Organizing and Mining HTS Data using Data Pipelining

Robert Brown, David Rogers and Andrei Caracoti, SciTegic Inc

ALA LabFusion 2004, Boston
Outline

• Introduction to data pipelining

• Methods
  – Extended connectivity fingerprints
  – Bayesian learning

• Case study
  – Data mining the NCI AIDS data set
  – Simulating screening prioritization
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components
Data Pipelining

- A powerful new paradigm for data processing
- Pipelines guide the flow of data through a network of modular computational components
Data Pipelining

• A powerful new paradigm for data processing
• Pipelines guide the flow of data through a network of modular computational components
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice
Data pipelining enables

- Processing of data from multiple disparate data sources
- Integration of disparate applications
- Rapid processing of large amounts of data
- Automated execution of routine processes
- Capture of best practice
Outline

• Introduction to data pipelining

• Methods
  – Extended connectivity fingerprints
  – Bayesian learning

• Case study
  – Data mining the NCI AIDS data set
  – Simulating screening prioritization
Extended Connectivity Fingerprints (ECFP)

• A new descriptor for molecular characterization

• Goals of the fingerprint
  – Be comprehensive – encode “all” features within a structure
    • do not rely on a pre-defined dictionary of features
    • encode tertiary/quaternary information (c.f. path fingerprints)
    • encode substitution patterns to the fragment
  – Create an interpretable model
    • Each bit in the fingerprint should represent a single decodable feature
  – Be fast to calculate
    • Model building and especially virtual screening should be fast processes
The FP Generation Process

• Process based on the Morgan algorithm
  – One of the first methods developed for computational chemistry

• Each atom is given an initial atom code
  – ECFP: Specific atom typing
  – FCFP: Abstract functional role of atom

• A number of iterations are performed
  – Each atom collects information from its neighbors
    • N iterations define structures 2N bonds wide
  – Resulting feature is mapped into a $2^{32}$ address space
Assignment of Initial Atom Codes

- ECFPs
  - Atom type
  - Atom charge
  - Atom mass
  - Valence
  - Number of bond to non-hydrogens
  - Number of bonds to hydrogens

- Variant is to use the 120 AlogP atom types
Extending the initial atom codes

- Record (bond-type, atom-type) codes for each neighbour
- Sort to avoid order dependency
- Apply hashing function to map to a single number in the $2^{32}$ address space (~4 billion bits)
- Chance of collisions is extremely low
ECFP: Generating the Fingerprint

- Iteration is repeated desired number of times
- Codes from all iterations are collected
- Duplicate bits are removed
- Information gain diminishes after a few iterations

<table>
<thead>
<tr>
<th>Molecule</th>
<th>ECFP_0</th>
<th>ECFP_2</th>
<th>ECFP_4</th>
<th>ECFP_6</th>
<th>ECFP_8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-182236392</td>
<td>-182236392</td>
<td>-182236392</td>
<td>-182236392</td>
<td>-182236392</td>
</tr>
<tr>
<td></td>
<td>642810091</td>
<td>642810091</td>
<td>642810091</td>
<td>642810091</td>
<td>642810091</td>
</tr>
<tr>
<td></td>
<td>-1100000244</td>
<td>-1100000244</td>
<td>-1100000244</td>
<td>-1100000244</td>
<td>-1100000244</td>
</tr>
<tr>
<td></td>
<td>-1074141656</td>
<td>-1074141656</td>
<td>-1074141656</td>
<td>-1074141656</td>
<td>-1074141656</td>
</tr>
<tr>
<td></td>
<td>1572579716</td>
<td>1572579716</td>
<td>1572579716</td>
<td>1572579716</td>
<td>1572579716</td>
</tr>
<tr>
<td></td>
<td>-1997021792</td>
<td>-1997021792</td>
<td>-1997021792</td>
<td>-1997021792</td>
<td>-1997021792</td>
</tr>
<tr>
<td></td>
<td>1996767644</td>
<td>1996767644</td>
<td>1996767644</td>
<td>1996767644</td>
<td>1996767644</td>
</tr>
<tr>
<td></td>
<td>-175146122</td>
<td>-175146122</td>
<td>-175146122</td>
<td>-175146122</td>
<td>-175146122</td>
</tr>
<tr>
<td></td>
<td>852414842</td>
<td>852414842</td>
<td>852414842</td>
<td>852414842</td>
<td>852414842</td>
</tr>
<tr>
<td></td>
<td>-2099970318</td>
<td>-2099970318</td>
<td>-2099970318</td>
<td>-2099970318</td>
<td>-2099970318</td>
</tr>
<tr>
<td></td>
<td>-932108170</td>
<td>-932108170</td>
<td>-932108170</td>
<td>-932108170</td>
<td>-932108170</td>
</tr>
<tr>
<td></td>
<td>-1564392544</td>
<td>-1564392544</td>
<td>-1564392544</td>
<td>-1564392544</td>
<td>-1564392544</td>
</tr>
<tr>
<td></td>
<td>1571214559</td>
<td>1571214559</td>
<td>1571214559</td>
<td>1571214559</td>
<td>1571214559</td>
</tr>
<tr>
<td></td>
<td>1451403962</td>
<td>1451403962</td>
<td>1451403962</td>
<td>1451403962</td>
<td>1451403962</td>
</tr>
<tr>
<td></td>
<td>284029635</td>
<td>284029635</td>
<td>284029635</td>
<td>284029635</td>
<td>284029635</td>
</tr>
<tr>
<td></td>
<td>-948152242</td>
<td>-948152242</td>
<td>-948152242</td>
<td>-948152242</td>
<td>-948152242</td>
</tr>
<tr>
<td></td>
<td>-1555299234</td>
<td>-1555299234</td>
<td>-1555299234</td>
<td>-1555299234</td>
<td>-1555299234</td>
</tr>
<tr>
<td></td>
<td>-281505363</td>
<td>-281505363</td>
<td>-281505363</td>
<td>-281505363</td>
<td>-281505363</td>
</tr>
<tr>
<td></td>
<td>-344170121</td>
<td>-344170121</td>
<td>-344170121</td>
<td>-344170121</td>
<td>-344170121</td>
</tr>
</tbody>
</table>
FCFP: Functional-Class Fingerprints

