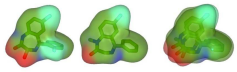
ParaFit



Clustering and Classifying HIV Entry Inhibitors

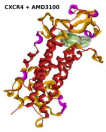
David Ritchie
INRIA, Nancy Grant Est

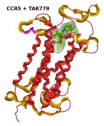
Violeta Pérez-Nueno
Institut Chimíque de Sarià

PERSONA CIENCIA EMPRESA
Universitat Ramon Llull

ESCUELA TÉCNICA SUPERIOR

Presentation Overview

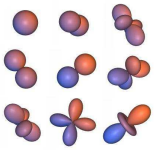




1. Introducing ParaFit
2. Summary of spherical harmonics plus SH clustering example
3. SH-based retrospective virtual screening of CXCR4 and CCR5 co-receptors
4. Introducing SH "consensus shapes"
5. Analysing CCR5 ligands and binding sub-sites using SH consensus shape clustering
6. Restrospective VS Results on the Berlex Dataset

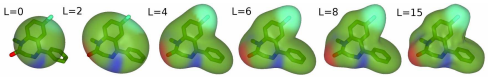
Spherical Harmonic Surfaces

- Use SHs as "building blocks," i.e. components of shape, etc.



- Real SHs: $y_{lm}(\theta,\phi)$
- Coefficients: a_{lm}
- Encode radial distances from origin as SH series…
- Solve coefficients by numerical integration…

$$r(\theta,\phi) = \sum_{l=0}^{15} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\phi)$$

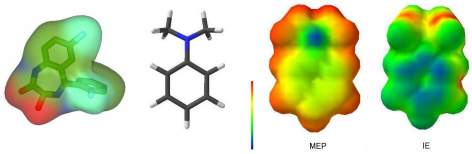


Ritchie, D.W. and Kemp, G.J.L. *J. Comp. Chem.* **1999**, *20*, 383–395.

ParaSurf – Quick Reminder

Shapes/Properties From Semi-Empirical QM

- From MOPAC or VAMP, calculate:
  - Density contours of 2x10^{-4}e/A^3 (i.e. approx = SAS)
  - MEP – electrostatic potential
  - IE_L – ionization energy
  - EA_L – electron affinity
  - α_L – polarizability
- Encode as Spherical Harmonic expansions to order L=15…



Lin, J.-H. and Clark, T. *J. Chem. Inf. Model.* **2005**, *45*, 1010–1016.

ParaFit – The Main Features

- Command-line program
- Available for Linux, Windows, SGI, etc.

- Reads and writes ParaSurf SDF files
- Superposes and compares SH molecular surfaces
- Works with other ParaSurf properties (+ combinations)
- Works with multi-molecule SDFs

- Four main operating modes:
  - Fitting
  - Matrix (all v's all fitting)
  - Canonical
  - Consensus

ParaFit – The Theory

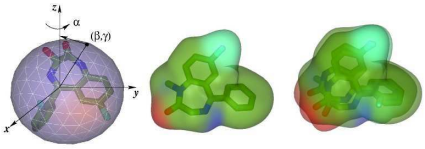
Mathematical Machinery

- **Distance:** $D = \int (r_A(\theta,\phi) - r_B(\theta,\phi)')^2 \mathrm{d}\Omega$
- **Orthogonality:** $D = |\underline{a}|^2 + |\underline{b}|^2 - 2\underline{a}.\underline{b}'$
- **Rotation:** $b'_{lm} = \sum_{m'} R^{(l)}_{mm'}(\alpha,\beta,\gamma) b_{lm'}$
- **Carbo:** $S = \underline{a}.\underline{b}'/(|\underline{a}|.|\underline{b}|)$
- **Hodgkin:** $S = 2\underline{a}.\underline{b}'/(|\underline{a}|^2 + |\underline{b}|^2)$
- **Tanimoto:** $S = \underline{a}.\underline{b}'/(\underline{a}|^2 + |\underline{b}|^2 - \underline{a}.\underline{b}')$
- **Multi-property:** $Q = pS + qS^{\mathrm{MEP}} + rS^{\mathrm{IE_L}} + \ldots$

Ritchie, D.W. and Kemp, G.J.L. *J. Comp. Chem.* **1999**, *20*, 383–395.

