Kinome-wide Activity Models from Diverse High-Quality Datasets

Stephan C. Schürer*,1 and Steven M. Muskal2

1Department of Molecular and Cellular Pharmacology, Miller School of Medicine and Center for Computational Science, University of Miami, Miami, FL 33136, USA.

2Eidogen-Sertanty, Inc. 3460 Marron Rd #103-475, Oceanside, CA 92056, USA.

Steven Muskal
Eidogen-Sertanty, Inc
smuskal@eidogen-sertanty.com
# FDA Approved Protein Kinase Inhibitors

Table 1. FDA Approved Protein Kinase Inhibitors (as of March 2012)

<table>
<thead>
<tr>
<th>generic (brand) name</th>
<th>year of approval</th>
<th>company</th>
<th>indication</th>
<th>target kinase</th>
</tr>
</thead>
<tbody>
<tr>
<td>imatinib (Gleevec)</td>
<td>2001</td>
<td>Novartis</td>
<td>chronic myeloid leukemia (CML)</td>
<td>Abl, c-Kit, PDGFRα/β</td>
</tr>
<tr>
<td>gefitinib (Iressa)</td>
<td>2003</td>
<td>AstraZeneca</td>
<td>non-small-cell lung carcinoma (NSCLC)</td>
<td>EGFR, EGFR</td>
</tr>
<tr>
<td>erlotinib (Tarceva)</td>
<td>2004</td>
<td>Genetech, OSI</td>
<td>NSCLC, pancreatic cancer</td>
<td>EGFR</td>
</tr>
<tr>
<td>sorafenib (Nexavar)</td>
<td>2005</td>
<td>Bayer, Onyx</td>
<td>hepatocellular carcinoma, renal cell carcinoma (RCC)</td>
<td>Raf, VEGFR2/3, c-Kit, PDGFRβ</td>
</tr>
<tr>
<td>sunitinib (Sutent)</td>
<td>2006</td>
<td>Pfizer</td>
<td>gastrointestinal stromal tumor (GIST), RCC</td>
<td>c-Kit, VEGFR, PDGFR, FLT3</td>
</tr>
<tr>
<td>dasatinib (Sprycel)</td>
<td>2006</td>
<td>Bristol-Myers Squibb</td>
<td>CML</td>
<td>Abl, c-Kit, PDGFR, Src</td>
</tr>
<tr>
<td>nilotinib (Tasigna)</td>
<td>2007</td>
<td>Novartis</td>
<td>CML</td>
<td>Abl, c-Kit, PDGFR, Src, ephrin</td>
</tr>
<tr>
<td>laptatinib (Tykerb)</td>
<td>2007</td>
<td>GlaxoSmithKline</td>
<td>breast cancer</td>
<td>EGFR, ErbB2</td>
</tr>
<tr>
<td>pazopanib (Votrient)</td>
<td>2009</td>
<td>GlaxoSmithKline</td>
<td>RCC</td>
<td>VEGFR, PDGFRα/β, c-Kit</td>
</tr>
<tr>
<td>vandetanib (Caprelsa)</td>
<td>2011</td>
<td>AstraZeneca</td>
<td>thyroid cancer</td>
<td>VEGFR, EGFR, RET</td>
</tr>
<tr>
<td>vemurafenib (Zelboraf)</td>
<td>2011</td>
<td>Roche, Plexxicon</td>
<td>CML</td>
<td>Abl, c-Kit, PDGFR, Src, ephrin</td>
</tr>
<tr>
<td>crizotinib (Xalkori)</td>
<td>2011</td>
<td>Pfizer</td>
<td>NSCLC (ALK +ve)</td>
<td>ALK, MET</td>
</tr>
<tr>
<td>ruxolitinib (Jakafi)</td>
<td>2011</td>
<td>Incyte</td>
<td>myelofibrosis</td>
<td>JAK1/2</td>
</tr>
<tr>
<td>axitinib (Inlyta)</td>
<td>2012</td>
<td>Pfizer</td>
<td>RCC</td>
<td>VEGFR, PDGFRβ, c-Kit</td>
</tr>
</tbody>
</table>

Outline

• Kinase Data (KKB)

• Regression Models
  ➡ Conclusions

• Naïve Bayesian Classifier Models
  ➡ Conclusions
Kinase Knowledgebase (Q2 2012) - “Hot Targets”

