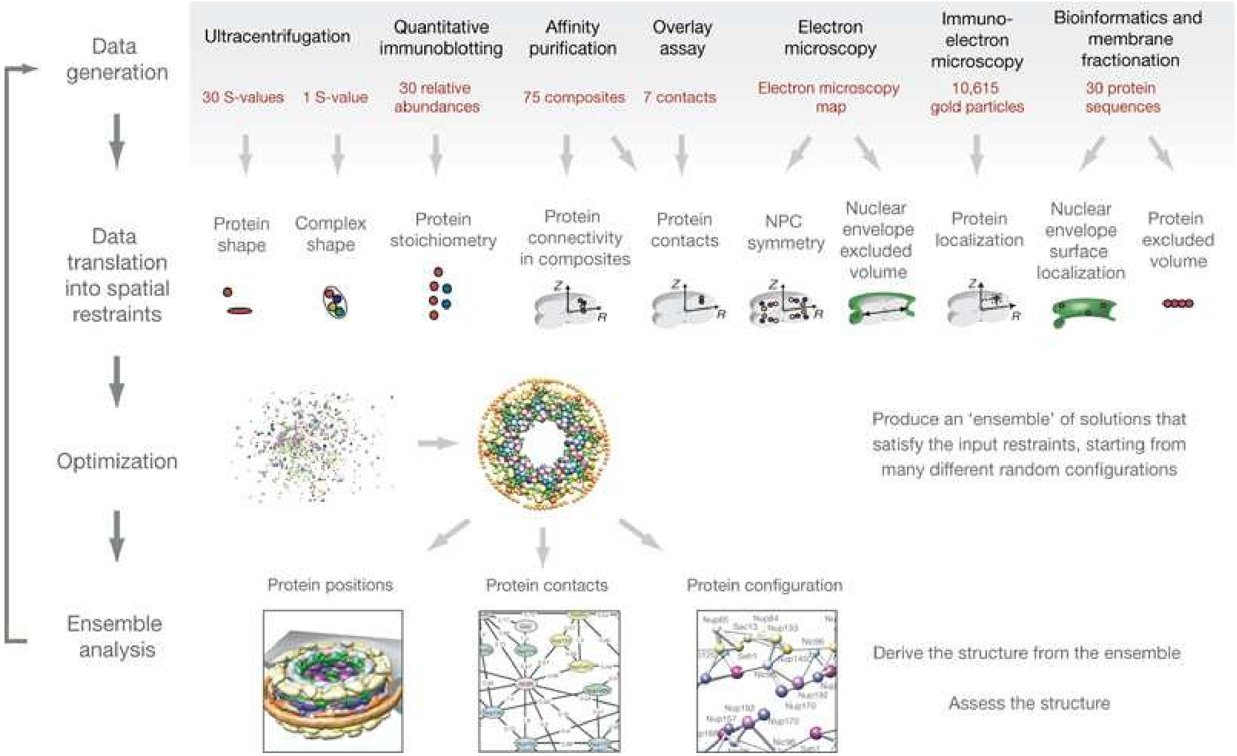
- But it is *highly* combinatorial search space, with missing/incomplete data...

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



Data generation

Ultracentrifugation: 30 S-values, 1 S-value

Quantitative immunoblotting: 30 relative abundances

Affinity purification: 75 composites

Overlay assay: 7 contacts

Electron microscopy: Electron microscopy map

Immuno-electron microscopy: 10,615 gold particles

Bioinformatics and membrane fractionation: 30 protein sequences

Data translation into spatial restraints

Protein shape

Complex shape

Protein stoichiometry

Protein connectivity in composites

Protein contacts

NPC symmetry

Nuclear envelope excluded volume

Protein localization

Nuclear envelope surface localization

Protein excluded volume

Optimization

Produce an 'ensemble' of solutions that satisfy the input restraints, starting from many different random configurations

Ensemble analysis

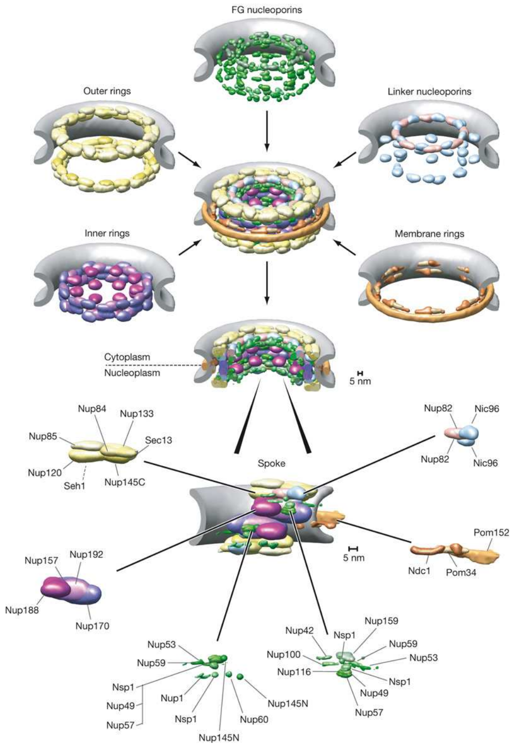
Protein positions

Protein contacts

Protein configuration

Derive the structure from the ensemble

Assess the structure



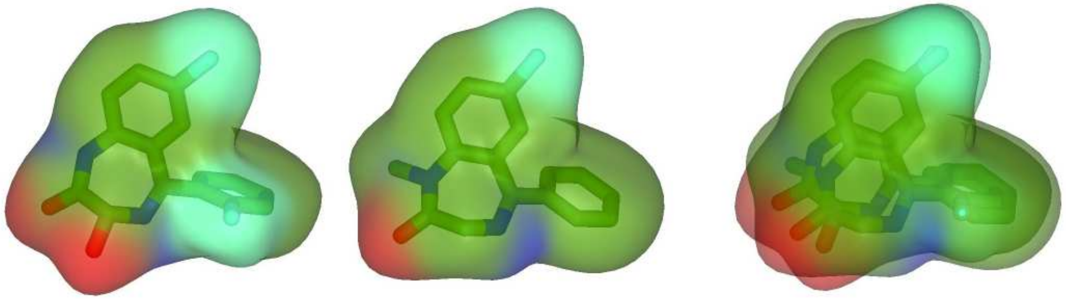
- It required an immense multi-disciplinary effort to build this model ...

- 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) – http://dx.doi.org/10.1002/prot.24092

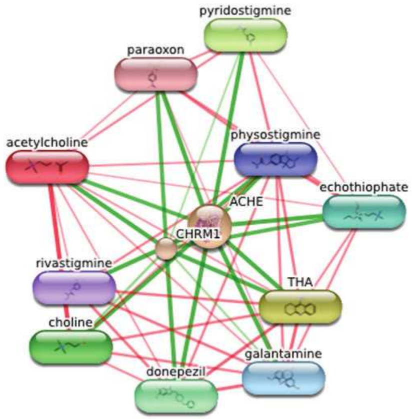
But What About the Virtual Screening ?

