Numerical Characterization of Structure-Activity Relationships from a Medicinal Chemists Point of View

Rajarshi Guha

School of Informatics
Indiana University

National Medicinal Chemistry Symposium
Pittsburgh, PA
15th June, 2008

Structure Activity Relationships

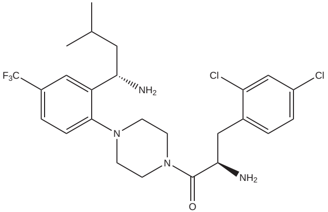
Assumptions

- Similar molecules will have similar activities
- Small changes in structure will lead to small changes in activity
- One implication is that SAR's are additive
- This is the basis for QSAR modeling

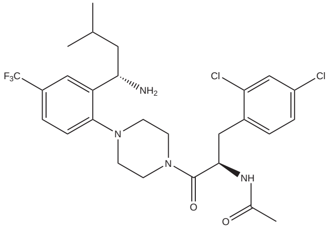
Martin, Y.C. et al., *J. Med. Chem.*, **2002**, *45*, 4350–4358

Structure Activity Landscapes

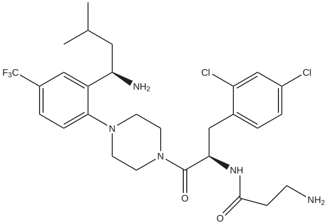
Melanocortin-4 receptor inhibitors



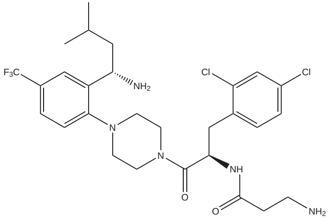
K_i = 39.0 nM



K_i = 1.8 nM



K_i = 10.0 nM



K_i = 1.0 nM

Tran, J.A. et al., *Bioorg. Med. Chem. Lett.*, **2007**, *15*, 5166–5176

Structure Activity Landscapes

Rugged gorges or rolling hills?

- Small structural changes associated with large activity changes represent steep slopes in the landscape
  - Activity Cliffs
- But traditionally, QSAR *assumes* gentle slopes
- Machine learning is not very good for special cases

Maggiora, G.M., *J. Chem. Inf. Model.*, **2006**, *46*, 1535–1535

Characterizing the Landscape

Converting activity cliffs to numbers

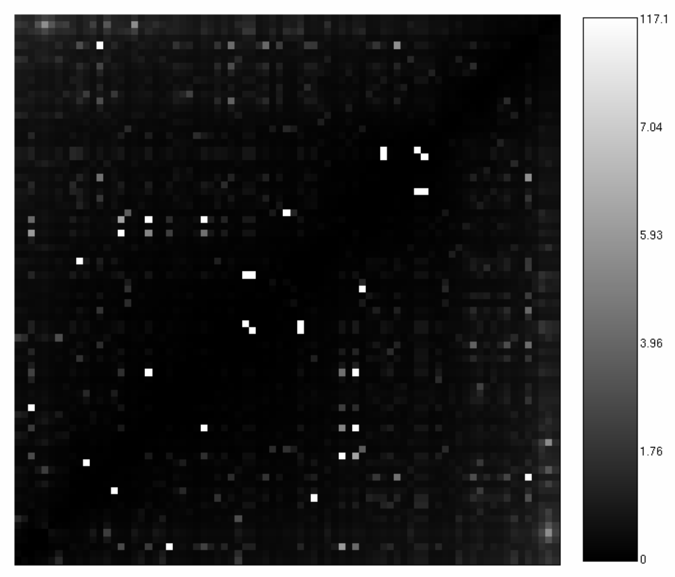
- A cliff can be numerically characterized
- Structure-Activity Landscape Index (SALI)

$$\mathrm{SALI}_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)}$$

- Cliffs are characterized by elements of the matrix with very large values

Guha, R.; Van Drie, J.H., *J. Chem. Inf. Model.*, **2008**, *48*, 646–658

Visualizing the SALI Matrix



Visualizing SALI Values

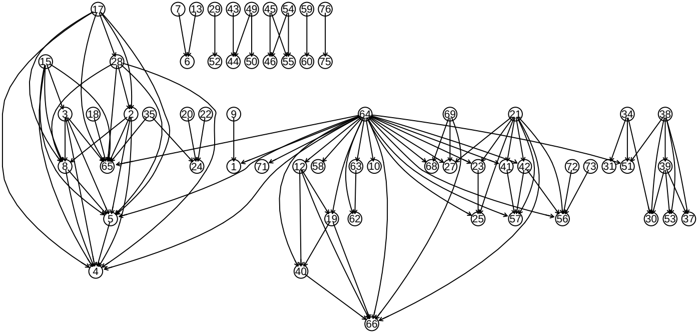
Alternatives?