Kinase Targets of Clinical Interest

Eidogen-Sertanty KKB SAR Data Point Distribution

> 649,000 SAR data points curated from
> 7915 journal articles and patents
Kinase inhibitor structures and SAR data mined from
> 7915 journal articles/patents

- **KKB Content Summary (Q2 2012):**
  - # of kinase targets: > 480
  - # of SAR Data points: > 649,000
  - # of unique kinase molecules with SAR data: > 241,000
  - # of annotated assay protocols: > 25,472
  - # of all kinase inhibitors (with or without bio-activity data): > 586,000

- **KKB Growth Rate:**
  - Average 15-20K SAR data points added per quarter
  - Average 20-30K unique structures added per quarter
Kinase Summary Statistics - Q2 2012

Articles covered: 2,307 (+30)
Patents and patent applications covered: 5,608 (+93)
Total Number of Bio-activity data points: 649,384 (+31,602)
Total Number of unique molecules: 586,610 (+8601)
Total Number of unique molecules w/ assay data: 241,680 (+8601)
Total Number of assay protocols: 25,472 (+322)

Targets with largest increase in Data Points in Q2-12

<table>
<thead>
<tr>
<th>Target</th>
<th># Data Points added</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFR3</td>
<td>4625</td>
</tr>
<tr>
<td>KDR</td>
<td>4482</td>
</tr>
<tr>
<td>FGFR1</td>
<td>4448</td>
</tr>
<tr>
<td>FLT3</td>
<td>3047</td>
</tr>
<tr>
<td>TTK</td>
<td>1634</td>
</tr>
<tr>
<td>FGFR2</td>
<td>1428</td>
</tr>
<tr>
<td>FGFR4</td>
<td>1364</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>1254</td>
</tr>
<tr>
<td>PIK3CD</td>
<td>955</td>
</tr>
<tr>
<td>JAK3</td>
<td>920</td>
</tr>
<tr>
<td>MTOR</td>
<td>827</td>
</tr>
<tr>
<td>JAK2</td>
<td>533</td>
</tr>
<tr>
<td>PTK2</td>
<td>453</td>
</tr>
<tr>
<td>RPS6KB1</td>
<td>425</td>
</tr>
<tr>
<td>JAK1</td>
<td>387</td>
</tr>
<tr>
<td>ALK</td>
<td>361</td>
</tr>
<tr>
<td>AKT1</td>
<td>357</td>
</tr>
<tr>
<td>ROCK2</td>
<td>335</td>
</tr>
<tr>
<td>SYK</td>
<td>305</td>
</tr>
<tr>
<td>BRAF</td>
<td>266</td>
</tr>
<tr>
<td>GSK3B</td>
<td>251</td>
</tr>
<tr>
<td>LRRK2</td>
<td>219</td>
</tr>
<tr>
<td>EGFR</td>
<td>211</td>
</tr>
<tr>
<td>BTK</td>
<td>197</td>
</tr>
<tr>
<td>TYK2</td>
<td>188</td>
</tr>
<tr>
<td>IRAK4</td>
<td>178</td>
</tr>
<tr>
<td>PIK3CB</td>
<td>158</td>
</tr>
<tr>
<td>PIK3CG</td>
<td>147</td>
</tr>
<tr>
<td>PIM1</td>
<td>147</td>
</tr>
<tr>
<td>IKBKB</td>
<td>130</td>
</tr>
<tr>
<td>CDK2</td>
<td>117</td>
</tr>
<tr>
<td>MAPK1</td>
<td>108</td>
</tr>
<tr>
<td>ERBB2</td>
<td>94</td>
</tr>
<tr>
<td>CSF1R</td>
<td>73</td>
</tr>
<tr>
<td>MET</td>
<td>72</td>
</tr>
<tr>
<td>TGFBR1</td>
<td>68</td>
</tr>
<tr>
<td>PLK1</td>
<td>55</td>
</tr>
<tr>
<td>PIM3</td>
<td>54</td>
</tr>
<tr>
<td>CDK9</td>
<td>52</td>
</tr>
</tbody>
</table>
Kinase Data Used in this Study - Q4 2009 [Q2 2012]

- Biological Activity Data Points: > 437,000 [> 649K]
- Unique kinase molecules w/assay data: >162,000 [> 241K]
- Unique kinase molecules patents/articles: > 507,000 [> 586K]
- Number of unique kinase targets with assay data: 394 [480]
- Number of annotated assay protocols: 20,593 [25,472]
Data Pre-Processing

