Clustering and Classifying HIV Entry Inhibitors

David Ritchie
INRIA, Nancy Grant Est

Violeta Pérez-Nueno
Institut Chimique de Sariá

Presentation Overview
1. Introducing ParaFit
2. Summary of spherical harmonics plus SH clustering example
3. SH-based retrospective virtual screening of CXC4 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: \( y_l^{lm}(\theta, \phi) \)
  - Coefficients: \( a_{lm} \)
  - Encode radial distances from origin as SH series...
  - Solve coefficients by numerical integration...

ParaSurf – Quick Reminder
- From MOPAC or VAMP, calculate:
  - Density contours of \( 2 \times 10^{-4} \) e/A
  - MEP – electrostatic potential
  - IE – ionization energy
  - EA – electron affinity
  - \( \alpha \) – polarizability
- Encode as Spherical Harmonic expansions to order \( L=15 \)...

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 (s combinations)
- Works with multi-molecule SDFs
- Four main operating modes:
  - Fitting
  - Canonical
  - Consensus

Mathematical Machinery
- Distance: \( D = \int (r_A(\theta, \phi) - r_B(\theta, \phi))^2 d\Omega \)
- Orthogonality: \( D = |a|^2 + |b|^2 - 2c \cdot d \)
- Rotation: \( b_{lm}' = \sum_{lm} R_{lm}(\alpha, \beta, \gamma) b_{lm} \)
- Carbo: \( S = a \cdot b / |(a) \cdot (b)| \)
- Hodgkin: \( S = 2a \cdot b / (|a|^2 + |b|^2) \)
- Tanimoto: \( S = a \cdot b / (|a|^2 + |b|^2 - a - b) \)
- Multi-property: \( P = pS + qS_{\text{MEP}} + rS_{\text{IE}} + ... \)
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
- 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% diff2jpg -d parafit.dif –o parafit.jpg
  a.sdf
  b.sdf
  c.sdf
  d.sdf

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

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...
Clustering SH shapes using ParaSurf, and cluster with ParaFit:


Clustering the Odour Dataset

unix% PS_mopac_run
unix% PS_Parasurf_run
unix% parafit –matrix –dif 

unix% dif2jpg –n10 o.dif *_p.psf
unix% eog o.jpg

Clustering SH shapes gives better clusters than using vibrational frequencies...

HIV and HIV Entry Inhibitors

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 and HIV Entry Inhibitors

Acquired Immune Deficiency Syndrome

It is not a hereditary disease

Group of symptoms and signs

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
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 derivatives
11) 4-hydroxypiperidine derivatives
12) Phenylcyclohexilamines
13) Anilide piperidine N-oxides
14) 1-phenyl-1,3-propanodiamines
15) AMD derivatives
16) Other

**CXCR4 antagonists (201):**
1) AMD derivatives
2) Microcycles
3) Total synthesis
4) Other

**PLUS...**
4094 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.

SH Ligand-Based VS Set-Up

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

![Superpositions](image)

- 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

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

- Consensus shapes for the three most active inhibitors
Consensus Shape-Based VS

Overall Results – CXCR4

Best scorers:
- ParaFit 3-Consensus
- ParaFit Tanimoto
- Fred Consensus
- ROCS Combo

Overall Results – CCR5

Best scorers:
- ParaFit 3-Consensus
- FRED Consensus
- ParaFit S-Consensus

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.
- Not all SDM locations affect the binding of all ligands.
- 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).

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.

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
- 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:

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

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

Screening the Berlex Dataset

• 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...

CCR5 VS with Berlex Dataset

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

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

ParaSurf + ParaFit:  [http://www.ceposinsilico.de/](http://www.ceposinsilico.de/)