Analysis of Activity Landscapes, Activity Cliffs, and Selectivity Cliffs

Jürgen Bajorath
Life Science Informatics
University of Bonn
Concept of Activity Landscapes

- “Activity landscapes”: biological activity hypersurfaces within chemical space; visualized as a 2D projection of chemical space with compound potency added as the third dimension
Idealized Activity Landscapes and SARs

Continuous SAR
gradual changes in structure result in moderate changes in activity

Discontinuous SAR
small changes in structure have dramatic effects on activity

“rolling hills”

“activity cliffs”
Activity Landscapes and SARs

Variable Activity Landscapes $\rightarrow$ Heterogeneous SARs

Combination of continuous and discontinuous SAR components

rugged regions

smooth regions
SAR Index (SARI) Scoring

Numerical SAR analysis function to support activity landscape assessment

\[
\text{cont} = \text{weighted mean}_{\{i,j| i>j\}} \left( \frac{1}{1 + \text{sim}(i,j)} \right) \\
\text{weight}_{ij} = \frac{P_i \cdot P_j}{1 + |P_i - P_j|}
\]

\[
\text{disc} = \text{mean}_{\{i,j| i>j, |P_i - P_j| > 1, \text{sim}(i,j) > 0.65\}} \left( P_i - P_j \right) \cdot \text{sim}(i,j)
\]

continuity score emphases structurally diverse compounds having similar potency

discontinuity score emphases similar compounds with large potency differences

GLOBAL SCORE
all possible compound pairs

\[
\text{SARI} = \frac{1}{2} \left( \text{cont} - (1 - \text{disc}) \right)
\]
Multi-Level Discontinuity Scoring

- **Activity class $A$**
  - SARI disc. score
    - global SAR character of an activity class

- **Cluster $C$**
  - cluster discontinuity scores
    - local SAR character
    - presence of activity cliffs

- **Compound $i$**
  - compound discontinuity scores
    - contributions of molecules to local SAR discontinuity

Mathematical expression:

$$\text{disc}_{\text{raw}}(A) = \text{mean}_{(i,j) \in A \mid \text{sim}(i,j) > 0.65, \mid P_i - P_j \mid > 1} \left( \mid P_i - P_j \mid \cdot \text{sim}(i, j) \right)$$
From Idealized to Calculated 3D Activity Landscapes

- “Coordinate-free” chemical space (MACCS Tc distances)
- 2D projection through multi-dimensional scaling
- x/y-plane: chemical space projection
- z-axis: interpolated potency values
- color code: surface elevation

Cathepsin S inhibitors
Activity Landscapes – 2D Projection

Squalene synthase inhibitors

- Reference space
  - MACCS Tc distances
  - MDS on Euclidean fingerprint distances

- Coordinate scaling
  - range scaling to [0,1]
  - multiplication with maximum fingerprint dissimilarity
3D Surface Generation

1. Add potency as 3\textsuperscript{rd} dimension
2. Interpolate on a regular grid
3. Generate 3D surface
Activity Landscapes – 3D Projection

- 80 x 80 grid
- Landscapes are comparable for different data sets
  - unified axes range
  - colored according to interpolated surface elevation:
    \[ \leq 5.8 \quad \rightarrow \quad \geq 8.7 \]
- Transparent areas are not populated with data points
Landscape Model Optimization

- Correlation between fingerprint dissimilarity and calculated 2D coordinate distances (80+%) 
- Interpolated 3D surface elevation and measured potency values (95+%)
Continuous Landscape

- MACCS Euclidean distance
- SARI: 0.92
- Mostly weakly active molecules
- Densely populated
- Data peaks do not correspond to SAR discontinuity
Discontinuous Landscape

Acetylcholine esterase

- MACCS Euclidean distance
- SARI: 0.17
- Cliff regions: similar molecules with high/low potency
Influence of Molecular Representations

MACCS

TGT

Molprint2D

1. 
\[
\begin{align*}
\text{N} & \text{H} \\
\text{H} & \text{N} \\
\text{N} & \text{H} \\
& \text{Cl} \\
\end{align*}
\]

\[0.02 \text{ nM}\]

2. 
\[
\begin{align*}
\text{N} & \text{H} \\
\text{N} & \text{H} \\
\text{O} & \text{N} \\
\text{Cl} & \\
\end{align*}
\]

\[590 \text{ nM}\]

