Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions

Rajarshi Guha, Debojyoti Dutta, Peter C. Jurs, Ting Chen

School of Informatics
Indiana University

Department of Computational Biology
University of Southern California

Department of Chemistry
Pennsylvania State University

14th September, 2006

Outline

Outline

QSAR Modeling

Global Models

- Traditional models are *global* since they encompass the whole dataset
- They can be influenced by molecules at the extremes of the activity range
- A complex SAR may not be characterized well by a single model

QSAR Modeling

Local Models

- Global models capture *major* trends
- These might overshadow minor features which are relevant to a small group of molecules
- To take into account local trends we can build models on clusters of compounds
  - Generally requires us to define the number of clusters *a priori*
  - Valid if there are distinct clusters

QSAR & Near Neighbors

- Rather than cluster *a priori*, detect near neighbors on the fly
- Fit a linear model to the neighborhood
  - Can be extended to higher order models
  - Dependent on the nature of the neighborhood
- This implies that there is no single model for the dataset
- Previously used for time series analysis

Birattari, M. et al., *Adv. Neural. Inf. Proc. Sys.*, **1999**, 375–381

Prior Work

- Genetic algorithm to search for subsets exhibiting a linear trend
- Clustering
  - Spanning trees
  - k-means
- Kriging

Barakat, N. et al., *Chemom. Intell. Lab. Systems*, **2004**, *72*, 73–82

Shen, Q. et al., *J. Chem. Inf. Model.*, **2004**, *44*, 2027–2031

Fang, K.T. et al., *J. Chem. Inf. Model.*, **2004**, *72*, 2106–2113

Outline

Local Linear Regression

- Traditional OLS

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\beta = (X^T X)^{-1} X^T Y$$

- Local regression

$$\beta_{x'} = (X^T_{NN(x')} X_{NN(x')})^{-1} X^T_{NN(x')}$$

- Models complex relationships using simple approximations
- Saves training time by deferring model building

Tools & Datasets

Tools and techniques

- R 2.2.0
- The `lazy` package
- k was automatically determined using LOO cross-validation
- Since the neighborhood is usually small, ridge regression is used

Datasets

- Artemisinin analogs
  - 179 molecules
  - Reduced pool of 65 descriptors
- DHFR inhibitors
  - 672 molecules
  - Reduced pool of 36 descriptors

Summary of the Results

		RMSE	
Dataset		Global Model	Local Model
Artemisinin	Whole Dataset	0.86	0.62
	Prediction Set	0.92	0.94
PDGFR	Whole Dataset	0.50	0.43
	Prediction Set	0.36	0.31
DHFR	Whole Dataset	2.14	1.56
	Prediction Set	2.16	2.01

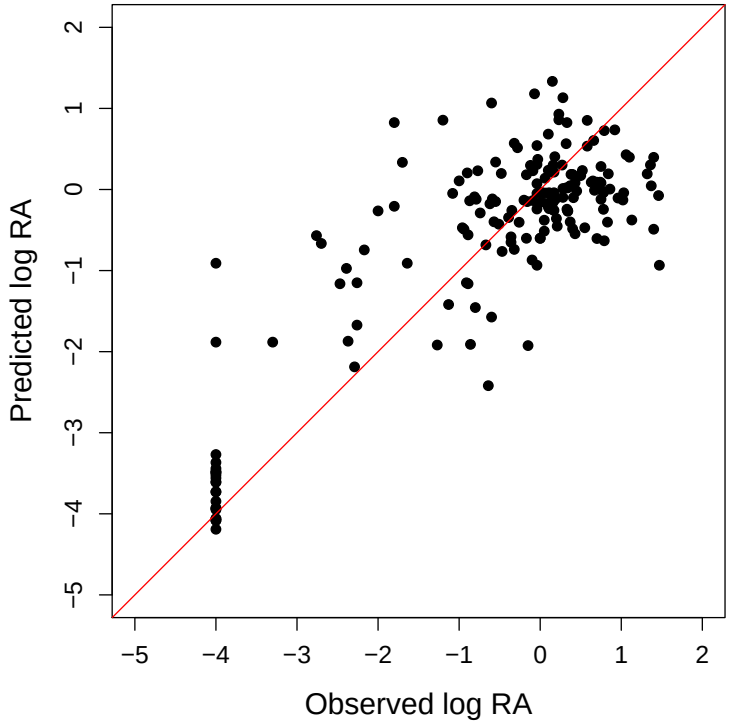
The range of the dependent variable for the artemisinin dataset was 2.53 log units, for the PDGFR dataset was 2.95 log units and for the DHFR dataset was 17.09 log units.