ParaFit Superposition Searches

- Uses icosahedral tessellation of sphere for Euler rotations



- Samples 22,000 orientations of about 8 degree steps
  - Refine with a 16x16x16 grid of 1 degree steps
- Approx 20 pair-wise superpositions/sec on 1.8GHz Xeon PC
- Rotates everything from ParaSurf SDF file –
  - SH coefficients, dipole, quadrupole, moments, etc.,
  - density matrix elements, NAO-PCs, etc.

Using ParaFit – Fitting Mode

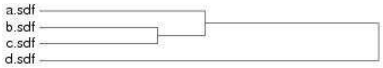
- One "reference" molecule, multiple moving molecules (equivalently: compare a query against a database)
- `unix% parafit -fit a.sdf b.sdf c.sdf`
  - creates `b_a.sdf c_a.sdf` (**b** in frame of **a**), etc.
  - `b.sdf, c.sdf` may be multi-molecule SDFs
- Output files contain rotated:
  - atom coordinates
  - dipole, quadrupole, octupole moments
  - NAO-PCs and density matrix elements
- Optimisation:
  - internally rotated **a** is compared against fixed **b, c, ...**
  - this gives about a 5-fold speed up
  - can achieve up to about 100 superpositions/second

Using ParaFit – Matrix Mode

- Matrix mode = all-versus-all fitting
  - useful for clustering, etc.
- `unix% parafit -matrix a.sdf b.sdf c.sdf d.sdf`
  - creates `b_a, c_a, d_a, a_b, c_b, d_b, etc.`
  - can suppress creation of output files with `-nosdf`
- `unix% cat parafit.pft`

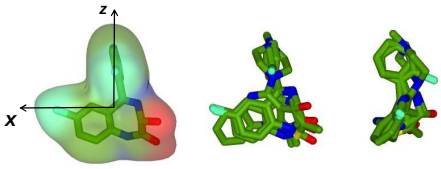
```
0.9974  c.sdf  b.sdf
0.9921  c.sdf  a.sdf
0.9917  b.sdf  a.sdf
```

- `unix% dif2jpg –d parafit.dif –o parafit.jpg`



Using ParaFit – Canonical Mode

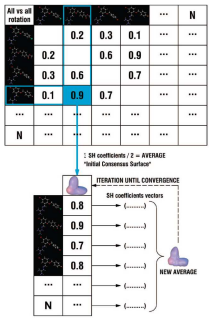
- Canonical mode = align molecules to coordinate axes
- Useful for visualisation (almost as good as fitting)
  - Similar to finding moments of intertia
  - But no ambiguity with respect to 180 degree flips
- `unix% parafit -canonical a.sdf b.sdf c.sdf d.sdf`



- Canonical mode is often almost as good as fitting

Using ParaFit – Consensus Mode

- `unix% parafit –consensus a.sdf b.sdf c.sdf ...`

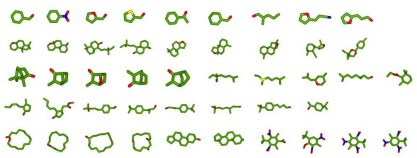


1. Do all-v-all SH comparison
2. Find best pair-wise match
3. Calculate SH average of pair
4. Treat average as new seed
5. Superpose all onto seed
6. Compute new average seed
7. Rotate all onto new seed
8. Iterate until convergence...
9. Result = SH pseudo-molecule

Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

ParaFit Clustering Example

- Takane *et al.* collected 47 odour molecules: in 7 classes:
- bitter, ambergris, jasmine camphor, rose, muguet, musk



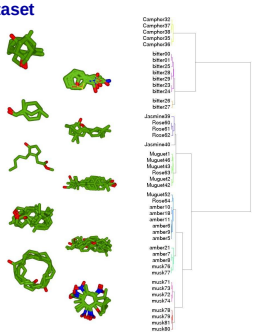
- Takane *et al.* clustered into 10 groups using eigenvector analysis of QM vibrational frequencies…

Takane S. and Mitchell J.B.O. *Org. Biomol. Chem.* **2004**, *2*, 3250–3255.

