Surveying ligand- and target-based similarities within the Kinome

Stephan Schürer & Steven Muskal
Kinase SAR Knowledgebase – Hot Targets

Kinase Targets of Clinical Interest

Eidogen-Sertanty KKB
SAR Data Point Distribution

>362,000 SAR data points curated from
>4,270 journal articles and patents
>130 Bayesian QSAR Models
About Eidogen-Sertanty

• Knowledge-Driven Discovery Solutions Provider
  • Formed in March 2005 when Sertanty (Libraria→Sertanty 2003) acquired Eidogen (Bionomix 2000)
  • >$20M Invested in Technology Development
  • 12 FTE’s
  • Worldwide Customerbase
  • Cash-Positive

• Chemogenomic Databases & Analysis Software
  • TIP™ - Structural Informatics Platform
  • KKB™ - Kinase SAR and Chemistry Knowledgebase
  • CHIP™ - Chemical Intelligence Platform

• DirectDesign™ Discovery Collaborations
  • In Silico Target Screening (“Target Fishing” and Repurposing)
  • Target and compound prioritization services
  • Fast Follower Design: Novel, Patentable Leads
TIP Algorithm Engine

> 400K Sequences

> 158K Chains & Models

> 388K Sites

> 33M Sequence Similarities

> 69M Structure Similarities

> 62M Site Similarities
STRUCTFAST™

Structure Realization Utilizing Cogent Tips From Aligned Structural Templates

Basic Principle: Gaps known to exist should not be strongly penalized.

Leverages experimental structure and structural alignment data to create better alignments


2) STRUCTFAST: Protein Sequence Remote Homology Detection and Alignment Using Novel Dynamic Programming and Profile-Profile Scoring Proteins. 2006 64:960-967
Geometric Site-Finding Algorithms Find Many Pockets
But they don’t know which pockets are important!

Evolutionary Trace Approach
Can’t clearly define site boundary
Not all conserved residues are functionally relevant

SiteSeeker combines both methods

Reliability & Confidence
We use proteins with apo- & co-crystal structures in the PDB to test the accuracy & reliability of method

Allows us to map SiteSeeker score to predict confidence!
(e.g. At this SiteSeeker score, 80% are “real” co-crystal sites)
→ Sites with <60% confidence are not stored in TIP
Weighted Clique Detection Algorithm
Importance of Points Related To Conservation In Multiple Sequence Alignment

Surface Atoms Assigned One of 5 Different Chemical Characters
Matching points increase the *SiteSorter* similarity score
**TIP Content**

- >75,000 Human Sequences
- >116,000 Total PDB chains (~50K PDBs)
- >42,000 Homology Models
- >194,000 PDB co-crystal sites
- >190,000 Predicted Sites (on PDBs & Models)

>33M Sequence Similarities
>69M Structural Similarities
>62M Site Similarities

Updated monthly with new PDBs and models:

**e.g. March 2006:**
- 661 new PDBs added
- 447 new models built
  - 153 had no previous structure in TIP
  - 294 had “better” models built

**e.g. July 2008:**
- 576 new PDBs added
- 1045 new models built

Automatically updated with new models as the PDB grows
Kinase Knowledgebase (KKB)

Kinase inhibitor structures and SAR data mined from

> 4278 journal articles/patents

- **KKB Content Summary (Q2-2008):**
  - # of kinase targets: >390
  - # of SAR Data points: > 362,000
  - # of unique kinase molecules with SAR data: >120,000
  - # of annotated assay protocols: >16,000
  - # of annotated chemical reactions: >2,300
  - # of unique kinase inhibitors: >465,000 (~340K enumerated from patent chemistries)

- **KKB Growth Rate:**
  - Average 15-20K SAR data points added per quarter
  - Average 20-30K unique structures added per quarter
Kinase Knowledgebase (KKB)

Kinase inhibitor structures and SAR data mined from > 4100 journal articles/patents

Kinase Validation Set

Three sizable datasets freely available to the research community

http://www.eidogen-sertanty.com/kinasednld.php

Average 20–30K unique structures added per quarter
The ATP site of LIMK1 shares a high level of homology with several well-studied kinases.
Kinome by Sequence
Kinase domain sequence similarities - MST
Kinome by SAR
Relating kinase targets by SAR