Guha, R. et al.; *J. Chem. Inf. Model.*, **2006**, *46*(4), 1836–1847

Artemisinin Analogs

Global Regression Model

- Built a 4-descriptor OLS model using the whole dataset
- Statistically valid
- RMSE = 0.86
- Significant variation for actives
- 23 *inactives* are not predicted well



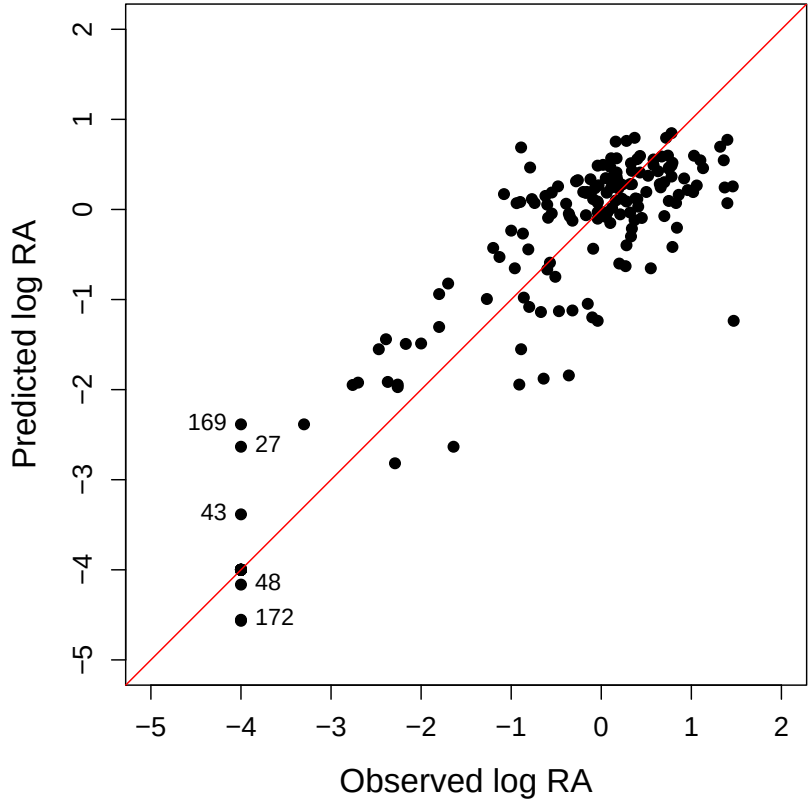
	Estimate	Std. Error	t value
(Intercept)	−60.17	5.03	−11.96
N7CH	−0.21	0.01	−17.30
NSB	0.22	0.02	10.30
WTPT-2	27.73	2.48	11.18
MDE-14	0.11	0.02	4.93

F-value $= 100.2\ (4, 174)$
$F_{crit} = 2.42\ (\alpha = 0.05)$

Artemisinin Analogs

Local Regression Model(s)

- Since we see two groups of molecules, local regression should work better
- Used descriptors from the global model
- Overall RMSE = 0.62
- Of the 23 inactives
  - 5 are mispredicted
  - exhibit increased variance