ParaFit Clustering Example

- **Clustering the Odour Dataset**
  - **Calculate SH shapes using ParaSurf, and cluster with ParaFit:**

```
unix% PS_mopac_run
unix% PS_Parasurf_run
unix% parafit –matrix –dif \
          o.dif *_p.psf
unix% dif2jpg –n10 o.dif
unix% eog o.jpg
```

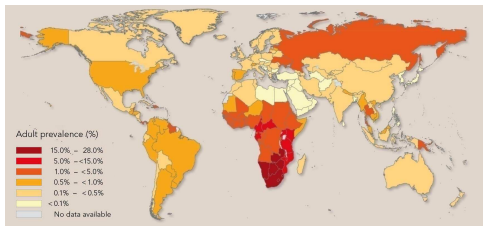


  - **Clustering SH shapes gives better clusters than using vibrational frequencies…**

Mavridis L., *Hudson B., Ritchie, D.W., J. Chem. Inf. Model.* **2007**, *45*, 1787–1796.

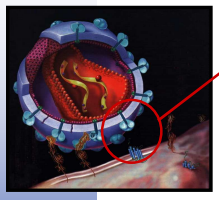
HIV and HIV Entry Inhibitors

A	Acquired	Group of symptoms and signs
I	Immune	Immunitary system
D	Deficiency	Weakening and/or destruction
S	Syndrome	It is not a hereditary disease

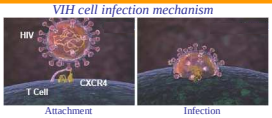


Number of people living with HIV in 2007	Total: 33,0 million (30–36)
People newly infected with HIV in 2007	Total: 2,7 million (2,2–3,2)
AIDS deaths in 2007	Total: 2,0 million (1,8–2,3)

HIV Cell Entry Mechanisms



VIH cell infection mechanism



VIH entry inhibition mechanism

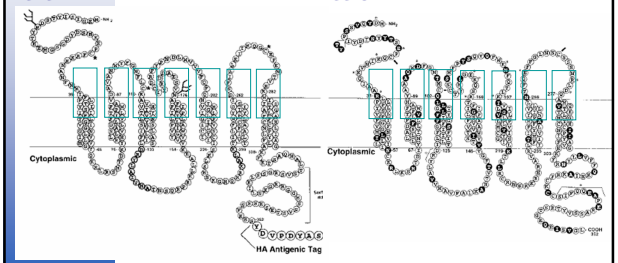
Target	*Mechanism*
CD4 (cell)	Block CD4 binding by gp120
gp120 (virus)	Block gp120 conformational changes needed to interact with the chemokine receptor
CCR5, CXCR4 (cell)	Block chemokine receptor binding by gp120
gp41 (virus)	Block gp41 structural changes needed for fusion
Membrane (cell or virus)	Block lipid bi-layer destabilization and mixing

Shaheen, F.; Collman, R.G. *Curr. Opin. Infect. Dis.* **2004**, *17*, 7–16.

Targeting the CXCR4 and CCR5 Co-Receptors

- **CXCR4 and CCR5 are members of the GPCR family**
- **We modelled them using bovine rhodopsin as template**

CXCR4 *CCR5*



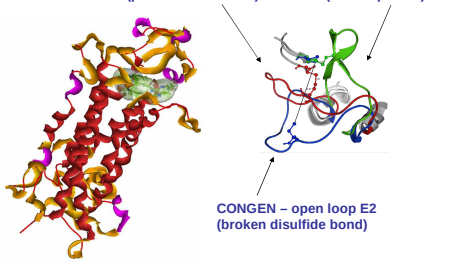
Cabrera, C. *et al. AIDS Res. Hum. Retrovir.* **1999**, *15*, 1535–1543.
Berson, J.F. *et al. J. Virol.* **2000**, *10*, 255–277.