- A heatmap is an easy to understand visualization
- Coupled with brushing, can be a handy tool
- A more flexible approach is to consider a network view of the matrix

The SALI graph

- Compounds are nodes
- Nodes i, j are connected if $\mathrm{SALI}_{i,j} > X$
- Only display connected nodes

Visualizing the SALI Graph



- Nodes are ordered such that the tail node in an edge has lower activity than the head node

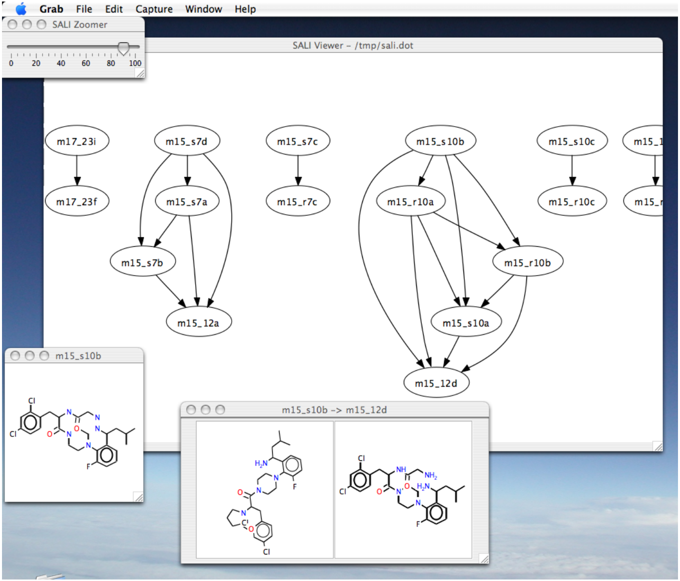
Better Visualization

SALIViewer

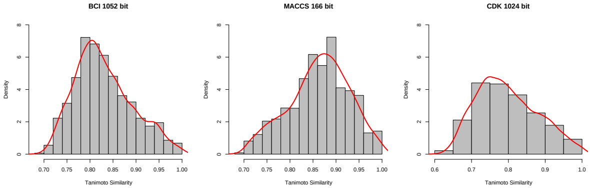
- Java application for generating and visualizing SALI graphs
- Create SALI graphs from SMILES and activity data, using the CDK fingerprints
- Easily examine SALI graphs at different cutoffs
- Provides 2D depictions for nodes and edges
- Generate SALI curves

http://cheminfo.informatics.indiana.edu/~rguha/code/java/salivis

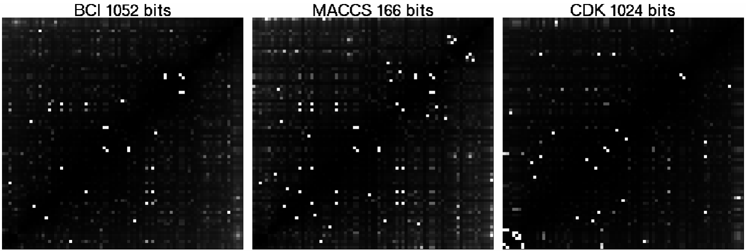
Better Visualization - SALIViewer



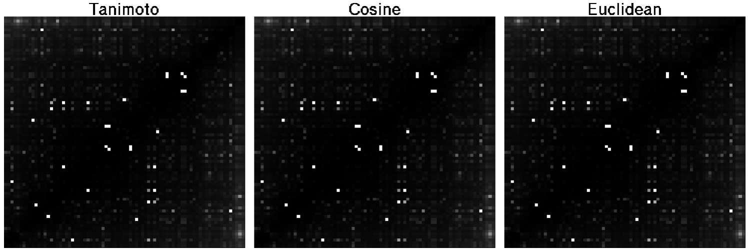
Varying Fingerprint Methods



- Shorter fingerprints will lead to more “similar” pairs
- Requires a higher cutoff to focus on significant cliffs