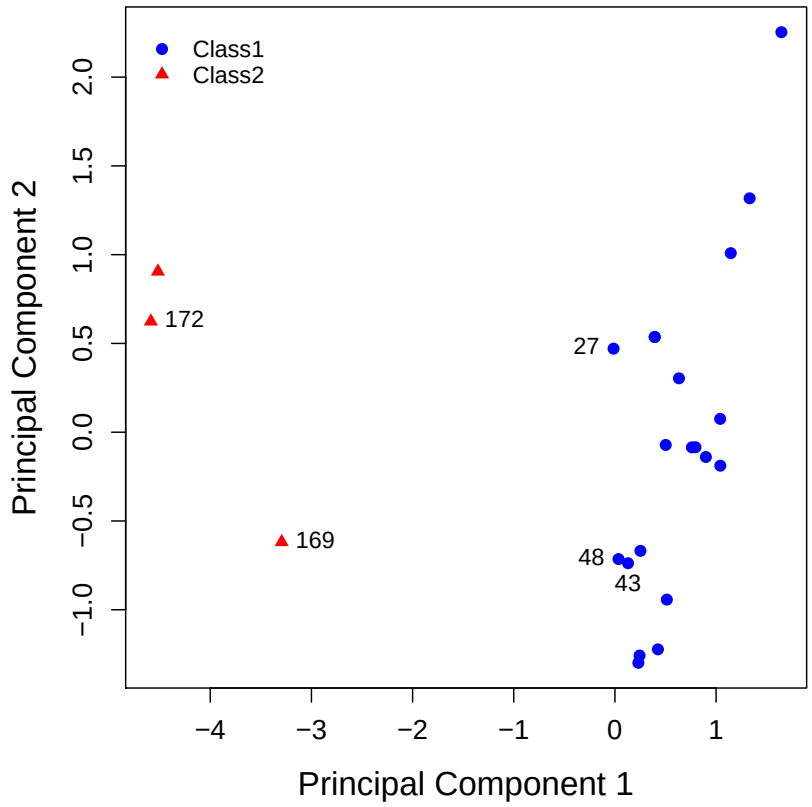


Artemisinin Analogs

Fuzzy Clustering of Inactives

- Evaluate principal components
- Colored based on cluster membership
- Number of cluster defined *a priori*

- **169** and **27** had the largest error
- Though **27** is close to a cluster, its silhouette value is very low compared to that of cluster average

