Using “Consensus clustering” to help understand how multiple ligands might bind to multiple target sites. How this helps to choose the Right Query in Shape-Based Virtual Screening

Violeta Pérez-Nueno, David Ritchie
Orpailleur Team, INRIA Nancy - Grand Est

LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications), INRIA Nancy – Grand Est, 615 rue du Jardin Botanique, 54506 Vandoeuvre-lès-Nancy, France
Presentation Overview

1. Calculating SH Shapes

2. Calculating SH Consensus Shapes and their performance in VS

3. Consensus Clustering: Exploring CCR5 Multiple Binding Sites

4. All-against-all DUD Dataset Shape Clustering vs DUD Cross Docking

5. Choosing the Right Query in DUD Shape-Based VS

6. Examples of how Consensus Clustering can identify Multi-Site Pockets
1. Calculating SH Shapes
Spherical Harmonic Surfaces

Surface shapes are represented as radial distance expansions of the molecular surface with respect to the center of the molecule.

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

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

ParaFit calculates superpositions between pairs of molecules by exploiting the special rotational properties of the SH functions

- **Distance:** \[ D = \int (r_A(\theta, \phi) - r_B(\theta, \phi'))^2 d\Omega \]

- **Orthogonality:** \[ D = |a|^2 + |b|^2 - 2a \cdot b' \]

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

- **Carbo:** \[ S = a \cdot b' / (|a| \cdot |b|) \]

- **Hodgkin:** \[ S = 2a \cdot b' / (|a|^2 + |b|^2) \]

- **Tanimoto:** \[ S = a \cdot b' / (a^2 + b'^2 - a \cdot b') \]

- **Multi-property:** \[ Q = pS + qS^{\text{MEP}} + rS^{\text{IEL}} + ... \]

2. Calculating SH Consensus Shapes and their performance in VS
Calculating Consensus Shapes

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

Virtual Screening Datasets

The consensus approach was first validated in a retrospective VS of two databases containing CCR5 and CXCR4 active compounds from the literature, and presumed inactive compounds from Maybridge.

**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-hydroxyperidine 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
Pseudo-molecules were obtained from the consensus shapes of the most active molecules for both CXCR4 and CCR5 targets, and used as VS queries against the database of known actives and decoys.
Consensus Shape-Based VS

CXCR4 families very similar shapes -> results do not substantially improve

CCR5 families very different shapes -> results considerably improve

3. Consensus Clustering: Exploring CCR5 Multiple Binding Sites
Exploring CCR5 Multiple Binding Sites: Clustering the 424 CCR5 Ligands

<table>
<thead>
<tr>
<th>CLUSTER</th>
<th>Compounds Found</th>
<th>Number of compounds</th>
<th>Consensus Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(8) 1,3,4-trisubstituted pyrrolidino-piperidines (3) 1,3,5-trisubstituted penta-cycles (5) 5-oxopyrrolidine-3-carboxamides (4) N,N'-diphenylureas (2) TAK derivatives (1) 4-piperidines (1) others (MERK-CMPD 167)</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(1) 1,3,4-trisubstituted pyrrolidino-piperidines (6) 1,3,5-trisubstituted penta-cycles (13) 1-phenyl-1,3, propanoamides (2) 4-piperidines (3) AMD derivatives (9) Diketopiperazines (1) SCH derivatives (2) Phenylcyclobutanamines (3) others (GSK, Merck2, Merck3)</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(22) Antile piperidine N-oxides (1) TAK derivatives (1) others (1- bendazepine)</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(21) 1-phenyl-1,3- propanoamides (5) Phenylcyclobutanamines</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(11) 1-phenyl-1,3- propanoamides</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>(12) 1-phenyl-1,3- propanoamides</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>(26) 4-amino-piperidine or tropans (6) 4-piperidines (2) Phenylcyclobutanamines (1) others (Merek)</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>(23) SCH derivatives</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>(20) SCH derivatives</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>(37) SCH derivatives</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>(22) SCH derivatives</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>(17) SCH derivatives</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>(19) TAK derivatives</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>(44) TAK derivatives</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>(33) Guanylhydrazone derivatives</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>(36) 4-hydroxy-piperidine derivatives</td>
<td>36</td>
<td></td>
</tr>
</tbody>
</table>