Varying the Similarity Metric



- The similarity metric does not affect the SALI values

SALI Graphs & Predictive Models

- The graph view allows us to view SAR's and identify trends easily
- The aim of a QSAR model is to encode SAR's
- Traditionally, we consider the quality of a model in terms of RMSE or R^2
- But in general, we're not as interested in RMSE's as we are in whether the model predicted something as more active than something else
  - What we want to have is the correct ordering
  - We assume the model is statistically significant

SALI Graphs & Predictive Models

Measuring model quality

- A QSAR model should easily encode the "rolling hills"
- A good model captures the most significant cliffs
- Can be formalized as

  How many of the edge orderings of a SALI graph does the model predict correctly?

- Define $S(X)$, representing the number of edges correctly predicted for a SALI network at a threshold X
- Repeat for varying X and obtain the *SALI curve*

SALI Graphs & Predictive Models

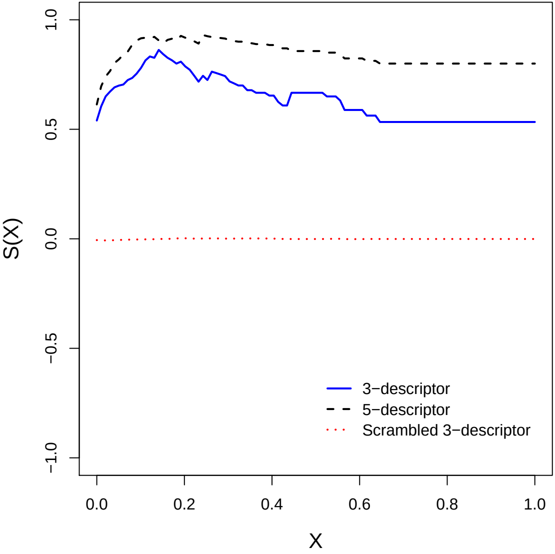
Measuring model quality

- A QSAR model should easily encode the “rolling hills”
- A good model captures the most significant cliffs
- Can be formalized as

> *How many of the edge orderings of a SALI graph does the model predict correctly?*

- Define $S(X)$, representing the number of edges correctly predicted for a SALI network at a threshold X
- Repeat for varying X and obtain the *SALI curve*

SALI Curves - An Example

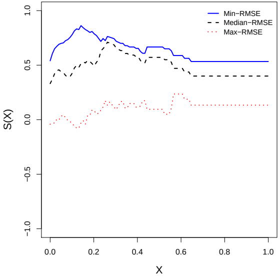


SALI Curves & Model Comparison

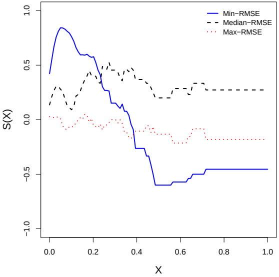
- Considered four datasets
- Developed linear regression models, using exhaustive search for feature selection
- Identify three models for each dataset
  - Minimum RMSE ("best")
  - Median RMSE
  - Maximum RMSE ("worst")
- Generate SALI curves for each model and summarize by dataset

Guha, R.; Van Drie, J.H., *J. Chem. Inf. Model.*, submitted

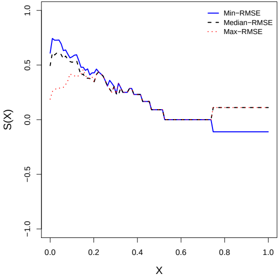
SALI Curves & Model Comparison



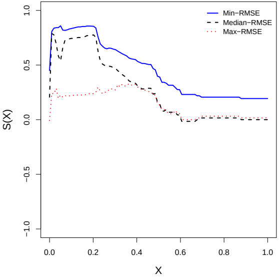
PDGFR (3)



Artemisinin (3)