- Relationships derived from Bayesian categorization models
  - Adopted from Schuffenhauer *Org Biomol Chem* 2004 3256

- Bayesian categorization models built within PipelinePilot:
  - Kinase enzyme assay data, activity cutoff pIC50 > 6.5; all other compounds “negative”
  - Functional group connectivity fingerprints length 4
  - ROC > 0.7

- Bayesian feature weights (~10,000 features) extracted for each model

- Correlation matrix determined between Bayesian vectors

- Visualization via minimum spanning trees (Kruskal algorithm)
Kinase SAR Bayesian models

130 Kinase Enzyme Models
FCFP_4 fingerprints scaled by number of actives
Kinase target relationships by SAR – MST

130 kinase models

MST – all “similarities” > 0.27
SAR-based similarity vs. Sequence identity

![Graph showing SAR-based similarity vs. Sequence identity with compounds CDC2A/CHUK and FGFR2/FGFR3]
CDC2A and CHUK: > 90 ligands with activity against both targets
FGFR2 / FGFR3: no similar ligands

FCFP4 Tanimoto (all pairs)

Top active FGFR2:

Top active FGFR3:

One active
Neither active
Both active
Kinome by structure binding site similarities
Relating kinases by ATP binding-site similarity

- Human Kinase domain sequences extracted (Sugen, Swissprot, PFAM)
- Human Kinome (500 sequences) modeled using STRUCTFAST
  - Multiple models per sequence (subset of 263 presented here)
- Binding sites for all models computed (SiteSeeker)
- Binding site similarity scores computed (SiteSorter)
- Similarity scores normalized: $AB_{\text{Norm}} := \frac{AB}{(AA + BB - AB)}$
  - $AB$ – Site Similarity between sites A & B
  - $AA / BB$ – “Self Site” Similarity Scores
- Analysis and visualization with MST
Kinase Site Similarity Relationships – MST

263 kinases; MST – all “similarities” > 0.6
Sequence vs. Site Similarity

MAP3K8 / NTRK1

MAST4 / PIK3R4

Same template
Different template
Similar sites – different sequences

- STE_STE11_MAP3K8: template 1u5rA
- TK_Trk_TRKA (NTRK1): template 1ir3A

MAP3K8

NTRK1

LGKGAY.V.A.K.V.E.V.MEFV.GGS.S.D.NN.M.D
LGEFGEF V A K - E V FE-M -GD - D -N L D
Similar sites – different site AA composition

- AGC_MAST_MAST4: template 1z5mA
- Other_VPS15_PIK3R4: template 1z5mA
- Site sequence similarity: 0.2
- Normalized (physicochemical) site similarity: 0.78
What did we learn?

- Expected global trend: Similar sequence results in physicochemical- and fold-similar binding sites
- Dissimilar sequences do not always result in different binding sites
- Binding site similarities group in “patches” by domain sequence similarity
  - Subtle differences in site relationships among groups and sub-types
- Modeling templates influence results:
  - For many kinases no experimental structures exist, but can be modeled
  - Growing body of structural information will optimize the picture
- Body of selective Kinase compounds continues to grow
- In principle, small molecules can be optimized to differentiate between very similar (sequence) kinases
Conclusions and Next steps

- Quantifying similarity relationships within the Kinome can provide insight in early Kinase drug development

- Similarity within the Kinome should consider SAR-based and structure-based binding site similarity (v. domain sequence-based similarity)

- Next steps include
  - Analyze trends with respect to DFG-In/DFG-out
  - Quantify template effects
  - Investigate effects of site size and predicted vs. templated sites
Acknowledgements

• Stephan Schürer
• Kevin Hambly
• Joe Danzer
• Brian Palmer
• Derek Debe
• Aleksandar Poleksic

• Accelrys/Scitegic - Shikha Varma-O'Brien/Ton van Daelen

Contact
Steven Muskal
Chief Executive Officer
smuskal@eidogen-sertanty.com