3. 
\[
\begin{align*}
\text{N} & \text{H} \\
\text{N} & \text{H} \\
\text{N} & \text{N} \\
\text{N} & \\
\end{align*}
\]

\[0.4 \text{ nM}\]

4. 
\[
\begin{align*}
\text{N} & \text{H} \\
\text{N} & \text{H} \\
\text{N} & \text{OH} \\
\end{align*}
\]

\[32 \mu\text{M}\]
Network-like Similarity Graph (NSG)

Annotated graph representation of similarity relationships in compound data sets

**Nodes:** represent compounds in the data set

**Edges:** connect nodes with high pairwise similarity

**Clusters:** Hierarchical clustering (gray background)

**Layout:** Fruchterman-Reingold

**Global scores**
- Continuity: 0.79
- Discontinuity: 0.99
- SAR Index: 0.40
NSG Information Layers

1. similarity relationships
2. potency distribution
3. compound discont. scores
4. clusters
5. cluster scores

Edges

- $T_c > x$ (eg 0.65)
- $T_c < x$
NSG Information Layers

Node color

1  similarity relationships
2  potency distribution
3  compound discont. scores
4  clusters
5  cluster scores
NSG Information Layers

Node size

- high compound score
- low compound score

1 similarity relationships
2 potency distribution
3 compound discont. scores
4 clusters
5 cluster scores
NSG Information Layers

Cluster islands

clusters obtained by Ward’s hierarchical clustering are highlighted

1 similarity relationships
2 potency distribution
3 compound discont. scores
4 clusters
5 cluster scores
NSG Information Layers

Annotation

clusters are annotated with global discontinuity scores calculated for cluster members

1 similarity relationships
2 potency distribution
3 compound discont. scores
4 clusters
5 cluster discont. scores
SAR Phenotypes - Lipoxygenase Inhibitors: Globally Continuous SAR

Many compound clusters with continuous local SARs

Global scores
continuity 0.986
discontinuity 0.037
SAR Index 0.974
SAR Phenotypes - Thrombin Inhibitors: Globally Discontinuous SAR

Clusters with large-magnitude activity cliffs

Global scores
- continuity: 0.081
- discontinuity: 0.665
- SAR Index: 0.208
SAR Phenotypes - Squalene Synthase Inhib.: Heterogeneous-relaxed SAR

Global scores
- continuity: 0.79
- discontinuity: 0.99
- SAR Index: 0.40
Activity Cliff Index

- NSGs provide interactive graphical access to prominent activity cliffs
- Cliff Index (CI) enables systematic mining and ranking of activity cliffs
- CI prioritizes pairs of similar compounds having large potency differences:

\[
CI(i, j) = (1 + \text{sim}(i, j))^2 \cdot |P_i - P_j|
\]
3D vs. 2D Activity Landscapes

Squalene synthase

Activity cliff region

Interpolated area

Smooth region

NSG cluster discontinuity: 1.00

NSG cluster discontinuity: 0.01
From Activity Cliffs to **Selectivity Cliffs**

- **Multi-target SARs**

- **Target-pair selectivity:** difference between logarithmic potency

$$ S_{A/B}(i) = -S_{B/A} = P_A(i) - P_B(i) $$

- **Structure-selectivity relationships (SSRs)**

$\text{pIC}50 = 9 \quad \frac{S_{L/B}}{S_{B/L}} = 2 \quad \text{pIC}50 = 7 $
SAR and SSR Network Analysis

Potency-based NSG

Potency:

10.4

3.0

Compound discontinuity score:

1  activity cliff markers

0

Cluster discontinuity score

1  “rough” SAR

0  “smooth” SAR
SAR and SSR Network Analysis

Selectivity-based NSG

Selectivity:

3.2 (L) – 3.2 (B)

Compound discontinuity score:

1 selectivity cliff markers

0

Cluster discontinuity score

1 “rough” SSR

0 “smooth” SSR

cathepsin L / cathepsin B
Local SSR Environments

cathepsin L / cathepsin B

discontinuous SSR
Activity Cliffs vs. Selectivity Cliffs

L: 15 nM
B: 3.5 µM
L/B: 1.4

L: 10 µM
B: 170 nM
L/B: -1.8
Local SSR Environments

cathepsin L / cathepsin B

0.72

0.73

discontinuous SSR
Activity Cliffs vs. Selectivity Cliffs

continuous SAR

discontinuous SSR

selectivity cliff markers

L: 3.6 µM
B: 102 µM
L/B: 1.5

L: 26 µM
B: 5.3 µM
L/B: -0.7
Selectivity Determinants

- Molecules with different selectivity are often found in the neighborhood of selectivity cliff markers.
- Selectivity rules can be formulated.

 halogens with increasing bulkiness and decreasing electronegativity shift selectivity toward cat L

<table>
<thead>
<tr>
<th>Halogen</th>
<th>Selectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>sel: 0.1</td>
</tr>
<tr>
<td>Cl</td>
<td>sel: 1.5</td>
</tr>
<tr>
<td>I</td>
<td>sel: 2.0</td>
</tr>
</tbody>
</table>
Narrowly Focused Landscape Views: Combinatorial Analog Graphs (CAGs)

- CAGs are designed to systematically analyze SARs in analog series

- Substitutions introducing SAR discontinuity or selectivity determinants can be identified
Activity Cliffs in Analog Series

Analog series

Graph

Discontinuity scoring

R-group decomposition
Combinatorial Analog Graph (CAG)

**edges**
- compounds in connected subsets share modifications at substitution sites

**site numbers**
- discontinuity score

**discontinuity score**
- compounds with variations at one site
- compounds with simultaneous variations two sites

**complete series**
- 0.45

**compounds with variations at one site**
- 1: 0.33
- 2: 0.33
- 3: 0.84

**compounds with simultaneous variations two sites**
- 1-2: 0.22
- 1-3: 0.65
- 2-3: 0.58
- 1-2-3: 0.78
Local Compound Similarity Assessment

19 substituent categories:

3 structural features \times 6 pharmacophore features + no substituent

- acyclic (Ac)
- aromatic (Ar)
- aliphatic (Al)
- donor (D)
- acceptor (A)
- donor & acceptor (DA)
- positively charged (P)
- negatively charged (N)
- featureless (-)
Similarity Assessment and Disc. Score

<table>
<thead>
<tr>
<th>Site</th>
<th>Mutation</th>
<th>Edit distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ac → Ac</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>(-) → Ac-DA</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>Ar → Ac</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sum 0.4$</td>
</tr>
</tbody>
</table>

Similarity: 1 – distance($i,j$)

$\Delta pK_i$: $pK_i(i) - pK_i(j)$

Disc$_{raw}$: $\text{sim}(i,j) \times \Delta pK_i(i,j)$
Discontinuity Score

Compounds with variations at substitution site 2:

- Discontinuity score: \( \text{disc} = \text{mean} \left( |P_i - P_j| \cdot \text{sim}(i, j) \right) \)

- SAR discontinuity contributions from individual sites or site combinations
SAR Hotspots and SAR Holes

**SAR Hotspots (red):**
Sites where substitutions significantly change potency

**SAR Holes (white):**
undersampled sites or site combinations
Local SAR Discontinuity / Activity Cliffs

Thrombin inhibitors

- 741 nM
- 160 nM
- 2 μM
- 204 nM
- 926 nM
- 82 nM
- 7 μM
- 1 nM
Structural Interpretation of Activity Cliffs

Factor Xa inhibitors

18 nM

950 nM

18 nM
Structural Interpretation of Activity Cliffs

Factor Xa inhibitors

Tautomerism disrupts π stacking

950 nM
Structural Interpretation of Activity Cliffs

Tie-2 kinase inhibitors

1 nM

153 nM
Structural Interpretation of Activity Cliffs

Tie-2 kinase inhibitors

Methylamine substituent forms strong hydrogen bond

1 nM
Conclusions

- Combining numerical and graphical analysis tools for activity landscape analysis and mining of SAR information

- NSGs for the systematic comparison of global and local SAR features in compound data sets

- From activity cliffs to selectivity cliffs through multi-target SAR analysis in NSGs

- Topology of calculated 3D landscapes for different data sets and molecular representations

- Combinatorial Analog Graphs to organize analog series and identify SAR hotspots and SAR holes
Acknowledgments

Preety Iyer
Qingli Liu
Eugen Lounkine
Lisa Peltason
Anne Mai Wassermann
Mathias Wawer