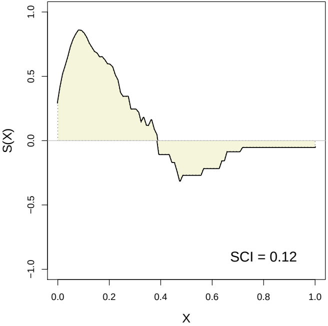
Melanocortin (6)



Benzodiazepine (5)

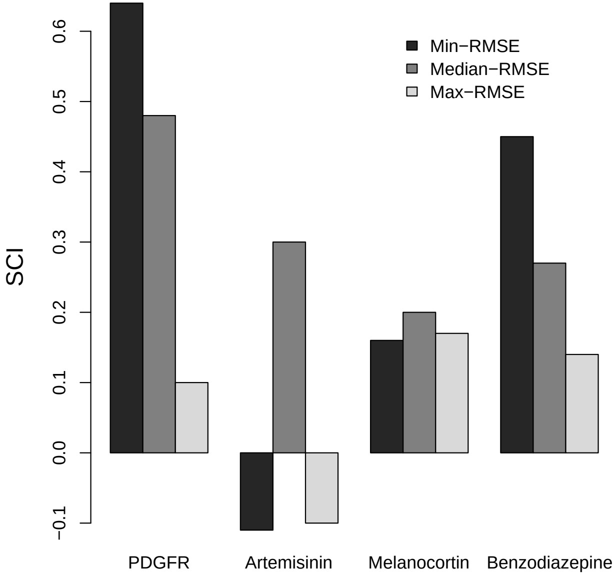
SALI Curves & Model Comparison

- The initial and final portions of the curve are of interest
- It's also useful to summarize the whole curve
- We evaluate the area between the curve and the X-axis (SCI)
    - -1 $\leq$ SCI $\leq$ 1



SALI **C**urve **I**ntegral

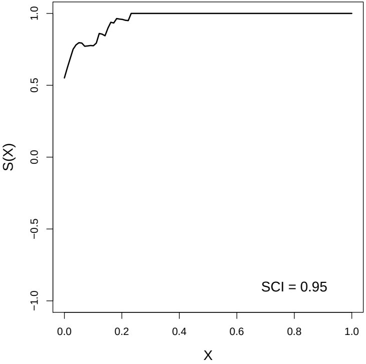
SALI Curves & Model Comparison



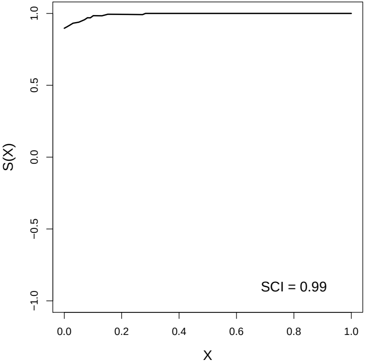
Examining Any Type of Model . . .

- Previous examples make use of predicted values from QSAR models
- We can consider any "prediction" that is supposed to track observed activity
  - Ranks
  - Energies
- Allows us to apply this approach to *any* type of computational model that predicts something
  - Docking
  - CoMFA
  - Pharmacophore

Docking & CoMFA Models



Docking



CoMFA

- Not surprising that 3D models capture more cliffs
- The CoMFA model is nearly perfect!

Holloway, K. et al, *J. Med. Chem.*, **1995**, *38*, 305–317

Cavalli, A. et al, *J. Med. Chem.*, **2002**, *45*, 3844–3853

Comparing Landscapes

- The SALI curve is a function of
  - dataset
  - descriptor space
- We can quantify a descriptor spaces ability to encode the structure-activity landscape using SALI graphs
  - What is the size of the graph as a function of SALI cutoff?
- The SALI approach allows us to investigate molecular representations that may not be directly accessible
- Work in progress