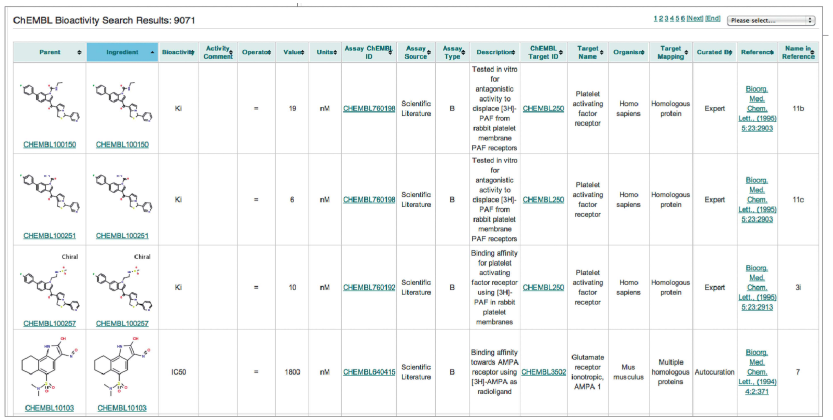


Protein-Drug Interaction Resources

> “The availability of interaction data between small molecule drugs and protein targets has increased substantially in recent years... We assembled a total of 4,767 unique interactions between 802 drugs and 480 targets, which means that on average every drug interacts with at least 6 targets...” Mestres et al., (2009).

- STITCH – Search Tool for Interactions of Chemicals – http://stitch.embl.de
  - 68,000 chemicals (including 2,200 drugs) linked to 1.5 million genes
- ChEMBL – database of drug-like bioactive molecules – https://www.ebi.ac.uk/chembldb/
  - Binding & toxicity data for 1.1 million compounds and 5,200 protein targets




ChEMBL Bioactivity Search Results: 9071

Parent	Ingredient	Bioactivity	Activity Comment	Operator	Value	Units	Assay ChEMBL ID	Assay Source	Assay Type	Description	ChEMBL Target ID	Target Name	Organism	Target Mapping	Curated By	Reference	Name in Reference
CHEMBL100150	CHEMBL100150	Ki		=	19	nM	CHEMBL760198	Scientific Literature	B	Tested in vitro for antagonistic activity to displace [3H]-PAF from rabbit platelet membrane PAF receptors	CHEMBL250	Platelet activating factor receptor	Homo sapiens	Homologous protein	Expert	Bioorg. Med. Chem. Lett., (1995) 5:23:2903	11b
CHEMBL100251	CHEMBL100251	Ki		=	6	nM	CHEMBL760198	Scientific Literature	B	Tested in vitro for antagonistic activity to displace [3H]-PAF from rabbit platelet membrane PAF receptors	CHEMBL250	Platelet activating factor receptor	Homo sapiens	Homologous protein	Expert	Bioorg. Med. Chem. Lett., (1995) 5:23:2903	11c
Chiral CHEMBL100257	Chiral CHEMBL100257	Ki		=	10	nM	CHEMBL760192	Scientific Literature	B	Binding affinity for platelet activating factor receptor using [3H]-PAF in rabbit platelet membranes	CHEMBL250	Platelet activating factor receptor	Homo sapiens	Homologous protein	Expert	Bioorg. Med. Chem. Lett., (1995) 5:23:2913	3i
CHEMBL10103	CHEMBL10103	IC50		=	1800	nM	CHEMBL640415	Scientific Literature	B	Binding affinity towards AMPA receptor using [3H]-AMPA as radioligand	CHEMBL3502	Glutamate receptor ionotropic, AMPA 1	Mus musculus	Multiple homologous proteins	Autocuration	Bioorg. Med. Chem. Lett., (1994) 4:2:371	7

Mestres et al. (2009) Molecular BioSystems, 5, 1051–1057

Kuhn et al. (2008) Nucleic Acids Research, 36, D684–D688 (STITCH)

Gaulton et al. (2012) Nucleic Acids Research, 40, D1100–D1107 (ChEMBL)

A Growing Interest in Drug Promiscuity and Drug Repositioning

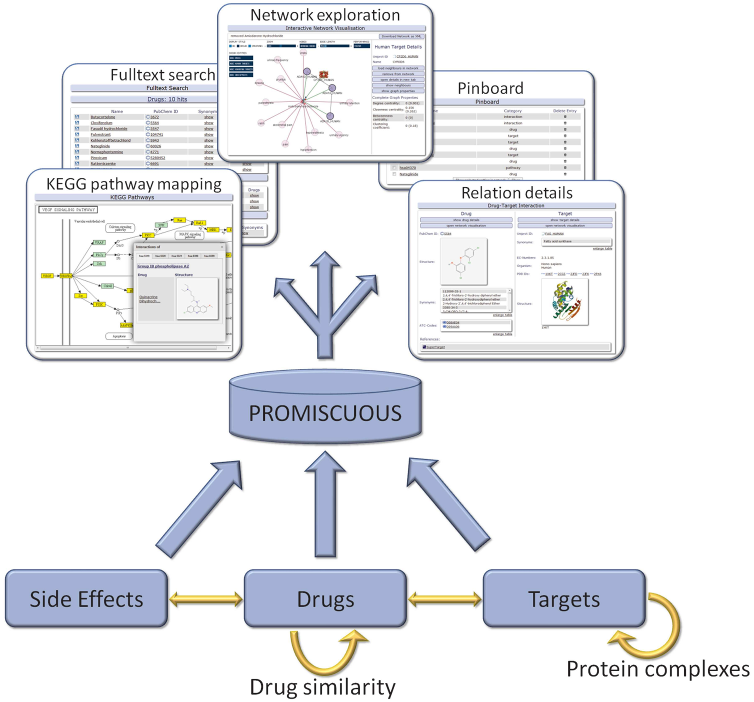
- Approx 90% of new drugs fail to reach market – often due to toxicity or lack of efficacy

  toxicity – e.g. from unwanted off-target interactions

  lack of efficacy – e.g. from robustness of biological network

- Example – PROMISCUOUS – http://bioinformatics.charite.de/promiscuous/

  “network-based drug repositioning” – 25K drugs, 21K drug-protein, 104K PPIs

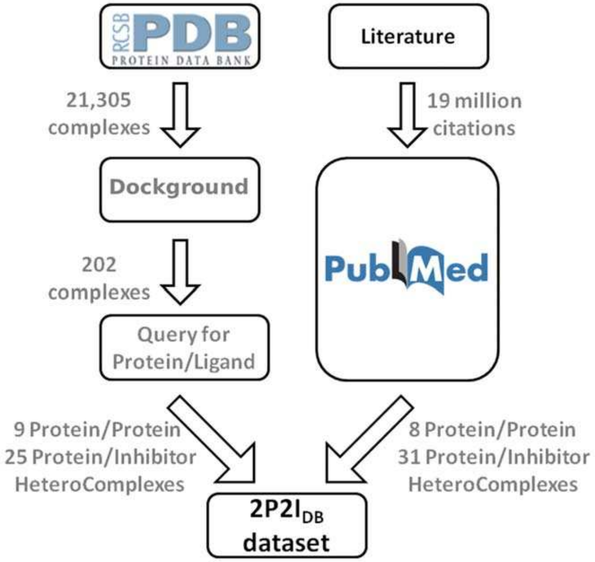


von Eichborn et al. (2011) Nucleic Acids Research, 39, D1060–D1066 (PROMISCUOUS)