Homology Modelling CXCR4/CCR5

- **The Co-receptor structures were built using Modeller**
- **But loop E2 was built with CONGEN + disulphide constraints**

CONGEN – open loop E2 (preserves disulfide)

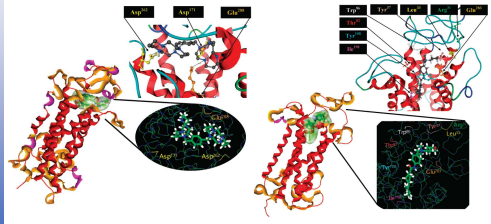
MODELLER – loop E2 (blocks pocket)



CONGEN – open loop E2 (broken disulfide bond)

Validating the Receptor Model Structures

- **The receptor models were validated by docking selected high-affinity ligands: AMD3100 (CXCR4) and TAK779 (CCR5)**



- **The binding modes from Autodock were consistent with the available SDM evidence on key ligand-binding residues**

Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

Virtual Screening Datasets

CCR5 Antagonists (424):

1) SCH-C derivatives
2) 1,3,5-trisubstituted pentacyclics
3) Diketopiperazines
4) 1,3,4-trisubstituted pyrrolidinepiperidines
5) 5-oxopyrrolidine-3-carboxamides
6) *N,N'*-Diphenylureas
7) 4-aminopiperidine or tropanes
8) 4-piperidines
9) TAK derivatives
10) Guanylhydrazone drivatives
11) 4-hydroxypiperidine derivatives
12) Phenylcyclohexilamines
13) Anilide piperidine N-oxides
14) 1-phenyl-1,3-propanodiamines
15) AMD derivatives
16) Other

CXCR4 antagonists (248):

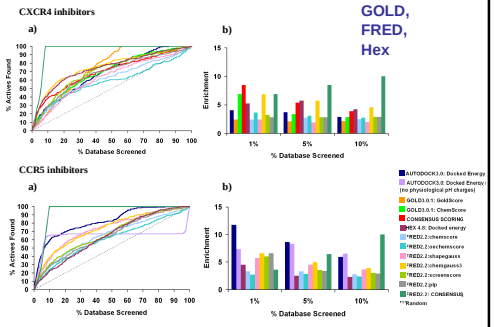
1) AMD derivatives
2) Macrocycles
3) Tetrahydroquinolinamines
4) KRH derivatives
5) Dipicolil amine zinc(II) complexes
6) Other

PLUS…

4696 inactive compounds from the Maybridge Screening Collection with similar 1D properties to the actives

Receptor-Based VS Enrichment Results

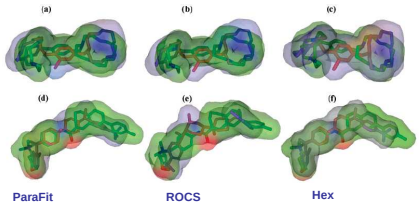
- Each ligand was docked and ranked using: Autodock, GOLD, FRED, Hex



Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

SH Ligand-Based VS Set-Up

- Each database compound was scored against the docked conformation of AMD3100 (CXCR4) and TAK779 (CCR5)

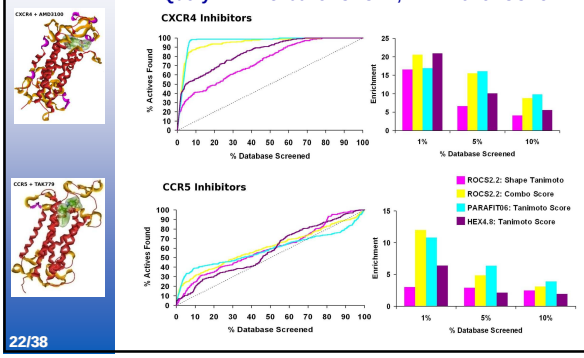


- This example shows the superpositions of (top) AMD3167 (blue), and (bottom) SCH417690) with the given queries
- NB. The database conformations were calculated by MOE FlexAlign… ROCS used Omega for 10 further conf.s