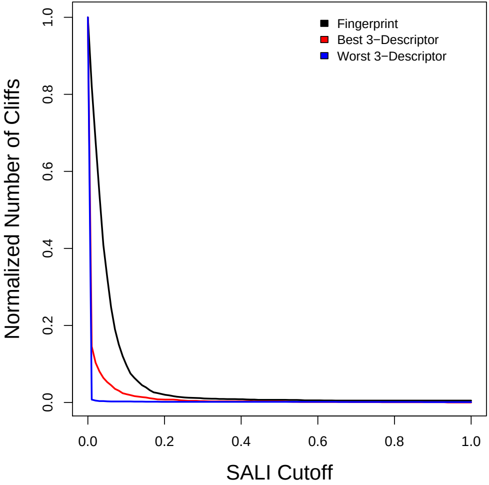
Comparing Landscapes

- The SALI curve is a function of
  - dataset
  - descriptor space
- We can quantify a descriptor spaces ability to encode the structure-activity landscape using SALI graphs
  - What is the size of the graph as a function of SALI cutoff?
- The SALI approach allows us to investigate molecular representations that may not be directly accessible
- Work in progress

Comparing Landscapes

- The SALI curve is a function of
  - dataset
  - descriptor space
- We can quantify a descriptor spaces ability to encode the structure-activity landscape using SALI graphs
  - What is the size of the graph as a function of SALI cutoff?
- The SALI approach allows us to investigate molecular representations that may not be directly accessible
- Work in progress

Comparing Landscapes



A Type 2 SALI curve for the PDGFR dataset, comparing 3 different molecular representations

What's Next?

- SALI graphs and curves represent a *framework* for exploring structure-$*$ landscapes

Open questions

- Weighted SALI graphs (ADMET, synthetic feasibility)
- Is it correct to identify cliffs using fingerprints, and then predict cliffs using different descriptors?
- Can we use SALI curves to compare 3D and 2D descriptor spaces?
- Can we use SCI for feature selection?

Conclusions

- The SALI is an effective way to numerically encode activity cliffs
- The network view of these values allows us to explore SAR's in an intuitive way
- Using the SALI curve allows us to compare predictive models in a manner that is intuitive for a medicinal chemist

Acknowledgments

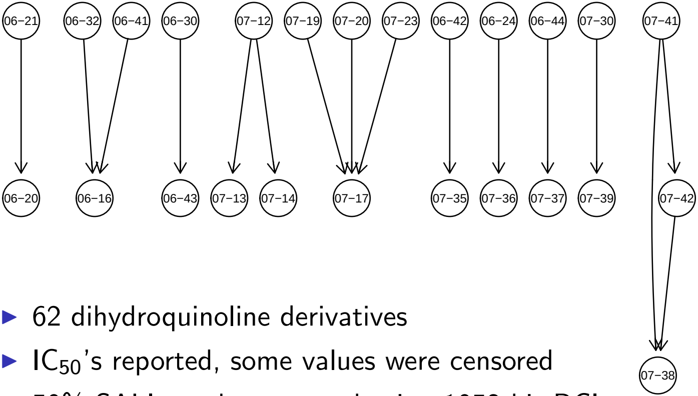
- John Van Drie

Making Use of the SALI Graph

- A little difficult with a non-interactive graph
- We can investigate a series of transformations that increase (or decrease) activity
- Identify two types of SAR's
  - Broad
  - Detailed
  - Depends on what cutoff we choose
- These correspond somewhat to the continuous and discontinuous SAR's described by Peltason et al.

Peltason, L. et al., *J. Med. Chem.*, **2007**, *50*, 5571–5578

Glucocorticoid Inhibitors

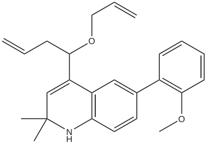


- 62 dihydroquinoline derivatives
- IC_{50}'s reported, some values were censored
- 50% SALI graph generated using 1052 bit BCI fingerprints

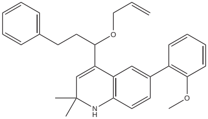
Takahashi, H. et al, *Bioorg. Med. Chem. Lett.*, **2007**, *17*, 5091–5095

Glucocorticoid Inhibitors

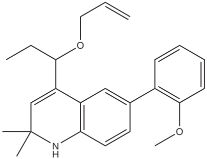
- Moving from ally or phenylethyl to ethyl causes a 6-fold increase in activity
- Reducing bulk at this position seems to improve activity
  - Pretty broad conclusion
- But ethyl is not much smaller than allyl
- We need more detail