Kuhn et al. (2011) Molecular Systems Biology, 6, 343 (SIDER)

Atomic Resolution Studies of Hetero PPIs and Inhibitors

- TIMBAL – http://www-cryst.bioc.cam.ac.uk/timbal/ – 27 structures, 104 small molecules
- 2P2I (= PPI inhibition) – http://2p2idb.cnrs-mrs.fr/ – 17 PPIs, 56 small molecules



RCSB PDB PROTEIN DATA BANK

Literature

21,305 complexes

19 million citations

Dockground

PubMed

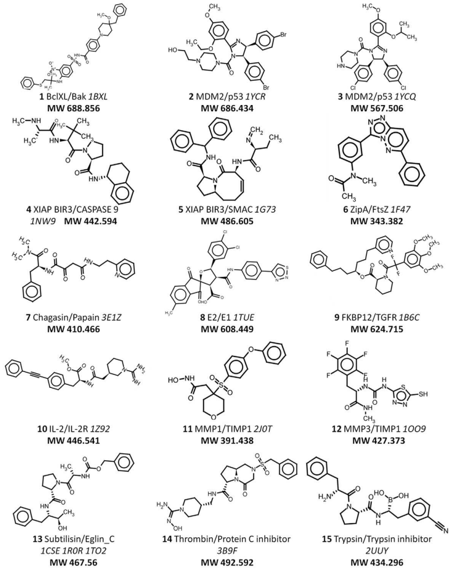
202 complexes

Query for Protein/Ligand

9 Protein/Protein
25 Protein/Inhibitor
HeteroComplexes

8 Protein/Protein
31 Protein/Inhibitor
HeteroComplexes

$2P2I_{DB}$ dataset



  - The ligands generally have high MW and are hydrophobic
  - The PPIs have few/no small interface pockets; small conformational changes on binding
- See also Dr. PIAS – http://www.drpias.net/ – SVM-based prediction of druggable PPIs

Higueruerlo et al. (2009) Chemical Biology Drug Design, 74, 457–467 (TIMBAL)

Borgeas et al. (2010) PLoS One, 4, e9598 (2P2I)

Sugaya, Furuya (2011) BMC Bioinformatics, 12, 50 (Dr. PIAS)

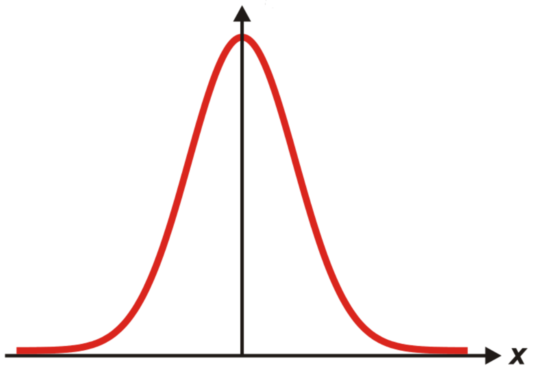
A Gaussian Representation of Molecular Shape

- Represent each atom in a molecule as a 3D Gaussian density function:

$$\rho_i(r) = \beta e^{-\gamma r^2/\sigma_i^2}$$

and choose β, γ such that: $\int \rho_i(r) \mathrm{d}\underline{x} = \frac{4}{3}\pi\sigma_i^3$

where σ_i = van der Waals radius of atom i

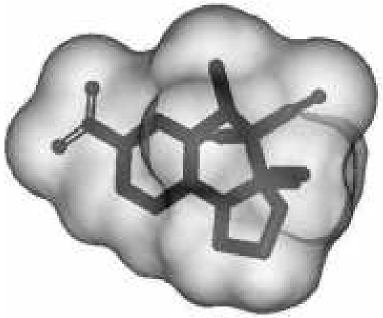


- Represent the "density" of a molecule as a sum of N atomic densities:

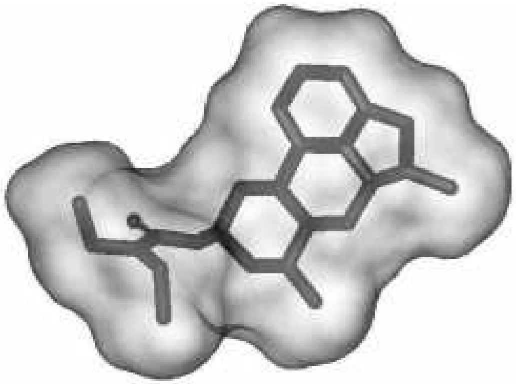
$$\rho = \sum_{i<N} \rho_i - \sum_{i<j<N} \rho_i\rho_j + \sum_{i<j<k<N} \rho_i\rho_j\rho_k - \ldots$$

$$= 1 - \Pi_{i=0}^{N-1}(1 - \rho_i)$$

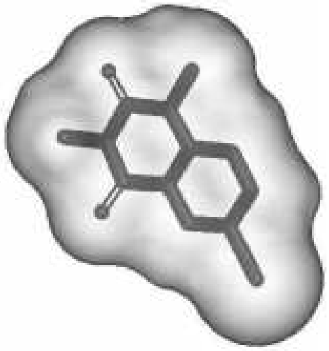
- Some examples:



HIFYOK



PEWHII



ABMQZD

Grant et al. (1996) J Computational Chemistry, 17, 1653–1666

Gaussian Overlap Volumes and Tanimoto Scores

- **The overlap volume between two atomic Gaussians is just another Gaussian:**

$$V_{ij} = \int \rho_i \rho_j \mathrm{d}\underline{x}$$

$$= \beta_i \beta_j \left(\frac{\pi}{\alpha_i+\alpha_j}\right)^{3/2} e^{-(\frac{\alpha_i\alpha_j}{\alpha_i+\alpha_j})R_{ij}^2}$$

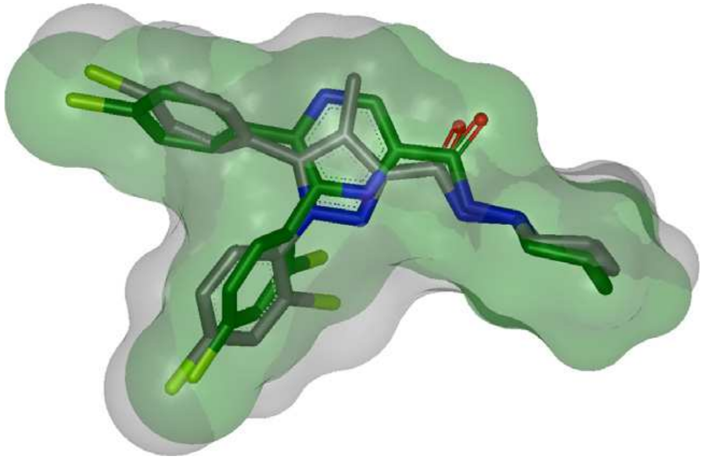
where R_{ij} is distance between the atom centres, and $\alpha_i = \gamma/\sigma_i^2$

- **Hence the overlap volume between two molecules can also be calculated easily...**

... and normalised to give a Tanimoto-like similarity score (with range $0 < S_{AB} \leq 1.0$):

$$V_{AB} = \int \rho_A \rho_B \mathrm{d}\underline{x}$$

$$S_{AB} = \frac{V_{AB}}{V_{AA}+V_{BB}-V_{AB}}$$



... and this is easy to optimise:

$$\frac{\delta V_{ij}}{\delta x} = -2\left(\frac{\alpha_i\alpha_j}{\alpha_i+\alpha_j}\right)(x_i - x_j)V_{ij}$$

etc.