- **Starting point:** KKB-Q2 2009
- Only enzymatic (homogeneous) assays with defined target
- Only high quality data (IC50, Ki, Kd)
- Standardized chemical structures (salt forms, stereochemistry, E/Z geometry, tautomers, ionization)
- Kinase target Entrez Gene names and SwissProt accessions
- 233,667 unique data points (411 kinases)
- 126,114 unique chemical structures
KinomeScan Data (Experimental Validation Data)

- NIH HMS LINCS DataBase (Harvard Medical School LINCS center) [http://lincs.hms.harvard.edu/resources/software/hms-lincs-database/](http://lincs.hms.harvard.edu/resources/software/hms-lincs-database/)
  - The LINCS program develops a library of molecular signatures based on gene expression and other cellular changes in response to perturbing agents across a variety of cell types using various high-throughput screening approaches

- 25,064 total datapoints downloaded:
  - 60 unique compounds (43 with defined/known chemical structure) against 486 targets
  - Kinase activity screened at 10 μM concentration
  - Targets mapped to KKB targets by UniProt accessions
  - Data not in KKB

- Result: 4,796 datapoints from 43 compounds
Outline

• Kinase Data (KKB)

• Regression Models
  ➔ Conclusions

• Naïve Bayesian Classifier Models
  ➔ Conclusions
Quantitative Regression Models

- K nearest neighbors (kNN) [20 nn, Gaussian weighting 0.5]
- Partial least squares (PLS) [components restricted to 20]
- Works best with congeneric series
- Sensitive to outliers and noise
- 168 kinase datasets have $\geq 20$ Dps
- PipelinePilot - ECFP4 (circular) fingerprints as descriptors
- Full 10 fold cross validation for both PLS and kNN (20 repetitions), report $R^2$, $q^2$, RMSE
Figure 5. Quantitative regression models developed from 168 kinase datasets. $q^2$ values of kNN vs. PLS regression models and the number of kinase activity data points indicated by the circle size and color.
Figure 7. kNN and PLS activity predictors for 91 kinases ($q^2$ kNN > 0.4 and $q^2$ PLS > 0.25) by the number of data points; datasets include protein and non-protein kinases and all major kinase groups. kNN $q^2$ is shown by number of (unique) structure-data points. Symbol indicates protein vs. non-protein kinase, size is scaled by PLS $q^2$, colored by kinase group, and annotated by HUGO kinase gene symbol.
Conclusions - Regression Models

• Work well for many kinase data sets

• kNN performs slightly better than PLS

• Larger numbers of data points improve both PLS and kNN models

• Best results for kinases with ≥ 50 data points

• Regression models improve with increased activity range
Outline

• Kinase Data (KKB)

• Regression Models
  ➡ Conclusions

• Naïve Bayesian Classifier Models
  ➡ Conclusions
Laplacien-modified Naïve Bayesian Classifiers

- Scale linearly, work in high-dimensional spaces (no over-fitting), good for structurally diverse cmpds, multiple activity classes, robust to outliers

- Define data sets by unique kinase gene IDs with active compounds defined as pIC50 ≥ 6

- 189 kinase data sets with at least 10 active molecules

- Data were treated in two ways:
  - Known Active - Known Inactive (KA-KI)
  - Presumed Inactive (PI): 126,114 unique chemical structures - N_{inact}

- ECFP4 (circular) fingerprints

- Leave-one-out cross-validation and repetitive train/test evaluation measuring ROC and enrichment factors
Figure S1. ROC score of leave-one-out vs. randomized 75/25 (train/test) cross validation for 187 kinase KA-KI classifiers with at least 10 active samples. Dot size is scaled by the number of total compounds in each dataset and colored by the number of actives. 129 classifiers are shown that correspond to datasets with at least 40 actives and 111 with at least 70 actives. ROC scores for randomized 75/25 training/test validations are averages of 10 repetitions. Compare table S1.
Figure S2. Characterization of all 189 kinase KA-PI protein and non-protein kinase classifiers, including all major protein kinase groups. ROC scores are shown as a function of active samples. Shape by protein vs. non-protein kinase, color-coded by kinase group, scaled by number of active data points, annotated by HUGO kinase gene symbol. ROC scores increase significantly for classifiers based on datasets with more than 25 active compounds.
Figure 1. Characterization of 141 kinase KA-PI protein and non-protein kinase classifiers, including all major protein kinase groups. ROC scores are shown as a function of active samples. Shape by protein vs. non-protein kinase, color-coded by kinase group, scaled by number of active data points, annotated by HUGO kinase gene symbol.

