Classification and data mining of high-throughput protein crystallization screens

Christian Anders-Cumbaa & Igor Jurisica

Ontario Cancer Institute, Univ. of Toronto and Northeast Structural Genomics Consortium

In collaboration with HWI HTS

HTP Protein Crystallization Setup

Systematic  Objective  Reproducible

Getting principles & Optimization of:
- screen
- crystallization

Planning of:
- crystallization

Image Analysis

DB of precipitation indices (PIs)

Data Mining

Case-Based Reasoning
Feature Extraction

- **Laplacian**: find edges & calculate smoothness
  - computes the difference between each point \((x,y)\) in an image and the average of its neighbors
  - computes the rate of change of the intensity gradient in an image, i.e. the second derivative of the intensity of an image
- **Radon transform**: find straight lines
  - useful for finding the straight edges of crystals and needle crystals
- **Correlation filters**: matching exemplars
  - useful for finding microcrystals
- **Quadtree decomposition**: measure smoothness
  - splitting the image into four squares and examining the difference between the min&max pixel values, until the difference < threshold
  - computes number of squares examined
- **Euler number**: measure the topology of an image
  - the total number of objects in the image minus the number of holes in those objects

Image Classification

A 1536-well plate with 15 wells containing crystals may be classified with an accuracy exceeding 99% by a method claiming that no wells contain crystals.

Human-Machine consistency - 85% and 88%

**Machine classification**

<table>
<thead>
<tr>
<th></th>
<th>P1 classification</th>
<th></th>
<th>P1 classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>X</td>
<td>N</td>
</tr>
<tr>
<td>P2</td>
<td>N 970</td>
<td>110</td>
<td>P2</td>
</tr>
<tr>
<td></td>
<td>X 84</td>
<td>372</td>
<td></td>
</tr>
<tr>
<td>Consistency - 87%</td>
<td></td>
<td>Consistency - 86%</td>
<td></td>
</tr>
<tr>
<td>Human-Machine consistency - 85%</td>
<td>88%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Misclassifications

- False positives
  - Speckled precipitate (looks like microcrystals)
  - Non-crystal straight edges (e.g., skin effect)
  - Genuine crystals (expert errors)

- False negatives
  - Feathery crystals (no straight edges)
  - Small, low-contrast crystals
  - Crystals confused with the edge

Time Course Classification
Data Mining

Cocktail Details

Protein Properties

Precipitation Index

SOM & k-means

Cocktail similarity

Protein similarity

Association Mining

- discovering rules in itemsets:
  - IF X THEN Y  
    [support, confidence]
  - Horizontal (protein data + PI rows)
    - Pyrococcus furiosus => A [10, 1.0]
    - Escherichia coli => B [21, 1.0]
    - Homo sapiens => E [13, 1.0]
    - X000001461 => X000001461 [85, 0.944]
    - X000001316, X000001318, X000001461 => X000001461 [20, 1.0]
  - Vertical (cocktail data + PI columns)
    - spermine tetra-hcl => sodium cacodylate [21, 1.0]
    - citric acid => ACIDIC [40, 1.0]
    - UC00295, UC00507 => UC00326 [10, 1.0]
- Piecewise (protein data + cocktail data + PI element)
  - ARCHAEA, NEUTRAL, 4000, magnesium chloride hexahydrate => crystal [5, 1]
  - NEUTRAL, 4000, magnesium chloride hexahydrate => crystal [12, 1]
  - ammonium bromide, sodium acetate => crystal [13, 1]
- Piecewise (protein data + cocktail data + PI element) (machine data)
  - ARCHAEA, 8000, sodium acetate => crystal [112, 0.86]
  - EUKARYA, lithium chloride, mes => crystal [40, 1]

# Plates reviewed  # Plates with hits

<table>
<thead>
<tr>
<th></th>
<th>16</th>
<th>6</th>
<th>37.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eukaryota</td>
<td>80</td>
<td>54</td>
<td>67.5%</td>
</tr>
<tr>
<td>Bacteria</td>
<td>8</td>
<td>7</td>
<td>87.5%</td>
</tr>
</tbody>
</table>
Generalizations

- **Crystal**
  - Magnesium Chloride and Lithium Sulfate: high for OCI & HWI experiments
  - Calcium Acetate and Magnesium Chloride: mid-range on both data sets

- **Precipitate**
  - Potassium Phosphate and Calcium Acetate: high in both datasets
  - Lithium Sulfate scoring: mid-range on both data sets
  - Calcium Chloride, Magnesium Chloride and Zinc Acetate: high in OCI and average in HWI
  - Ammonium Sulfate: high in HWI but mid-range on OCI dataset

- **Concentration**
  - Calcium Chloride (OCI) - direct trend
  - NH4H2PO4 (HWI) - inverse trend
  - Sodium Cacodylate (HWI) - inverse trend
  - PEG4000 - direct trend

**Number of Proteins Crystallized**

CBR for Crystallization

![CBR Diagram]

1. PI for individual case as problem description
2. PI of a new protein
   - modified k-nearest neighbor retrieval
   - computational integration of results
   - solution

- distance of a given case from the others
Future Directions

- New image features
- Re-analysis of historical data
- Data integration
- Data mining
- Planning

Acknowledgments

M. Sultan, C. Cumbaa, X. Zhang, P. Rogers, R. Lu, D. Otasek, M. Popov, A. Patel

C. Arrowsmith, A. Edwards, G. Prive


J. Wrana, T. Brown, M. Tyers, A. Jurisicova

M. Kostyar, E. Xia

N. Arshadi, L. Zhang, J. Chae

J. Glasgow

IRIS, NSERC, OCRN

Percarn, NIH, IBM, GenomeCanada,

NCIC, CIHR, NASA, NIH,
The J.R. Oishei Foundation,
M.L. Wendt Foundation

URL: http://www.cs.utoronto.ca/~juris
http://www.hwi.buffalo.edu/Research/Facilities/CrystalGrowt.html