A Tiered Screening Protocol for the Discovery of Structurally Diverse HIV Integrase Inhibitors

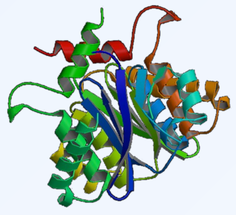
Rajarshi Guha, Debojyoti Dutta, Peter C. Jurs, Ting Chen,
Raveendra Dayam, Nouri Neamati

Department of Chemistry
The Pennsylvania State University

Department of Computational Biology
University of Southern California

School of Pharmacy
University of Southern California

1 Introduction



HIV Integrase Core Domain (1BIS)

Background

- Most drugs target HIV reverse transcriptase or protease
- HIV integrase is vital for viral replication
- No drugs have been approved for HIV integrase

Previous Approaches

- Pharmacophores
- Docking

Disadvantages

- Requires a reliable receptor structure
- Computationally intensive
- Not necessarily diverse

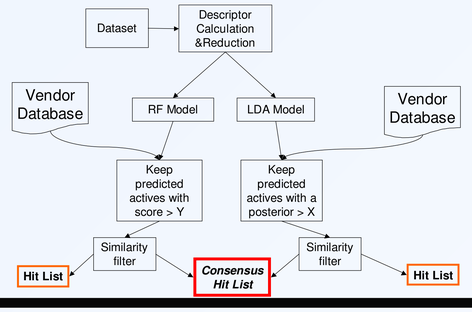
2 Goals

- High Speed
  - Be able to process large libraries rapidly
  - Avoid docking till it is required
  - Try and use connectivity information only
- Reliability
  - Use a consensus approach for predictions
  - Use molecular similarity
- Novelty
  - Try to obtain a diverse set of hits
  - Try to obtain hits suitable for lead hopping

3 Datasets & Tools

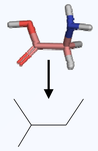
- Training Data
  - Curated dataset
  - 529 inactives
  - 395 actives
- Vendor Database
  - 50,000 compounds
  - 2D Structures
- Software
  - MOE
  - R 2.2.0 & 1.9.0
  - Various packages: *MASS, randomForest, fingerprint*
- *Hardware*
  - *Hammer for DB screening*
  - *LionXO for model search*

4 Overview of the Screening Protocol



5 Descriptor Calculations

- Restricted to topological descriptors
- Calculated 142 descriptors
- Removed correlated and low variance descriptors
- Final reduced pool had 45 descriptors



- Topological descriptors only consider 2D connectivity
- This converts a molecular structure to a directed graph
- Atom / bond identity may be considered by vertex / edge weights
- Very fast to compute

6 Predictive Models

Linear Discriminant Analysis | ***Random Forest***

Why?

- •Simple
- •May be sufficient

- •No feature selection
- •Does not overfit
- •May capture non-linearities

How?

- •Used a GA to search for descriptor subsets
- •Used a 6-descriptor model

- •Number of trees = 500
- •Number of features sampled = 6

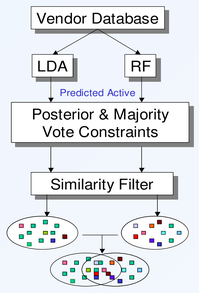
Accuracy?

On whole dataset	72%
On TSET/PSET	72% / 71%
With leave-10%-out	71%

On whole dataset	72%
On TSET/PSET	75% / 30%

7 Consensus Predictions

- We first consider compounds predicted active by both models
  - For the LDA model we consider compounds with a posterior > P
  - For the RF model we consider compounds which had a majority vote > M
- We then apply the similarity filter
- This results in two hit lists
- A final hit list is obtained from the intersection of the hit lists for the individual model



8 Similarity Filter (I)

- The goal is to predict actives from the vendor database
- Evaluate 166 bit MACCS fingerprints
- The Tanimoto Similarity between two compounds is defined by