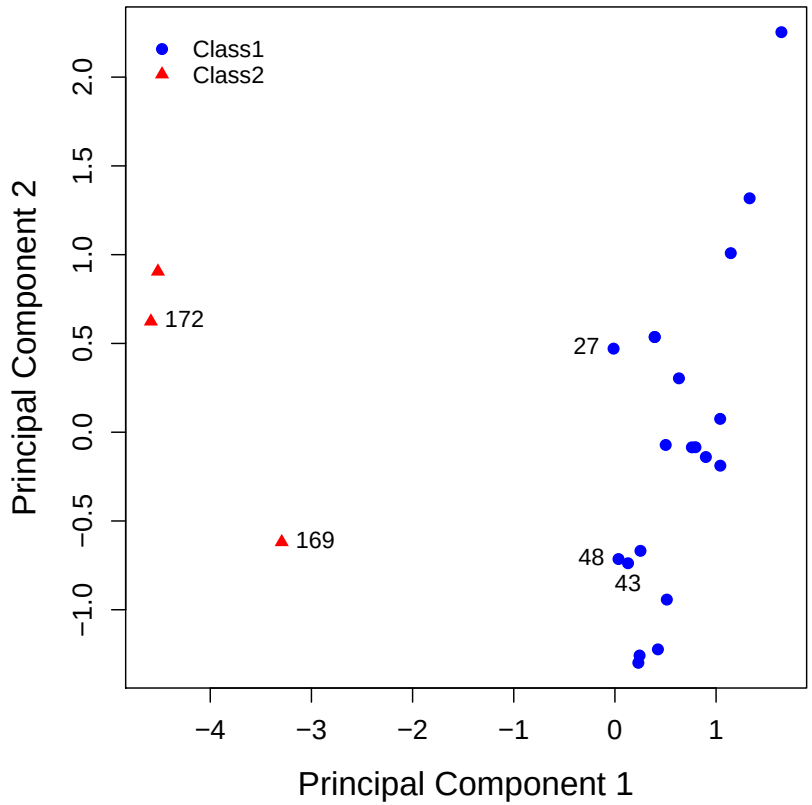


Artemisinin Analogs

Fuzzy Clustering of Inactives

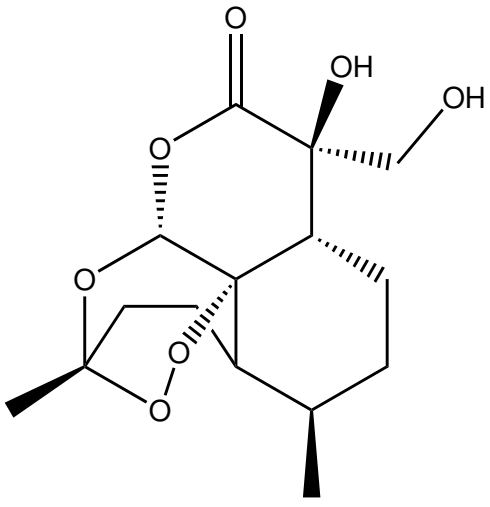
- Evaluate principal components
- Colored based on cluster membership
- Number of cluster defined *a priori*

- **169** and **27** had the largest error
- Though **27** is close to a cluster, its silhouette value is very low compared to that of cluster average