Haigh, Pickup (2005), J Chemical Information & Modeling, 45, 673–684

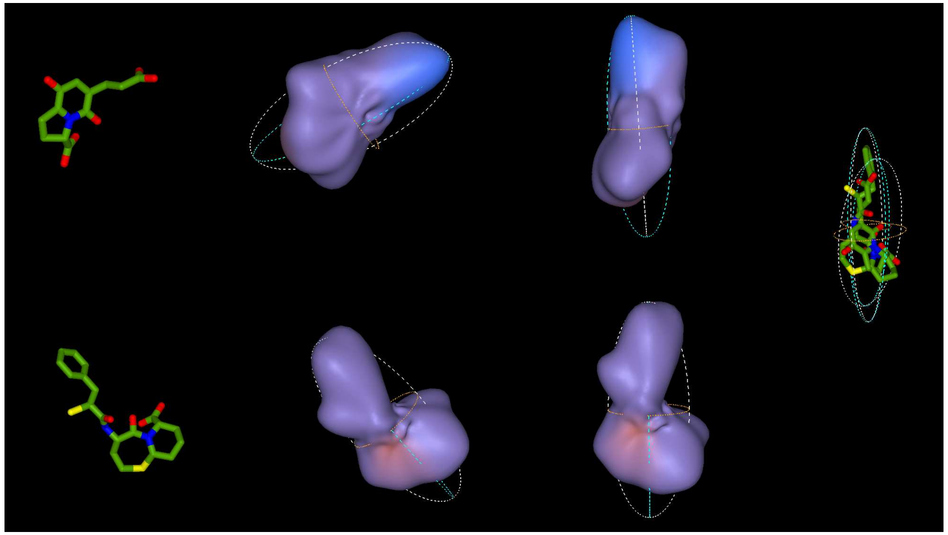
ROCS – "Rapid Overlay of Chemical Structures"

- ROCS is a commercial implementaton of the Gaussian representation of Grant at al.
- ROCS initially uses "steric multipoles" to align molecules with the Cartesian axes

  $M_{pq} = \int pq\rho \mathrm{d}\underline{x}$, where each p and q stands for x, y, or z

  Diagonalising M is equivalent to finding the principal ellipsoidal axes...

- ROCS then maximises the Gaussian overlap starting from 4 different orientations (axis flips)



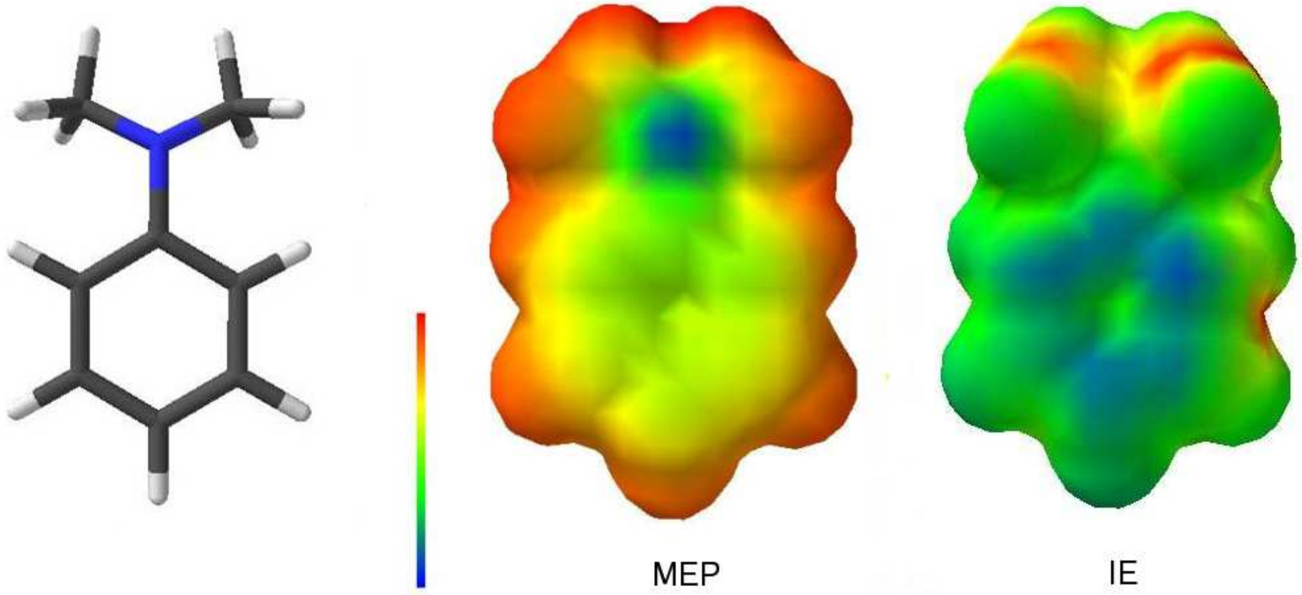
- Recently, "FASTROCS" (GPU-based version) – up to 10^6 superpositions/second !!

http://www.eyesopen.com/rocs/

Haque, Pande (2009), J Computational Chemistry, 31, 117–132 (open source GPU version)

ParaSurf – SH Surfaces & Properties from Semi-Empirical QM

- From MOPAC or VAMP calculate:
  - Density contours of $2 \times 10^{-4}\mathrm{e}/\mathrm{\AA}^3$ ($\sim$ 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 Chemical Information & Modeling, 45, 1010–1016

Clark (2004), J Molecular Graphics, 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: $S = \underline{a}.\underline{b}'/(|\underline{a}|.|\underline{b}|)$

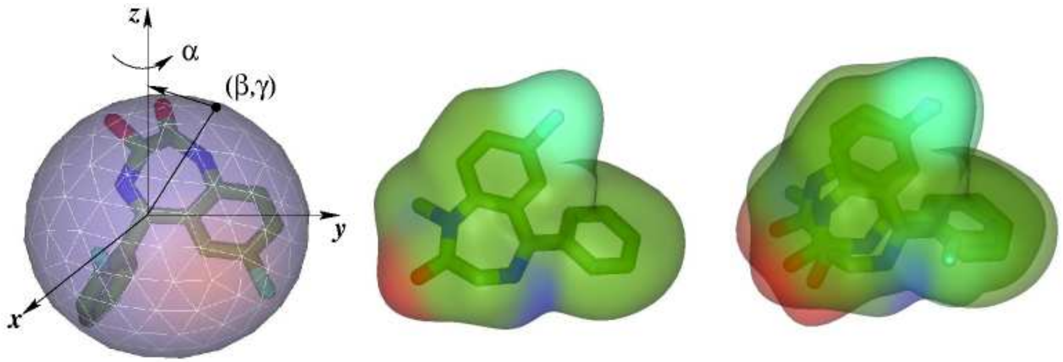
Tanimoto: $S = \underline{a}.\underline{b}'/(|\underline{a}|^2 + |\underline{b}|^2 - \underline{a}.\underline{b}')$