- 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)
From Consensus Shapes to Super-Consensus Clusters

• Another round of Ward’s clustering proposed four super-consensus clusters.
Using Super-Consensus Shapes as VS Queries

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

  - VS Super Consensus A: AUC=0.785
  - VS Super Consensus B: AUC=0.413
  - VS Super Consensus C: AUC=0.905
  - VS Super Consensus D: AUC=0.630

- 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 (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)

To confirm that the SC shapes were properly matched to their predicted target sites, the three proposed binding sites were treated as if they were separate targets for docking-based VS:

- 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)

As before, merging SCs worsens the AUCs…

- SC docking => no less than THREE CCR5 pocket sub-sites

4. All-against-all DUD Dataset Shape Clustering vs DUD Cross Docking
Cross-docking on DUD
Huang et al. work

| Protein       | er            | er_agonist    | er_antagonist | gr            | grn           | grn_agonist   | grn_antagonist | his           | his_agonist   | his_antagonist | hsp90        | hsp90_agonist | hsp90_antagonist | p38           | p38_agonist   | p38_antagonist | p53           | p53_agonist   | p53_antagonist | pdgfra        | pdgfra_agonist | pdgfra_antagonist | pxr           | pxr_agonist   | pxr_antagonist | src           | src_agonist   | src_antagonist | ts            | ts_agonist   | ts_antagonist | vegfc2        | vegfc2_agonist | vegfc2_antagonist | hvea          | hvea_agonist | hvea_antagonist | thrombin      | thrombin_agonist | thrombin_antagonist | tpsi1         | tpsi1_agonist | tpsi1_antagonist | acoa          | acoa_agonist | acoa_antagonist | ada           | ada_agonist   | ada_antagonist | const         | const_agonist | const_antagonist | pole5         | pole5_agonist | pole5_antagonist | dhti          | dhti_agonist | dhti_antagonist | gart          | gart_agonist | gart_antagonist | acne          | acne_agonist | acne_antagonist | air2          | air2_agonist | air2_antagonist | ampc          | ampc_agonist | ampc_antagonist | cox1          | cox1_agonist | cox1_antagonist | cox2          | cox2_agonist | cox2_antagonist | gilb          | gilb_agonist | gilb_antagonist | hyst         | hyst_agonist | hyst_antagonist | hmgc          | hmgc_agonist | hmgc_antagonist | mthf          | mthf_agonist | mthf_antagonist | na            | na_agonist   | na_antagonist | pare          | pare_agonist | pare_antagonist | pnp           | pnp_agonist | pnp_antagonist | cmnh          | cmnh_agonist | cmnh_antagonist |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|----------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|

All-against-all DUD Dataset Shape Clustering vs DUD Cross Docking

- Select a threshold that emphasizes the overall similarity to the cross-docking experiment.

- Count the membership over the threshold for ligands belonging to each of the targets for all the 40 clusters and sum memberships.

- Normalize this sum by the total number of ligands for each target. This gives the cross-shape matching score.
Ligands for a given target are split into 2 or 3 sub-groups, which could suggest they might bind to different sub-sites of the same target.

P38 ligands split in C36, C30, C29
Ligands from different targets are grouped together. It suggests that they could bind to the same targets.