- Use the role of an atom in the initial atom code rather than the atom type
  - Halogens give the same code
  - Hydrogen bond donors equivalent
  - Hydrogen bond acceptors equivalent

FCFP Atom code bits from:
1: Has lone pairs
2: Is H-bond donor
4: Is negative ionizable
8: Is positive ionizable
16: Is aromatic
32: Is halogen
FCFP: Generating the Fingerprint

- Again, the information gained by reaching out further diminishes.
ECFPs and FCFPs

- New class of fingerprints for molecular characterization
  - Each bit represents the presence of a structural (not substructural) feature
  - Multiple levels of abstraction contained in single FP

- Large but sparse
  - Typical molecule generates 100s - 1000s of bits
  - Typical library generates 100K - 10M different bits.

- Fast
  - Generated at 10,000 mols/sec (2GHz PC)
  - Tanimoto pairwise similarities at ~500K comparisons/sec
Outline

• Introduction to data pipelining

• Methods
  – Extended connectivity fingerprints
  – Bayesian learning

• Case study
  – Data mining the NCI AIDS data set
  – Simulating screening prioritization
Bayesian Learning

• Build a model which estimates the likelihood that a given data sample is from a "good" subset of a larger set of samples (classification learning)

• Ideal for vHTS applications
  – Efficient:
    • Fast & scales linearly with large data sets
  – Robust:
    • works for a few as well as many ‘good’ examples
  – Unsupervised:
    • no tuning parameters needed
  – Multimodal:
    • can model broad classes of compounds
    • multiple modes of action represented in a single model
The Model

- Input is a training set with descriptors, a response variable and a test for good
- A feature is a binary attribute of a data record
  - For molecules: a property range or a fingerprint bit
- A count of each feature is kept:
  - Over all the samples
  - Over all samples that pass the test for good
- Normalized probability is calculated for each feature
  - log(Laplacian corrected probability)
- The normalized probabilities are summed over all features to give the relative score
An example model

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Equation &quot;MAOInhibitorLike&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Features from: (&quot;FCFP_6&quot; LongFingerprintType)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Features from: (&quot;ALogP&quot; DoubleType)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Features from: (&quot;Molecular_Veight&quot; DoubleType)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Features from: (&quot;Num_H_Donors&quot; LongType)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Features from: (&quot;Num_H_Acceptors&quot; LongType)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Features from: (&quot;Num_RotatableBonds&quot; LongType)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Feature Statistics:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Property &quot;FCFP_6&quot;:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Total # of features in all samples: 250988 in subset: 4004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>POSITIVE BINS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Bin ID</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
</tr>
<tr>
<td>16</td>
<td>Bin Value</td>
<td>-1290796621</td>
<td>-821137192</td>
<td>638618274</td>
</tr>
<tr>
<td>17</td>
<td>Feature Co</td>
<td>23</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>Subset Co</td>
<td>15</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>Normalized</td>
<td>2.46003</td>
<td>2.323228</td>
<td>2.309748</td>
</tr>
<tr>
<td>51</td>
<td>NEGATIVE BINS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Bin ID</td>
<td>B1</td>
<td>B2</td>
<td>B3</td>
</tr>
<tr>
<td>53</td>
<td>Bin Value</td>
<td>565968762</td>
<td>-1549163031</td>
<td>394124770</td>
</tr>
<tr>
<td>54</td>
<td>Feature Co</td>
<td>759</td>
<td>360</td>
<td>325</td>
</tr>
<tr>
<td>55</td>
<td>Subset Co</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>56</td>
<td>Normalized</td>
<td>-2.585342</td>
<td>-1.908514</td>
<td>-1.82208</td>
</tr>
</tbody>
</table>
Normalized Probability

• Given a set of $N$ samples

• Given that some subset $A$ of them are good (‘active’) – Then we estimate for a new compound: $P(\text{good}) \sim A / N$ – For a new feature to the model, this is our base estimate.