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

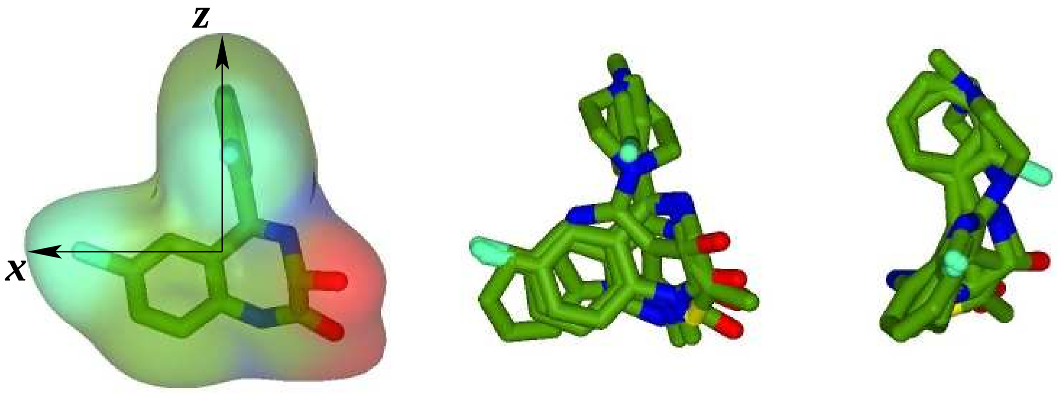
Perez-Nueno et al. (2010), Molecular Informatics, 30, 151–159

Brute-Force Spherical Harmonic Surface Superpositions

- Generate 22,500 Euler rotations from icosahedral tesselation of sphere



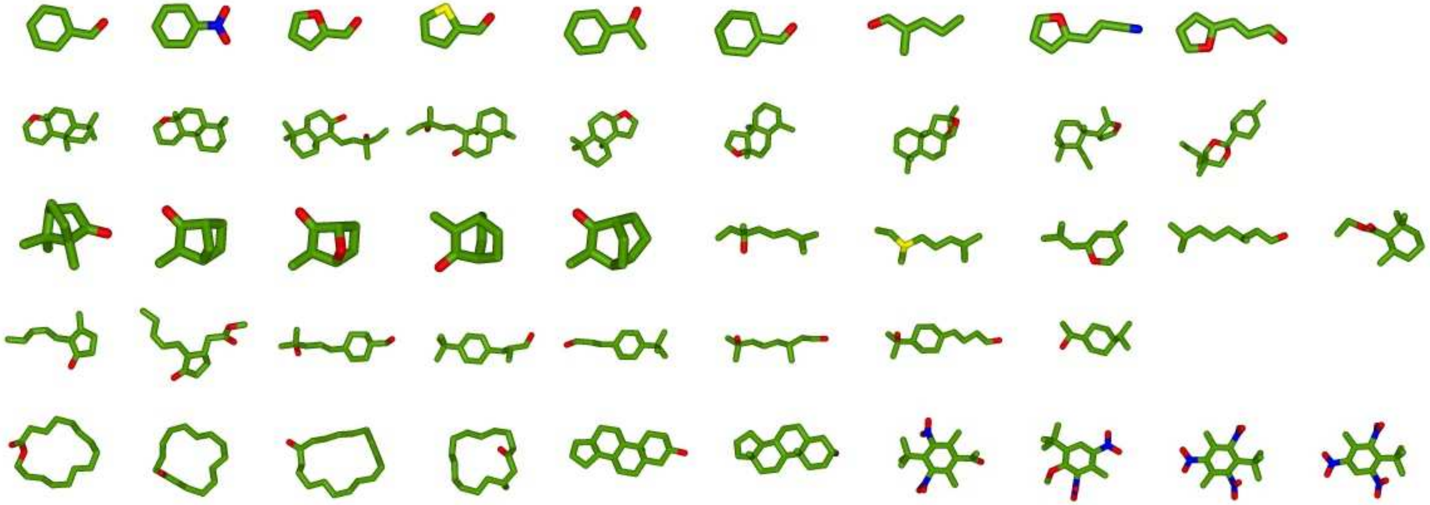
- Refine with $16 \times 16 \times 16$ grid of 1 degree steps (gives about 50 molecules / second)
- Can also pre-process a set of molecules by aligning them to the principal axes

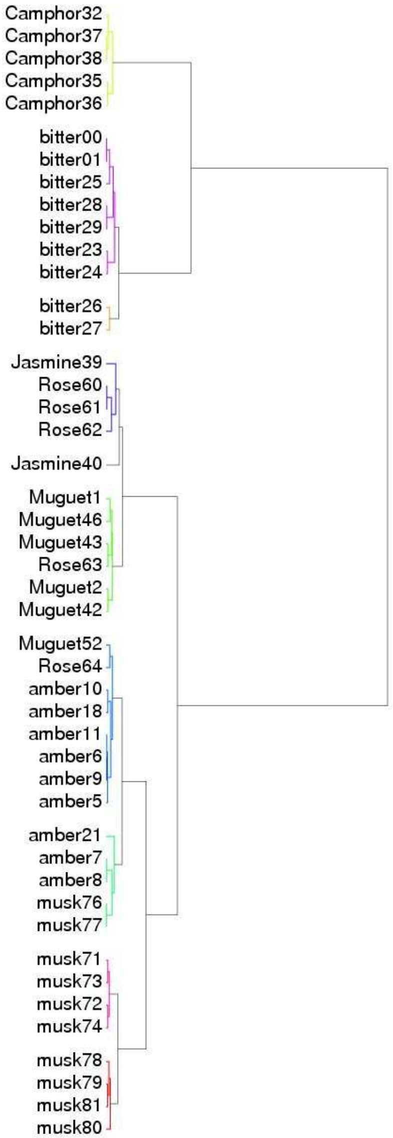


- Pre-aligned "canonical" orientations of similar molecules often overlay very well ...