07-20, 2000 nM

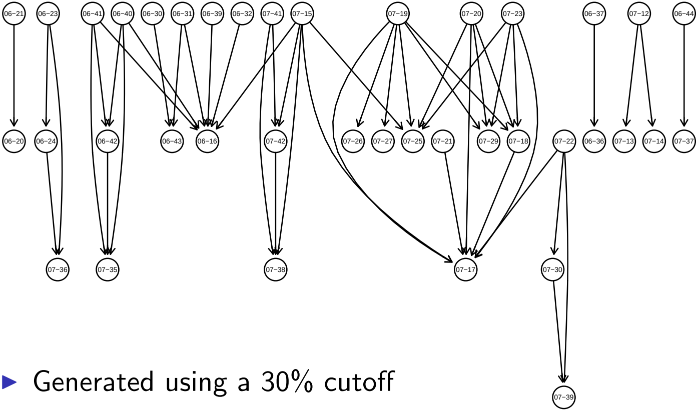


07-23, 2000 nM



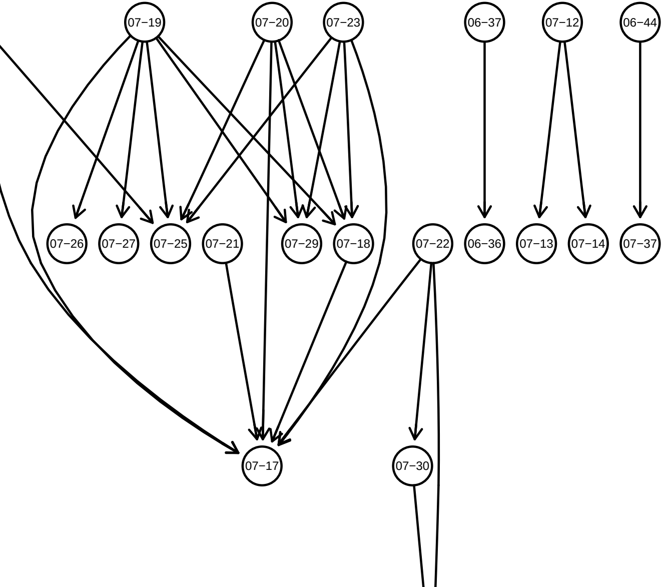
07-17, 355 nM

Glucocorticoid Inhibitors

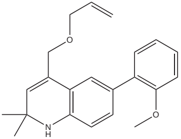


- Generated using a 30% cutoff

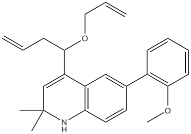
Glucocorticoid Inhibitors



Glucocorticoid Inhibitors

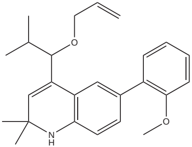


07-15, 2000 nM

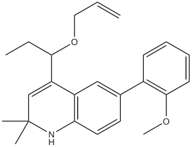


07-20, 2000 nM

- Suggests that electron density is also important
- Lower π density possibly correlates to increased activity
- Confirmed by **07-23** → **07-18**
- **07-15** → **07-17** is interesting since the change *increases* the bulk



07-18, 710 nM

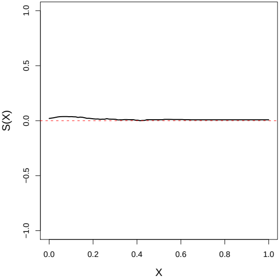


07-17, 355 nM

Glucocorticoid Inhibitors

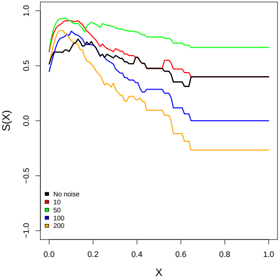
- These observations match those made by Takahashi et al.
- More detailed graphs exhibit longer paths that focus on the bulk of side chains at the C4-α position
- A number of paths consider changes to the epoxide substitution
  - Usually of length 1
  - Highlights the fact that bulk at the C4 α has greater impact on activity than epoxide substitutions
- The SALI graph stresses the non-linearity of the SAR

SALI Curves - Control Experiments



Scrambling

- Scramble the Y-variable and rebuild the model
- Evaluate the SALI curve
- Repeat 50 times and take the mean of the counts for a given cutoff



Noise

- Add uniform noise to each descriptor, rebuild the model
- We expect little variation in the plateau