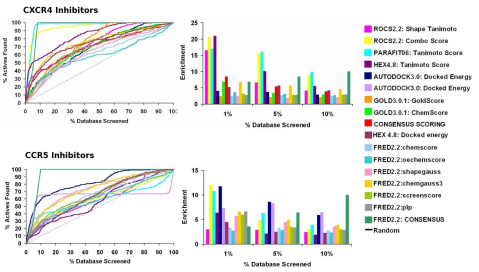
Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

SH Ligand-Based VS Enrichment Results

- Query = AMD3100 for CXCR4; TAK779 for CCR5



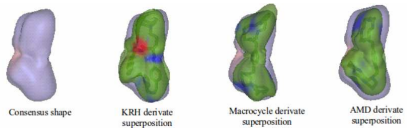
Comparing Ligand-Based and Receptor-Based VS



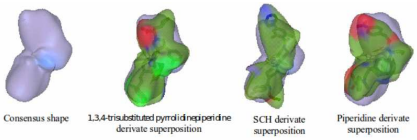
- Docking enrichments are better for CXCR4 than CCR5
- But shape-based scoring gives better overall enrichments

SH Consensus Shapes of the Three Most Active Inhibitors

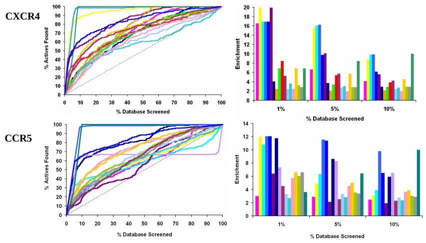
CXCR4



CCR5

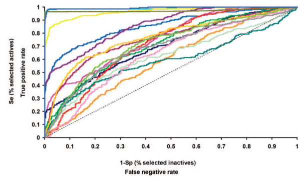


Consensus Shape-Based VS



Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

Overall Results – CXCR4



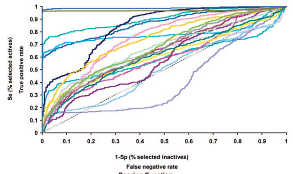
Scoring Function	AUC
PARAFIT08: Consensus 3 queries Shape Tanimoto	0.994
PARAFIT08: Shape Tanimoto	0.992
FRED2.2: CONSENSUS Chemscore, Oechemscore, Shapegauss, Chemgauss3, Screenscore, Plp	0.974
ROCS2.2: Combo Score (Shape Tanimoto + Scaled Color Score)	0.966
HEX4.8: Shape Tanimoto	0.862
PARAFIT08: Consensus all compounds queries Shape Tanimoto	0.857
FRED2.2: Chemgauss3	0.768
ROCS2.2: Shape Tanimoto	0.761
HEX4.8: Docked Energy	0.756
FRED2.2: Screenscore	0.717
CONSENSUS SCORING: Autodock Docked Energy, Gold GoldScore, Gold ChemScore	0.717
GOLD3.0.1: ChemScore	0.712
AUTODOCK3.0: Docked Energy	0.699
FRED2.2: Plp	0.687
FRED2.2: Shapegauss	0.637
FRED2.2: Chemscore	0.623
GOLD3.0.1: GoldScore	0.595
FRED2.2: Ochemscore	0.588

Best scorers:

- **ParaFit 3-Consensus**
- **ParaFit Tanimoto**
- **Fred Consensus**
- **ROCS Combo**

Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

Overall Results – CCR5



Scoring Function	AUC
PARAFIT08: Consensus 3 queries Shape Tanimoto	0.991
FRED2.2: CONSENSUS Chemscore, Oechemscore, Shapegauss, Chemgauss3, Screenscore, Plp	0.971
PARAFIT08: Superconsensus C Shape Tanimoto	0.905
PARAFIT08: Consensus all cmpds queries Shape Tanimoto	0.866
AUTODOCK3.0: Docked Energy	0.843
PARAFIT08: Superconsensus A Shape Tanimoto	0.785
FRED2.2: Shapegauss	0.750
FRED2.2: Chemgauss3	0.747
FRED2.2: Screenscore	0.683
FRED2.2: Chemscore	0.660
FRED2.2: Plp	0.659
HEX4.8: Docked Energy	0.653
PARAFIT08: Superconsensus D Shape Tanimoto	0.630
ROCS2.2: Combo Score (Shape Tanimoto + Scaled Color Score)	0.615
FRED2.2: Ochemscore	0.611
ROCS2.2: Shape Tanimoto	0.600
PARAFIT08: Shape Tanimoto	0.594
HEX4.8: Shape Tanimoto	0.581
PARAFIT08: Superconsensus B+D Shape Tanimoto	0.505
PARAFIT08: Superconsensus B Shape Tanimoto	0.413

Best scorers:

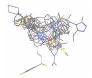
- **ParaFit 3-Consensus**
- **FRED Consensus**
- **ParaFit S-Consensus**

Pérez-Nueno *et al. J. Chem. Inf. Model.* **2008**, *48*, 2146–2165.

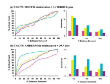
Experimental Evidence for Multiple CCR5 Binding Sites

There is strong evidence that there are multiple sub-sites within the CCR5 extracellular pocket:

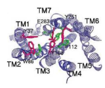
▪It is very difficult to superpose all the different families of CCR5 active compounds.



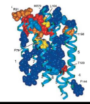
▪VS enrichment results are strongly dependent on the conformation of the query molecule.



▪Site directed mutagenesis evidence suggests a large pocket (the SDM residues are spatially well distributed around the pocket).



▪Not all SDM locations affect the binding of all ligands.



Exploring the CCR5 Multiple Binding Site Hypothesis

- **There is a hypothesis that *the CCR5 ligands form two or more groups, i.e., they have two or more binding modes…***

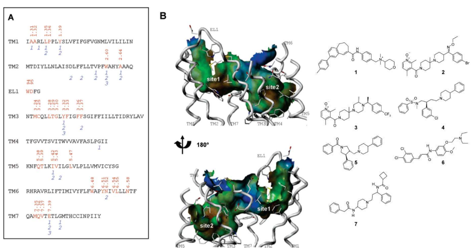


Figure 2. CCR5 TM binding cavity. A. Amino acid sequence. Residues with side chain pointing toward the cavity are written in red and pointed using Ballesteros numbering,[61] except in the extracellular domain 1 (EL1). Numbers indicated underneath the sequence summarize experimental mapping of receptor interaction site for nonpeptide antagonists **1**−**5**; *1*: residues important for the efficiency of **1**,[22,23,24] *2*: residues important for the efficiency of **2** and Schering-Plough compound AD101,[23−25] and *3*: residues important for the binding **4** and **5**.[26] B. The Connolly surface of the CCR5 receptor cavity (colored according to the lipophilic potential) is displayed together with the ribbon diagram of the seven TM helices. Side chains of key residues highlighted in the sequence are depicted using line representation. The bottom view is rotated about a vertical axis by 180° relative to top view.

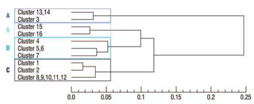
Kellenberg *et al. J. Med. Chem.* **2007,** *50,* 1294-1303.