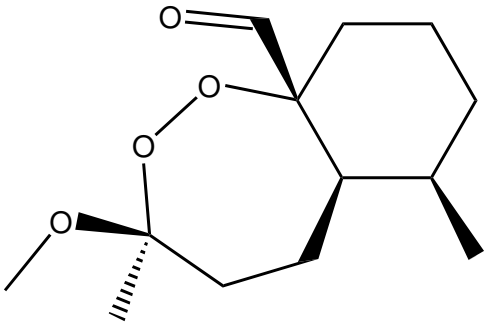


Artemisinin Analogs

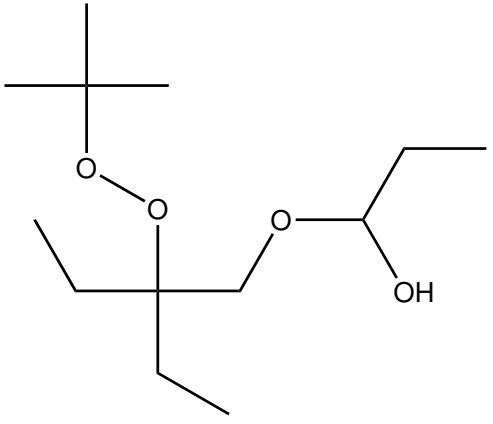
Inactive Outliers



27



169



172

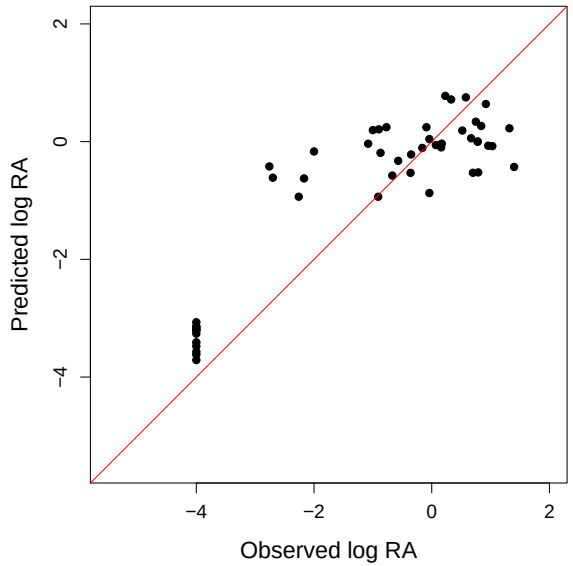
Artemisinin Analogs

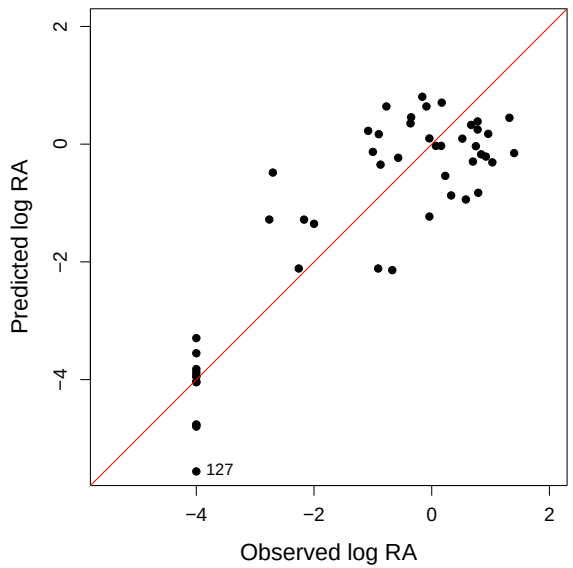
TSET/PSET Split

- Split the dataset into TSET (129) and PSET (50)
- Built a global model using the TSET, predict the PSET
- Used local regression to get predictions for the PSET

Results

- Global, RMSE = 0.92
- Local, RMSE = 0.94
- Removing *127* results in RMSE = 0.91