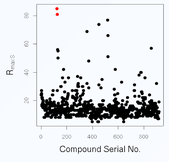
$$S = \frac{N_c}{N_a + N_b - N_c}$$

- For a compound predicted active
  - Calculate average similarity to TSET actives (S_1)
  - Calculate average similarity to TSET inactives (S_2)
  - Select compounds where

$$S_1 - S_2 \geq \varepsilon$$

9 Suggesting Good Leads

- How can we further prioritize the hit list?
  - Look for TSET actives that are not in the bulk of the dataset
  - These compounds may be good starting points for further modifications
  - Evaluate similarity of the hit list compounds to these isolated TSET compounds
- Hit list compounds most similar to the most isolated TSET active may be a good lead



A RNN plot which displays the dataset in terms of the *outlyingness* of each compound. Points at the top of the plot are the most isolated in the dataset. The two points in red are active. Thus new compounds similar to the red points may be good starting points for lead hopping.

Saeh, J.C.; Lyne, P.D.; Takasaki, B.K.; Cosgrove, D.A.; *J. Chem. Inf. Comput. Sci.*, **2005**, *45*, 1122-1133.
Guha, R.; Dutta, D.; Jurs, P.C; Chen, T.; *J. Chem. Inf. Model.*, submitted

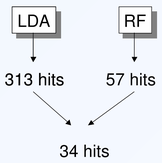
10 Time Considerations

- Model development
  - Descriptor calculation is rapid and a one time event
  - Building individual LDA or RF models is fast
  - Time required to obtain *optimal* model can be large when a GA is used (partly due to interpreted code)
  - Predictions for 50,000 compounds < 1min
- Similarity calculation is very time consuming
- Detecting spatial outliers is slow
  - Can be improved with approximate NN algorithms

Datar, M; Immorlica, N.; Indyk, P.; Mirrokni, V.S.; *Proc. 20th Symp. Comp. Geom.*, **2004**, ACM Press, pages 253-262
Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T.; *J. Chem. Inf. Model.*, **2006**, *46*, 321-333

11 Results

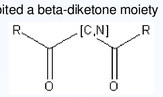
- Parameter Settings
  - P, M > 0.7
  - $\varepsilon \geq 0.01$
- Of the 34 hits, 7 compounds have a similarity > 0.5 with the most outlying TSET active
- We obtained 66 more hits from an ensemble of LDA models



- Average similarity of hits = 0.64
- Not significantly diverse
- None of the hits were in common with the pharmacophore models

12 The Search for Beta Diketones

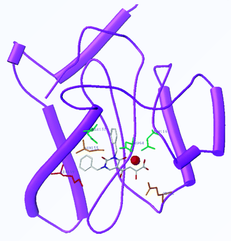
- A recently published inhibitor exhibited a beta-diketone moiety
- None of our hits had this moiety

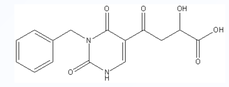


- Searched the vendor DB for structures with this feature and predicted them
  - LDA model : 244 actives (posterior > 0.8)
  - RF model : 4 actives (score > 0.85)
  - The intersection set contained 4 compounds
- We missed these hits due to our similarity constraints being too strict

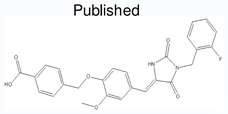
Nair, V.; Chi, G.; Ptak, R.; Neamati, N.; *J. Med. Chem.*, **2006**, *45*, 445-447

13 Similarity to the Published Compound





Published



Predicted

Nair, V.; Chi, G.; Ptak, R.; Neamati, N.; *J. Med. Chem.*, **2006**, *45*, 445-447

14 Future Work

- Investigate similarity to known inhibitors in terms of pharmacophore similarity
- Dock our best hits (may not be conclusive)
- Build predictive models using *local* techniques such as local lazy regression
- Investigate the distribution of vendor compounds in descriptor space
- Cluster the vendor database and predict representative members of clusters
- Perform assays!

Atkeson, C.G.; Moore, A.; Schall, S.; "Locally Weighted Learning", *Artificial Intelligence Review*, **1995**, *11*, 11-73
Bontempi, G.; Birattari, M.; "Local Learning for Iterated Time-Series Prediction", *Int. Conf. Mach. Learn.*, **1999**, pages 32-38
Guha, R.; Dutta, D.; Jurs, P.C.; Chen, T.; *J. Chem. Inf. Model.*, submitted