Evidence of cox2 cross-docked with hsp90

Evidence of sahh cross-docked with tk and ada
All-against-all DUD Dataset Shape Clustering vs DUD Cross Docking

8.45 h

several months (from 20h to 12d/target)


DUD Cross-shape matching score
(sum of normalized number of actives found in all clusters superior to 0.1)

Poor (S < 0.4)
Medium (0.4 ≤ S ≤ 0.6)
Good (0.6 ≤ S < 0.8)
Very good (S ≥ 0.8)

DUD Cross-docking enrichments
Poor (ETmax < 10 and ET20 < 2)
Medium (20 > ETmax ≥ 10 and 2.5 > ET20 ≥ 2)
Good (30 > ETmax ≥ 20 and 3 > ET20 ≥ 2.5)
Very good (ETmax ≥ 30 and ET20 ≥ 3)
5. Choosing the Right Query in DUD Shape-Based VS
Choosing the Right Query in DUD Shape-Based VS

Queries used:

- PARAFIT_B: PARAFIT bound conformation.
- PARAFIT_CM: PARAFIT consensus molecule.
- PARAFIT_C1_CM - PARAFIT_C3_CM: consensus molecules for SH clusters 1, 2, 3.
- PARAFIT_RCM: PARAFIT real center molecule closest to consensus.
- PARAFIT_C1_RCM - PARAFIT_C3_RCM: real center molecules for clusters 1, 2, 3.
- PARAFIT_SHEF_C: best SHEF molecule that fits pocket as PARAFIT query.

- ROCS_B: ROCS bound conformation.
- ROCS_RCM: real center molecule from PARAFIT as ROCS query.
- ROCS_C1_RCM - ROCS_C3_RCM: real PARAFIT center molecules as ROCS queries.
- ROCS_SHEF_C: best SHEF molecule that fits pocket as ROCS query.

- SHEF shape-based docking.
- GOLD conventional docking.

ROC plots
Bar graphs
Choosing the Right Query in DUD
Shape-Based VS
Choosing the Right Query in DUD
Shape-Based VS

ALR2

AUC

0.0 0.2 0.4 0.6 0.8 1.0

PARAFIT D
PARAFIT CM
PARAFIT C1 CM
PARAFIT C2 CM
PARAFIT ROM
PARAFIT C1 ROM
PARAFIT C2 ROM
PARAFIT SHEF C
ROCS D
ROCS ROM
ROCS C1 ROM
ROCS C2 ROM
ROCS SHEF C
SHEF
GOLD

[Bar chart showing AUC values for different methods]
6. Examples of how Consensus Clustering can identify Multi-Site Pockets
Multi-Site Pockets: p38

• 353 p38 DUD ligands clustered using Wards hierarchical clustering of chemical fingerprints

• Kelley’s method to find the optimal number of clusters (15)

• SH consensus shapes were calculated for the 15 groups

• These were then compared in ParaFit (all-vs-all)

• Another round of Ward’s clustering proposed three super-consensus clusters
Multi-Site Pockets: p38

SC_A

SC_B

SC_C

Glu 71
Asp 168
Met 109
Glu 71
Asp 168
Met 109
Glu 71
Asp 168
Met 109
Multi-Site Pockets: p38

Multi-Site Pockets: p38

SITE 1 (ATP site)  SITE 2 (Allosteric site)

SC_A  SC_B  SC_C

diaryl urea inhibitors  pyridinyl-imidazole inhibitors
Multi-Site Pockets: Alr2

- 26 alr2 DUD ligands clustered using Wards hierarchical clustering of chemical fingerprints
- Kelley’s method to find the optimal number of clusters (5)
- SH consensus shapes were calculated for the 5 groups
- These were then compared in ParaFit (all-vs-all)
- Another round of Ward’s clustering proposed three super-consensus clusters
Multi-Site Pockets: Alr2

SC_A

SC_B

SC_C
Multi-Site Pockets: Alr2

Multi-Site Pockets: Alr2

Small ligand
Sub-site 1
(catalytic cleft)

Super ligand
Sub-site 2
(whole pocket)
Conclusions

- Consensus shape provides good queries. It is also a good way to select a center molecule.

- Cross-shape matching is much faster than cross-docking in identifying possible cross-docked targets.

- Consensus clustering allows to detect multi-site targets.