* At least 25 actives
Figure 9. Aggregated predicted probabilities (EstPGood) of compounds being active against kinases (based on the KA-PI classifiers) as a function of the actual KINOMEScan percent inhibition ranges; by category of kinase group and protein vs. non-protein kinase; 4,796 activity data points for 43 compounds mapped to KA-PI models (not all compounds tested against the same number of targets).
Figure 8. Probability (EstPGood) of compounds being active against a kinase based on KA-PI kinase classifiers and actual KINOMEScan percent inhibition values (at 10 μM); compare supporting table S6. Kinase classified by groups and protein vs. non-protein kinases.
Conclusions - Classification Models

• Work very well for known actives - known inactives (KA-KI)

• Relevant and applicable for real world, highly unbalanced data sets (KA-PI)

• Leave-one-out ROC is a good guide of model quality

• Naïve Bayesian classification is excellent for the majority of kinases (>140)

• Performance increases markedly with ≥50 active compounds

➡ Very useful for virtual screening and rapid profiling
Eidogen’s iPhone, iPad, and Android Apps

See: eidogen.com, kinasedb.com or kinasedata.com
MobileApps Support Real Scientific Workflows

Bioactivity searching (e.g. kinase SAR)

Commercial availability

Synthesis planning
Acknowledgements

- Dr. Rajan Sharma and Prof. Stephan Schurer
  - iKinase
  - iKinasePro
  - iOncology

- Dr. Maurizio Bronzetti
  - Mobile Reagents
  - Reaction101
  - SPRESImobile

- Dr. Alex Clark
  - MMDSLib:
    - Mobile Reagents
    - iProtein
    - Reaction101
    - Yield101
    - SPRESImobile

- Dr. Tony Yuan
  - JSDraw:
    - iKinasePro
    - Mobile Reagents
    - iProtein

- Dr. Peter Löw, Dr. Josef Eiblmaier, et al.
  - Mobile Reagents
  - iProtein
  - SPRESImobile
Welcome to the Kinase Data Portal

The Kinase KnowledgeBase
The most accurate database of kinase inhibitors now available also in Instant IChEM format
Q1-2012 now released!

>617,000 Biological Data Point

>233,000 unique kinase inhibitors with annotated assay data

483 unique kinase targets with assay data

25,150 annotated Assay protocols

======= WATCH A DEMO ========

Last Updated on Thursday, 03 May 2012 05:47
Naïve Bayesian with Laplacien Correction

Naïve Bayes (features are conditionally independent):

\[ P(C \mid F_1, \ldots, F_n) = k \prod_{i=1}^{n} \frac{P(C \mid F_i)}{P(C)} \]

\[ P(C) = \frac{A}{T} \]

\[ P(C \mid F_i) = \frac{A_{Fi}}{T_{Fi}} \]

\( A \): Number of active samples
\( T \): Total number of sample
\( A_{Fi} \): Active sample with feature \( F_i \)
\( T_{Fi} \): Total sample with feature \( F_i \)

Adding virtual samples for each feature:

\[ P_{corr}(C \mid F_i) = \frac{A_{Fi} + P(C)K}{T_{Fi} + K} \]

Estimating active virtual samples using baseline probability

Sample frequency 1/P(C) or T/A (Laplacien correction):

\[ P_{final}(C \mid F_i) = \frac{A_{Fi} + 1}{T_{Fi} + T/A} \]

Pipeline Pilot implementation:

\[ \log(P(C \mid F_1, \ldots, F_n)) = K + \sum_{i=1}^{n} \log(P_{final}(C \mid F_i)) \]
Classification Model Evaluation

Receiver operating characteristic:

\[ ROC = \frac{S}{1 - SP} \]

Sensitivity (true positive rate):

\[ S = \frac{TP}{N_{act}} \]

Specificity (true negative rate):

\[ SP = \frac{TN}{N_{inact}} \]

Enrichment:

\[ EF = \frac{TP + FP}{\frac{TP}{N_{act}} + \frac{FP}{N}} \]

TP: True positives
FP: False positives
\( N_{act} \): Number of active samples
\( N_{inact} \): Number of inactive samples

Report enrichment at 0.1 %, 0.5 %, 1%, etc.
MobileApps: Worldwide Marketing Vehicles!

~ 30,000 People Use Eidogen Mobile Apps