Clustering the 424 CCR5 Ligands

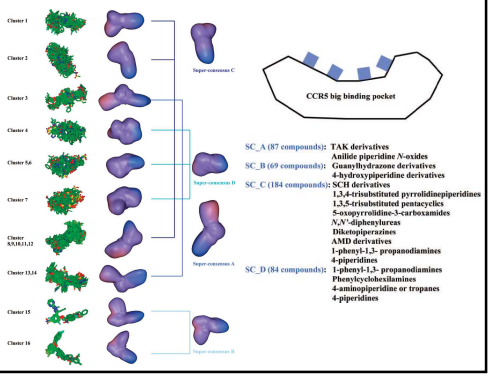
- Because it is not clear *a priori* which ligands might belong to which group, we first performed Wards hierarchical clustering of chemical fingerprints…
- We then used Kelley's method to find the optimal number of clusters (16)
- These were manually merged to 10 groups based on known CCR5 families

CLUSTER	*Compounds Found*	*Number of compounds*	*Consensus Shape*
1	*(8) 1,3,4-trisubstituted pyrrolidinepiperidines* *(3) 1,3,5-trisubstituted pentacyclics* ***(5) 5-oxopyrrolidine-3-carboxamides*** ***(4) N,N'-diphenylureas*** (2) TAK derivatives (1) 4-piperidines (1) others (MRK-1 CMPD 167)	24	
2	*(1) 1,3,4-trisubstituted pyrrolidinepiperidines* *(6) 1,3,5-trisubstituted pentacyclics* (13) 1-phenyl-1,3- propanodiamines (3) 4-piperidines ***(3) AMD derivatives*** ***(9) Diketopiperazines*** (1) SCH derivatives (2) Phenylcyclohexilamines (3) others (GSK, Merck2, Merck3)	41	
3	***(22) Anilide piperidine N-oxides*** (1) TAK derivatives (1) others (1-benzazepine)	24	
4	(21) 1-phenyl-1,3- propanodiamines (5) Phenylcyclohexilamines	26	
5	(11) 1-phenyl-1,3- propanodiamines	11	
6	(12) 1-phenyl-1,3- propanodiamines	12	
7	***(26) 4-aminopiperidine or tropanes*** (6) 4-piperidines (2) Phenylcyclohexilamines (1) others (Merck1)	35	
8	(23) SCH derivatives	23	
9	(20) SCH derivatives	20	
10	(37) SCH derivatives	37	
11	(22) SCH derivatives	22	
12	(17) SCH derivatives	17	
13	(19) TAK derivatives	19	
14	(44) TAK derivatives	44	
15	***(33) Guanylhydrazone derivatives***	33	
16	***(36) 4-hydroxypiperidine derivatives***	36	

- SH consensus shapes were calculated for the 10 groups
- These were then compared in ParaFit (all-vs-all)
- Another round of Ward's clustering proposed four super-consensus clusters

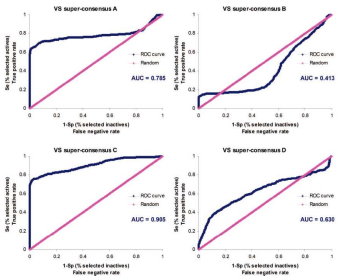


From Consensus Shapes to Super-Consensus Clusters



Using Super-Consensus Shapes as VS Queries

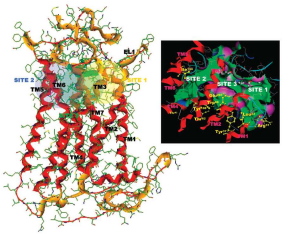
- Each SC pseudo-molecule was used as a VS query:



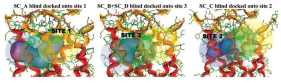
- NB. merging SC shapes significantly worsens the AUCs…
- SC queries => CCR5 ligands form no less than FOUR groups

Hex Blind Docking of SC Pseudo-Molecules to CCR5

- 3D pseudo-molecules were created as the union of all superposed ligands in each SC family for docking in Hex



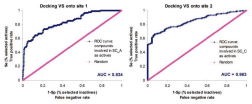
- SC-A docks to Site-1 (TMs 1, 2, 3, 7)
- SC-C docks to Site-2 (TMs 3, 5, 6)
- B and D dock to Site-3 (TMs 3, 6, 7)