- Multi-site targets are more common than previously assumed. We recommend clustering ligands if poor performance is observed with the single query.
Acknowledgements

• Dave Ritchie
• Vishwesh Venkatraman
• Lazaros Mavridis

• INRIA Nancy - Grand Est
• IQS, Universitat Ramon-Llull

• Agence Nationale de la Recherche (ANR-08-CEXC-017-01)
• Generalitat de Catalunya-DURSI (FI2008 and BE-DGR2009)

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

ParaSurf + ParaFit: http://www.cephesinsilico.de/
Thank you!
From MOPAC or VAMP, calculate:
- Density contours of $2 \times 10^{-4} \text{e}/\text{Å}^3$ (i.e. approx = SAS)
- MEP – electrostatic potential
- IEL – ionization energy
- EAL – electron affinity
- aL – polarizability

- Encode as Spherical Harmonic expansions to order $L=15$…

Choosing the Right Query in DUD Shape-Based VS

<table>
<thead>
<tr>
<th>Target</th>
<th>PF_B</th>
<th>PFT_CM</th>
<th>PFT_C1_CM</th>
<th>PFT_C2_CM</th>
<th>PFT_C3_CM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR5</td>
<td>0.594</td>
<td>0.866</td>
<td>0.785</td>
<td>0.905</td>
<td>0.630(B)</td>
</tr>
<tr>
<td>ALR2</td>
<td>0.508</td>
<td>0.609</td>
<td>0.520</td>
<td>0.617</td>
<td>0.419</td>
</tr>
<tr>
<td>COMT</td>
<td>0.371</td>
<td>0.351</td>
<td>0.508</td>
<td>0.674</td>
<td>0.378</td>
</tr>
<tr>
<td>TRYPsin</td>
<td>0.207</td>
<td>0.728</td>
<td>0.740</td>
<td>0.639</td>
<td>0.639</td>
</tr>
<tr>
<td>PDGFBRB</td>
<td>0.448</td>
<td>0.742</td>
<td>0.667</td>
<td>0.678</td>
<td>0.579</td>
</tr>
<tr>
<td>P38</td>
<td>0.498</td>
<td>0.825</td>
<td>0.552</td>
<td>0.809</td>
<td>0.753</td>
</tr>
<tr>
<td>CDK2</td>
<td>0.524</td>
<td>0.752</td>
<td>0.661</td>
<td>0.769</td>
<td>0.510</td>
</tr>
<tr>
<td>COX1</td>
<td>0.574</td>
<td>0.758</td>
<td>0.652</td>
<td>0.716</td>
<td>0.581</td>
</tr>
<tr>
<td>PR</td>
<td>0.591</td>
<td>0.797</td>
<td>0.752</td>
<td>0.741</td>
<td>0.676</td>
</tr>
<tr>
<td>THROMBIN</td>
<td>0.623</td>
<td>0.806</td>
<td>0.426</td>
<td>0.808</td>
<td>0.775</td>
</tr>
<tr>
<td>VEGFR2</td>
<td>0.598</td>
<td>0.609</td>
<td>0.472</td>
<td>0.556</td>
<td>0.629</td>
</tr>
<tr>
<td>ACE</td>
<td>0.572</td>
<td>0.769</td>
<td>0.804</td>
<td>0.452</td>
<td>0.642</td>
</tr>
<tr>
<td>RXR_ALPHA</td>
<td>0.908</td>
<td>0.960</td>
<td>0.689</td>
<td>0.904</td>
<td>0.892</td>
</tr>
<tr>
<td>COX2</td>
<td>0.919</td>
<td>0.945</td>
<td>0.935</td>
<td>0.947</td>
<td>0.881</td>
</tr>
<tr>
<td>SAHH</td>
<td>0.859</td>
<td>0.944</td>
<td>0.871</td>
<td>0.931</td>
<td>0.923</td>
</tr>
</tbody>
</table>
Choosing the Right Query in DUD Shape-Based VS
Choosing the Right Query in DUD Shape-Based VS