Clustering the Odour Dataset using 2D SH Surface Shapes

(Takane et al. (2004) Org. Biomol. Chem. 2 3250–3255)



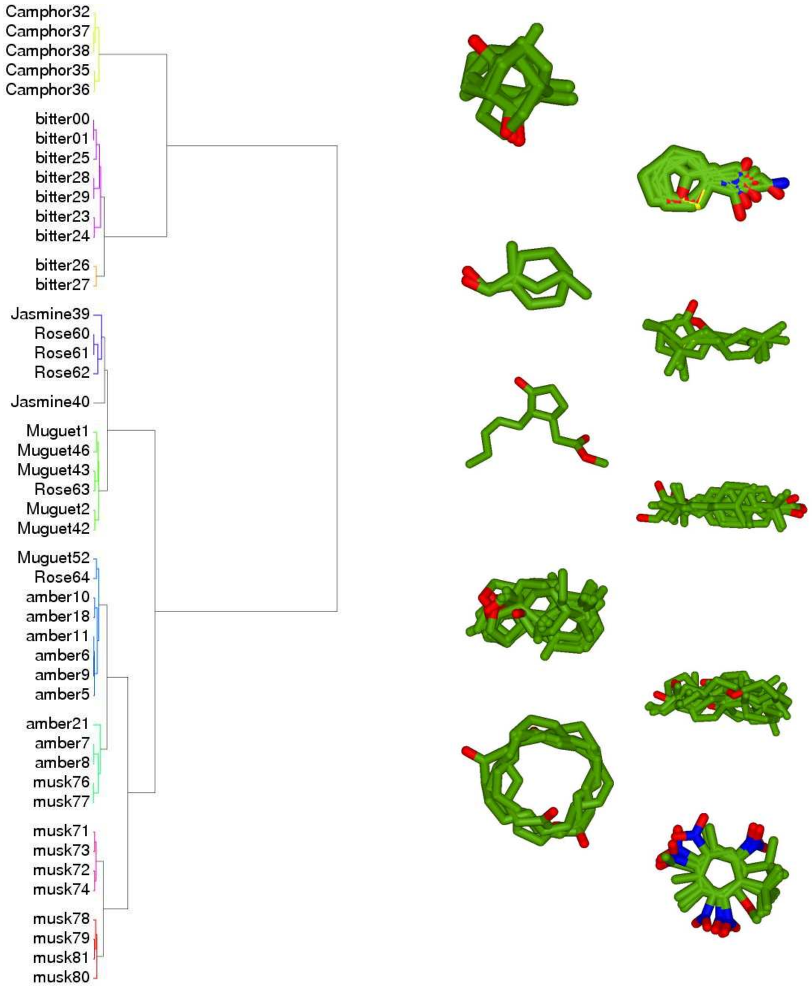


- **Seven classes: bitter, ambergris, camphoraceous, rose, jasmine, muguet, musk**
- **Following Takene et al., we clustered into 10 group using ParaSurf & Parafit:**

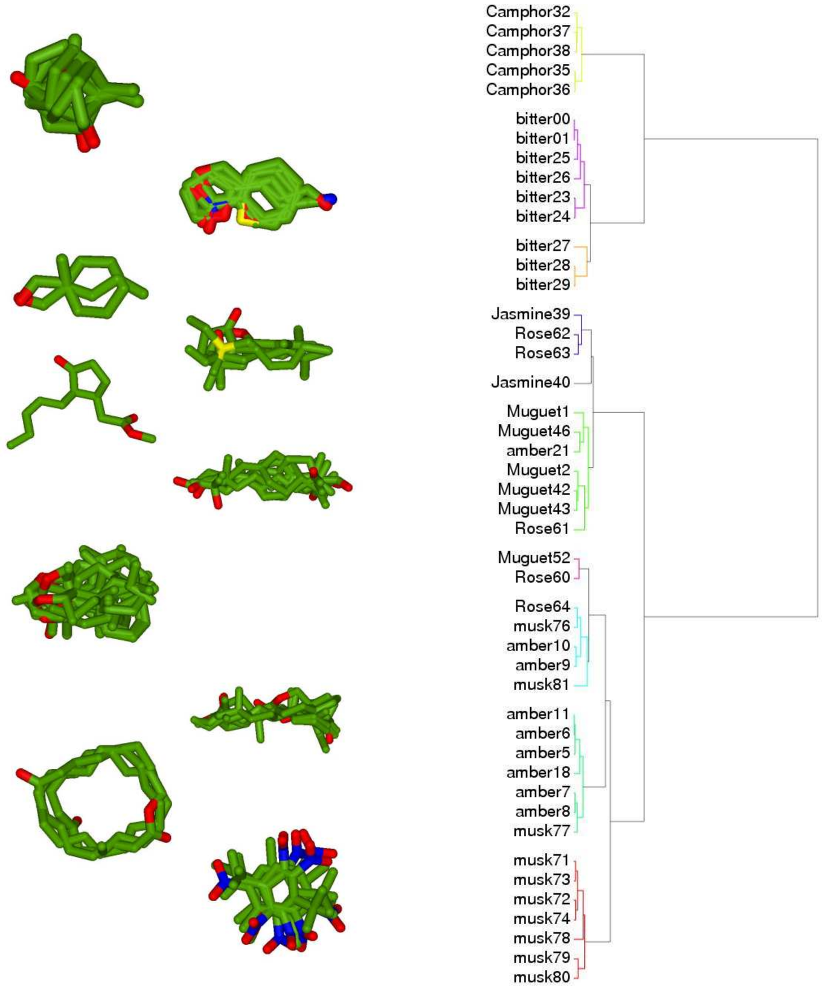
```
unix% PS_mopac_run

unix% PS_parasurf_run

unix% parafit -matrix -dif odour_data.dif *_p.sdf

unix% dif2jpg -n 10 -d odour_data.dif

unix% eog odour_data.jpg
```