Autodock Docking VS w.r.t. Three CCR5 Sub-Sites

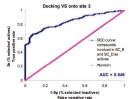
- To confirm the SC shapes were matched to their predicted target sites, docking based VS was repeated for each ligand using:
- SC-As treated as actives for Site 1 (SCs B, C, D treated as inactives)
- SC-Cs treated as actives for Site 2 (SCs A, B, D treated as inactives)
- SC-B/Ds assumed active for Site 3 (SCs A and C treated as inactives)

A -> Site-1



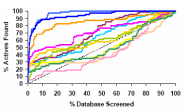
C -> Site-2

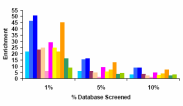
B,D -> Site-3



- As before, merging SCs worsens the AUCs…
- SC docking => no less than THREE CCR5 pocket sub-sites

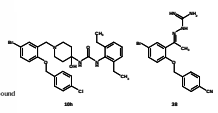
Screening the Berlex Dataset





PARAFIT06: Shape Tanimoto
PARAFIT06: Shape consensus Tanimoto (3 active compounds)
PARAFIT06: Shape consensus Tanimoto (all active compounds)
HEX 4.0: Shape Tanimoto
ROCS 2.2: Shape Tanimoto
ROCS 2.2: Shape Tanimoto 10 conf/database compound
ROCS 2.2: Shape Tanimoto 10 query conf & 10 conf/database compound
ROCS 2.2: Combo score (shape + chemistry)
ROCS 2.2: Combo score (shape + chemistry) 10 conf/database compound
ROCS 2.2: Combo score (shape + chemistry) 10 query conf & 10 conf/database compound
QIKPROP/QIKSIM (one query descriptors): Tanimoto Score
QIKPROP/QIKSIM (actives averaged descriptors): Tanimoto Score

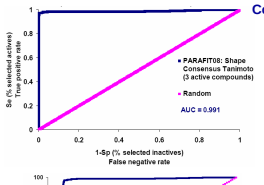
- Berlex Science recently synthesised 69 guanyl-hydrozone and 4-piperidine-hydrazone derivatives which showed activity as CCR5 antagonists
- We performed retrospective VS against 3388 decoys from Maybridge Screening Collection, with similar 1D properties to the actives using:
- One high affinity query
- Consensus of the 3 most active
- Consensus of all actives...



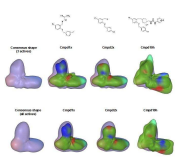
Carrieri *et al. ChemMedChem* **2009**, 4(7), 1153-1163.

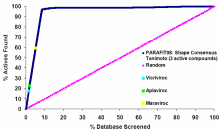
CCR5 VS with Berlex Dataset

- Using Berlex actives as queries to previous 424/4696 dataset:



Consensus of top 3 Berlex actives





Maraviroc = 1-phenyl-1,3-propanodiamine
Vicriviroc = SCH417690 (Schering Plough)
Aplaviroc= diketopiperazine

Conclusions

- SH surfaces allow fast comparison and clustering
  - SH-based clustering of Odour dataset superior to EVA clustering
- Our models of CXCR4 and CCR5 are consistent with SDM
- We built a VS library of 248 CXCR4 and 424 CCR5 inhibitors
- Ligand-based VS gives better enrichments than docking
- ParaFit and ROCS give the best overall VS enrichments
- Docking & SH-based VS results for CXCR4 better than CCR5
  - CXCR4 has smaller pocket and fewer ligands than CCR5
- Consensus clustering of CCR5 ligands -> FOUR super-families
- Docking CCR5 SC pseudo-molecules -> THREE sub-sites
- Good retrospective VS results on the Berlex actives

Acknowledgments

- Violeta Pérez-Nueno
- Antonio Carrieri
- Lazaros Mavridis
- Brian Hudson
- Vishwesh Venkatraman

- EPSRC
- University of Aberdeen
- IQS, Universitat Ramon-Llull

Papers: http://www.loria.fr/~ritchied/

ParaSurf + ParaFit: http://www.ceposinsilico.de/