• Given a set of binary features $F_i$ – For a given feature $F$: • It appears in $N_F$ samples • It appears in $A_F$ good samples
  – Can we estimate $P(\text{good} \mid F) \sim A_F / N_F$? • Different features are sampled different numbers of times • Error gets worse as $N_F \to \text{small}$
Normalized Probability

• Solution: renormalize probabilities to baseline
  – Can be thought of as adding a single sample at baseline probability

• \[ P'(\text{good} \mid F) = \frac{AF + P(\text{good})K}{NF + K} \]
  – \( P'(\text{good} \mid F) \rightarrow P(\text{good}) \) as \( NF \rightarrow 0 \)
    • Assume: Most features have no relationship with activity
  – \( P'(\text{good} \mid F) \rightarrow \frac{AF}{NF} \) as \( NF \rightarrow \text{large} \)
    • Assume: more instances of the observation the more likely it is not an artifact

• If \( K = 1/P(\text{good}) \) this is the Laplacian correction
Outline

• Introduction to data pipelining

• Methods
  – Extended connectivity fingerprints
  – Bayesian learning

• Case study
  – Data mining the NCI AIDS data set
  – Simulating screening prioritization
Case Study: NCI AIDS data

- ~32,000 compounds selected for HTS
- Whole-cell assay
- Found 230 confirmed hits (“CA”)
- Represent 7 “activity classes” (modes of activity)
Results of Bayesian modeling

- Data split 50/50
  - Trained on 16,000 samples w/ 115 hits
  - FCFP_6, AlogP, MW, #HBA, #HBD, #Rot Bonds
- Results:
  - Would have discovered 80% of actives screening ~600 cmpds
  - Model learned multiple modes of activity
…ask more of your data

Robust to small numbers of hits

- Data split 5/95
  - Trained on ~1,600 samples, 14 hits
- Results:
  - Would have discovered 80% of actives screening ~3,000 cmpds
Robust to noise in hits

- Data split 50/50
  - 5% of negatives in training set reassigned as *false positives*
  - Data contained 115 true actives and ~800 false actives

- Results:
  - Would have discovered 80% of actives screening ~1,500 cmpds
Robust to weak actives for training

- Data split 50/50
  - All confirmed actives (CA) removed to test set
  - Trained on 130 confirmed *moderately active* (CM) compounds

- Results:
  - Weak actives aided in discovery of highly-active compounds
Search for false negatives

- False negatives problematic
  - Costly to retest negatives
  - Can disrupt SAR studies
- Experiment:
  - Take half of 230 hits and mark them as inactive
  - Build model with data set
  - Sort negatives for retest
Search for false negatives

- 85% found in top 5% of negatives
Outline

• Introduction to data pipelining

• Methods
  – Extended connectivity fingerprints
  – Bayesian learning

• Case study
  – Data mining the NCI AIDS data set
  – Simulating screening prioritization
Screening Prioritization

• HTS Screening strategies
  – Screen the entire compound collection
  – Iterative screening
    • Screen the entire collection in ordered subsets
    • Screen the collection in ordered subsets and stop when returns are diminishing (each screening point costs US$0.25 upwards)

• Iterative screening
  – Screen a subset
    • Random / Ordered
    • Build a model of the screening results
    • Prioritize the remaining compounds and select the next subset to screen
    • Update the model and select the next subset
    • Repeat until
      – No more compounds
      – Hit rate falls below a set level
Example

• Using the same NCI AIDS data set
  – Select a subset at random (384, 768, 1536, 3072)
  – “Screen” (i.e look up # actives)
  – Build a Bayesian model
  – Score the remaining compounds
  – Sort by score
  – Select the next subset of the same size and “screen”
  – Repeat until all molecules are “screened”

• Additional experiment
  – Restrict the initial random subset to weakly actives
Initial set contains CA (Actives)
...ask more of your data

Initial set contains only CM (Weak Actives)
Using the models

- Models can be used as virtual screens to filter
  - Virtual combichem libraries
  - Vendor files e.g. Maybridge
  - Vendor databases e.g. ACD
  - Corporate databases
Summary

• New fingerprint for molecular characterization
  – Fast, comprehensive and interpretable

• Bayesian learning
  – Successfully model HTS data
  – Robust to low hit rate and noise
  – Able to identify false negatives for retest

• Screening prioritization
  – Can identify most actives early in a screen

• Data pipelining provides the infrastructure for successful deployment of virtual screening