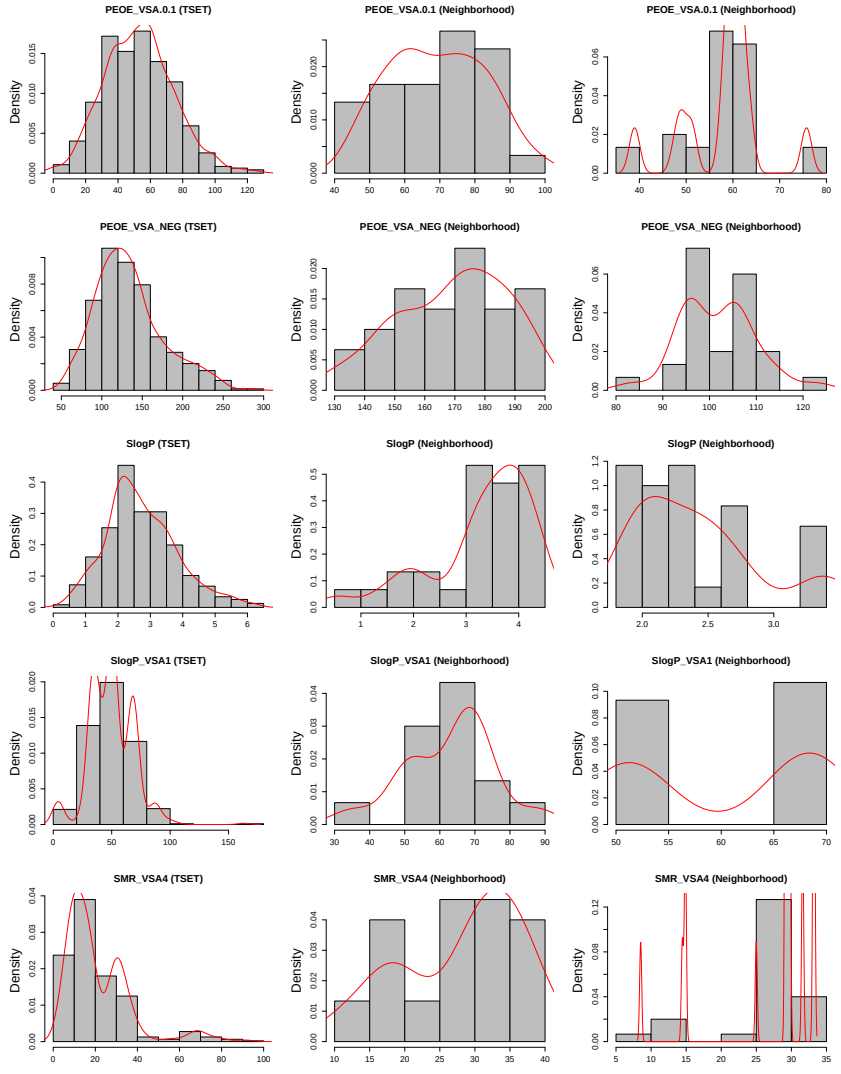




Local Descriptor Distribution

DHFR Inhibitors

- If the local distribution is similar to the population distribution one may expect reliable predictions
- Skewed distributions will not lead to significant models
- If the underlying relationship between descriptors and observed activity is not linear in the neighborhood, the local model will be poor

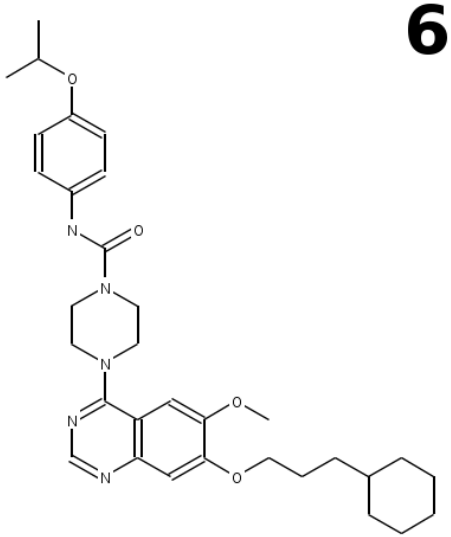


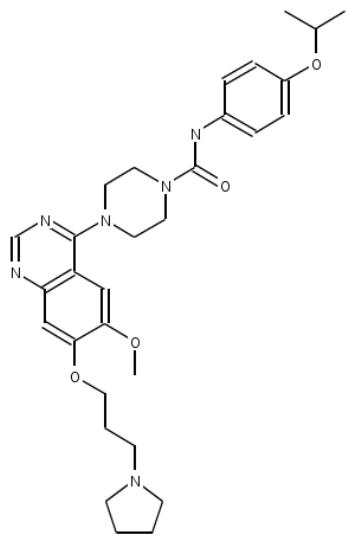
Neighborhood Similarity

PDGFR Inhibitors

- The prediction for **64** is not improved by local regression
- Neighborhood contained 19 molecules, average Tanimoto similarity was 0.94
- **58** has a Tanimoto similarity of 1.0 to **64**
- Range of activities was 1.41 log units
- Given significant structural similarity but differing activity values, local regression fails

64





58

Some caveats . . .

- Sparse training data can lead to larger prediction errors
- Requires that there be structural differences between groups of molecules
- Due to its focus on individual query molecules and the absence of an *a priori* model, structure-activity trends cannot be extracted

Outline

Future Work

- Develop a strategy to choose separate descriptor sets for the neighborhood detection and regression steps
  - Investigate fingerprints for the neighborhood selection
  - Use a large set of descriptors to select the neighborhood and use stepwise selection for the regression step
- Investigate radius based NN methods for neighborhood detection
  - Might be better for sparse datasets
  - Could indicate that no model can be built
- Flag query points with highly distorted neighborhood descriptor distributions
  - Use the Shapiro–Wilk test for normality

Summary

- Local regression generally shows improved accuracy
- Dependent on the nature of the neighborhood around a query point
- Can be performed very rapidly making it good for large datasets that are increasing in size (HTS data)
- No need to perform *apriori* clustering
- The correlation between neighborhood similarity and predictive ability is not significant