Visualising The Odour Dataset Clustering Results

Clustering Superposed Pairs



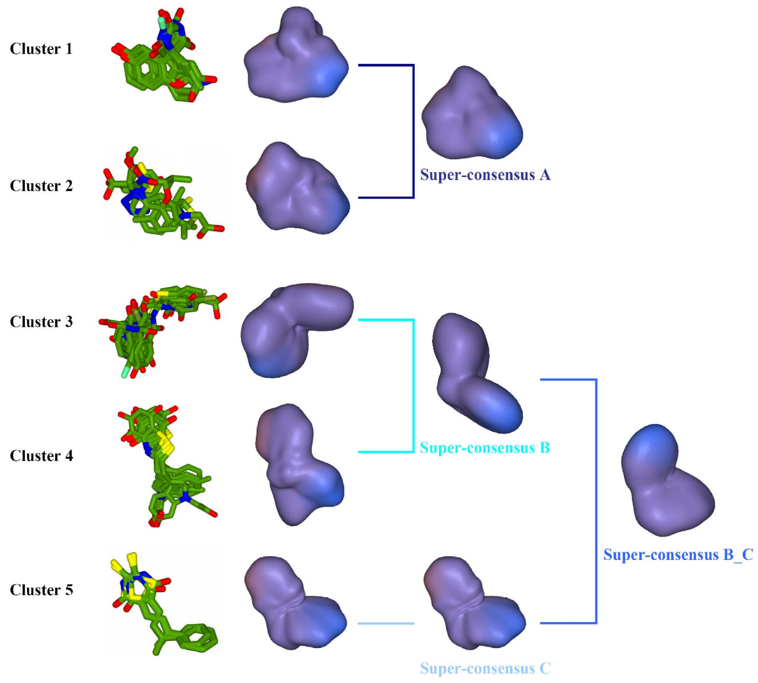
Clustering Canonical Orientations



Mavridis et al. (2007), J Chemical Information & Modeling, 45, 1787–1796

Promiscuous Protein Targets Seem to be Rather Common

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



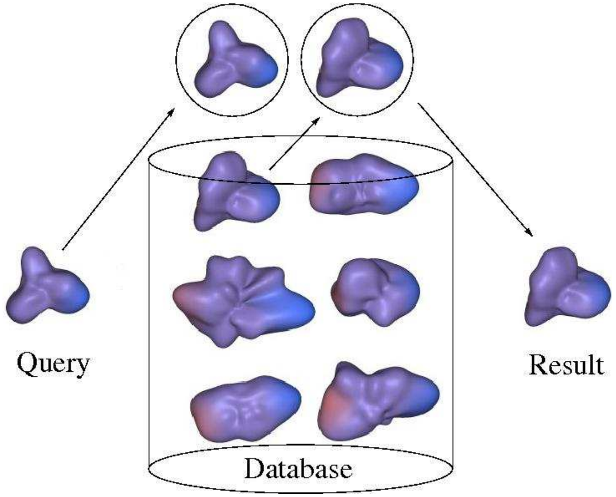
- Several other promiscuous targets in the literature:
  - the $\alpha 1\beta 1$ and $\alpha 2\beta 1$ integrins,
  - factor H, LRP6, PPAR-γ, LXR-β,
  - ACHE, P38, FXA, VEGFR2, PXR,
  - β-secretase, thrombin, CDK2,
  - LAIR-1, LAIR-2, LTBLP-2, NS2B-NS3.

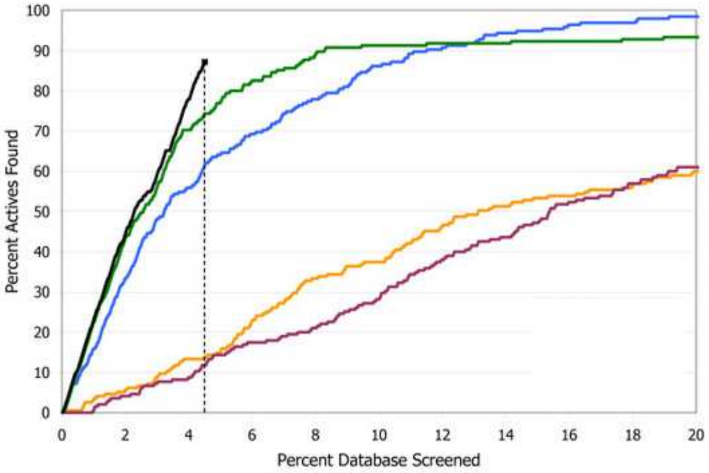
- For ligand-based virtual screening, these examples suggest:
  - cluster the 3D shapes of any known ligands before performing VS ...
  - compare shape-based VS performance with and without clustering ...
  - ... any large differences could suggest a promiscuous (multi-site?) substrate.

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

Ligand-Based VS (LBVS) Principals

- LBVS aims to find new actives by similarity to one or more existing actives
- Usually LBVS has two phases – retrospective (i.e. "training mode"), and prospective
- Main purpose of retrospective VS is to find the best algorithm + query (molecule/conformation)
    - Prerequisites: some know actives + a good set of decoys (similar mol wt, chemistry)
- Historically, enrichment plots have been popular for analysing ranked lists of VS results



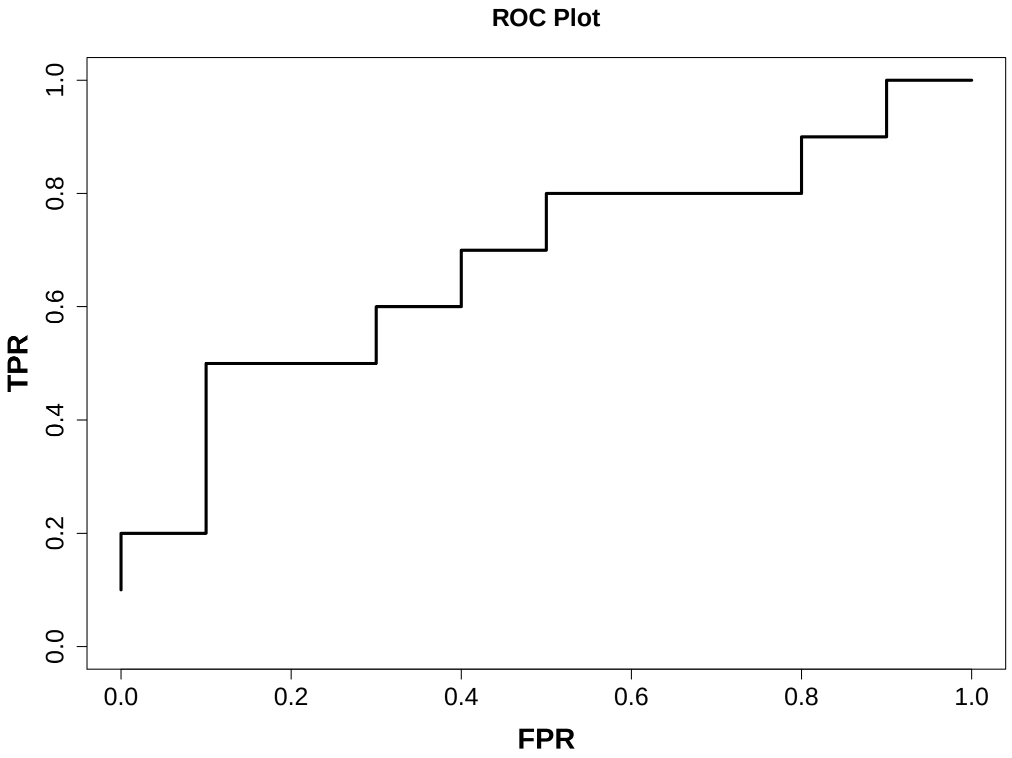


- Disadvantages:
    - enrichment plots (or enrichment factors) depend on the no. of actives
    - it is difficult to compare different enrichment plots

Receiver-Operator-Characteristic (ROC) Plots

- ROC plots show the ability of a classifier to distinguish postive and negative instances
  - A ROC plot shows the true positive rate (TPR) against the false positive rate (FPR)
  - Suppose 10 positive and 10 negative instances have been scored by a classifier...

```
Score Class Npos Nneg  TPR  FPR
----- ----- ---  ---   ---  ---
0.90  pos    1    0    0.1  0.0
0.80  pos    2    0    0.2  0.0
0.70  neg    2    1    0.2  0.1
0.60  pos    3    1    0.3  0.1
0.55  pos    4    1    0.4  0.1
0.54  pos    5    1    0.5  0.1
0.53  neg    5    2    0.5  0.2
0.52  neg    5    3    0.5  0.3
0.51  pos    6    3    0.6  0.3
0.50  neg    6    4    0.6  0.4
0.40  pos    7    4    0.7  0.4
0.39  neg    7    5    0.7  0.5
0.38  pos    8    5    0.8  0.5
0.37  neg    8    6    0.8  0.6
0.36  neg    8    7    0.8  0.7
0.35  neg    8    8    0.8  0.8
0.34  pos    9    8    0.9  0.8
0.33  neg    9    9    0.9  0.9
0.32  pos   10    9    1.0  0.9
0.31  neg   10   10    1.0  1.0
```



- The area under the curve (AUC) gives a good overall measure of classifier performance
- A random classifier gives a diagonal line: TPR=FPR (AUC=0.5)
- A perfect classifier gives TPR=1.0 for all FPR (AUC=1.0)

Fawcett (2006) Patttern Recognition Letters 27, 861–874; en.wikipedia.org/wiki/Receiver_operating_characteristic

Several Other Common VS Quality Measures

- Suppose there are n actives in a total of N molecules, and the scoring function is used to produce a ranked list of molecules: $i = 1, 2, 3, ...N$.

- Often we are most interested in the quality of the top (e.g. top 1%) of the ranked list

  - Enrichment Factor: $EF_{x\%} = \frac{n_a/N_{x\%}}{n/N}$

  - ROC AUC: $AUC = \frac{1}{n}\sum_{i=1}^{n}(1 - f_i)$

  - ROC $AUC_{x\%}$: $AUC_{x\%} =$ calculate graphically

  - Balanced ROC: $BAROC = \frac{1}{n}\sum_{i=1}^{n} e^{-\alpha f_i}$

  - Sum of Logs of Rank: $SLR = -\sum_{i=1}^{n} log\left(\frac{r_i}{N}\right)$

  - Normalised SLR: $NSLR = \sum_{i=1}^{n} log\left(\frac{r_i}{N}\right) / \sum_{i=1}^{n} log\left(\frac{i}{N}\right)$

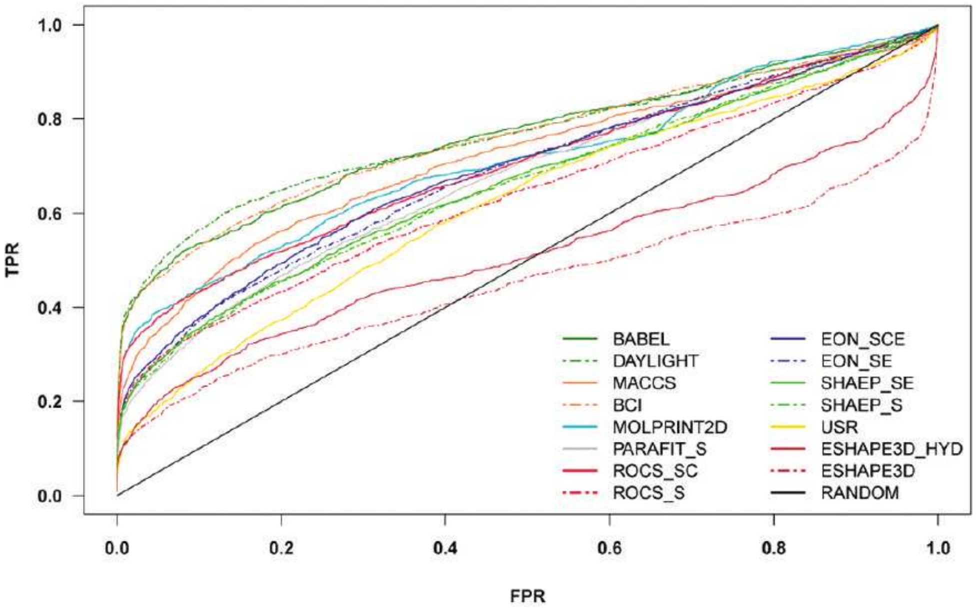
  (here, the top $x\%$ of the list contains $N_{x\%}$ molecules and n_a actives, r_i is the rank of the i^{th} active, and f_i is the fraction of inactives ranked higher than i)

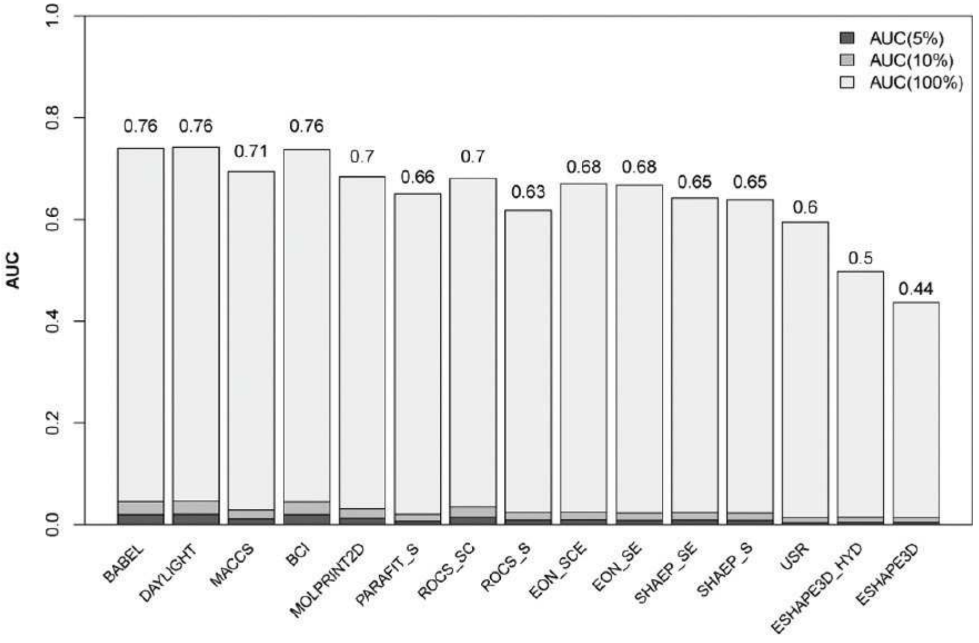
- Which is best? Debatable! These days, ROC AUC, $AUC_{5\%}$, $AUC_{10\%}$ are quite popular...

Venkatraman et al. (2010), J Chemical Information & Modeling, 50, 2079–2093

Comparing Ligand-Based Virtual Screening Methods

- We calculated aggregate ROC plots to compare several VS methods on the "DUD" dataset
  - DUD = Directory of Useful Decoys – http://dud.docking.org/ – 40 targets, 100K decoys
  - 2D methods = Babel, Daylight, MACCS, MCI, Molprint2D
  - 3D methods = Parafit, ROCS, SHAEP, USR, Eshape3D





- The fingerprint methods perform remarkably well (!)
- Suggests need to improve 3D methods – better query conformations ? shape clustering ?

Venkatraman et al. (2010), J Chemical Information & Modeling, 50, 2079–2093

Conclusions

- Modeling flexibility during docking is still a major challenge
- Cross-docking can detect protein-protein partners remarkably often

- 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

- Fast 3D virtual screening algorithms are becoming available
- All-vs-all 3D protein docking and ligand shape-matching now feasible ?
- Choosing a good query conformation still a challenge in ligand-based VS

Acknowledgments

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

BBSRC, EPSRC, ANR

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

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