KDD-2001 Cup
The Genomics Challenge

Christos Hatzis, Silico Insights
David Page, University of Wisconsin
Co-chairs

August 26, 2001

Special thanks: DuPont Pharmaceuticals Research Laboratories for providing data set 1, Chris Kostas from Silico Insights for cleaning and organizing data sets 2 and 3

http://www.cs.wisc.edu/~dpage/kddcup2001/

acm SIGKDD Knowledge Discovery and Data Mining

The Genomics Challenge

- High throughput technologies in genomics, proteomics and drug screening are creating large, complex datasets
- Bioinformatics datasets are typically under-determined
  - very large number of features (complex domain)
  - small number of instances (high cost per data point)
- Multi-relational nature of data
  - reflect complex interactions between molecules, pathways and systems
  - Hierarchical organization of interacting layers
- Current tools and approaches do not adequately address the Genomics Challenge

Overview

- Cup organization
- Dataset description
  - Thrombin binding
  - Gene function/localization prediction
- Statistics
- Tasks and highlights
- Winners talk (3x10 min)

Cup Organization

- KDD-2001 Cup web site
  - Posting of datasets, Q&A, answer keys
- Schedule
  - Training dataset available: May 31
  - Question period 1: June 1-10
  - Test set available: July 13
  - Question period 2: July 13-24
  - Entries due: July 26
  - Winners notified: August 1
  - Results to participants: August 7
- Evaluation criteria
  - Task 1: weighted accuracy (average of true pos, true neg)
  - Tasks 2, 3: non-weighted accuracy

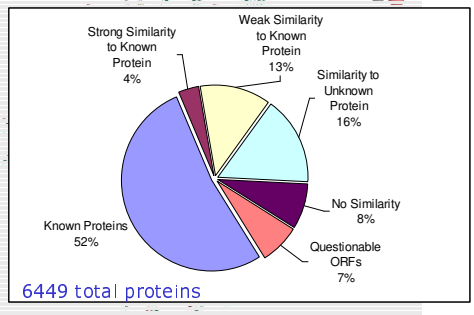
Dataset 1: Molecular Bioactivity

Dataset provided by DuPont Pharmaceuticals for the KDD-2001 Cup competition

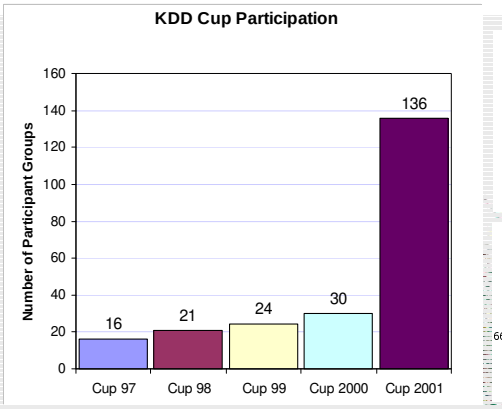
- Activity of compounds binding to thrombin
- Library of compounds included:
  - 1909 known molecules (42 actively binding thrombin)
- 139,351 binary features describe the 3-D structure of each compound
- 636 new compounds with unknown capacity to bind thrombin

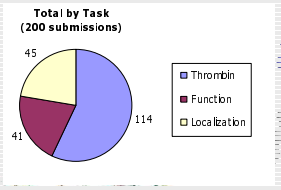
Dataset 2: Protein Functional Annotation

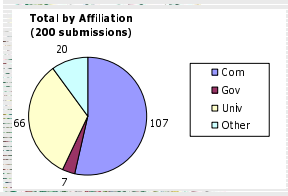
- Yeast Genome dataset
  - Data on the protein-protein interactions from MIPS database (Munich Information Centre for Protein Sequences)
  - Expression profiles: DeRisi et al. (1997) Science 278: 680
- Relational dataset
  - Gene information
  - Interaction information
- Predict function, localization of unknown proteins



Statistics: I. Participation

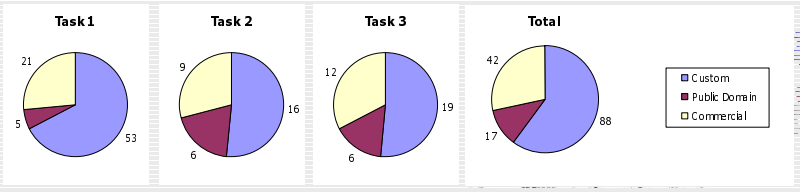






- 136 unique groups, 200 total entries by about 300-400 participants
- Almost 5-fold increase over previous years
- More than half of the entries from commercial sector

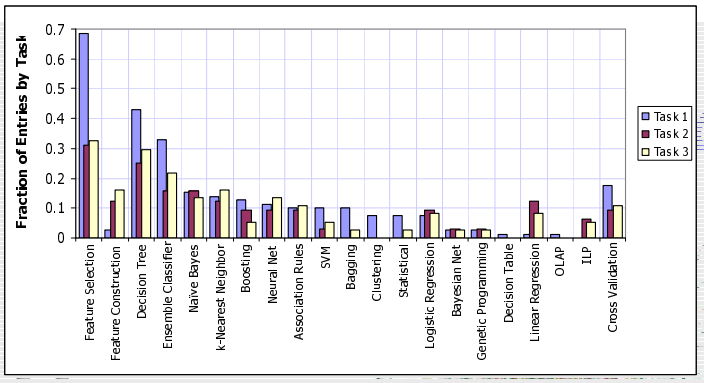
Statistics: II. Data Mining Software



Note: Statistics from 157 responders who provided details on their approach

- Mostly custom software was used
- Especially for task 1, where the number of features was too large for most commercial systems
- Gap points to need for commercial tools that can cope with bioinformatics datasets

Statistics: III. Algorithms

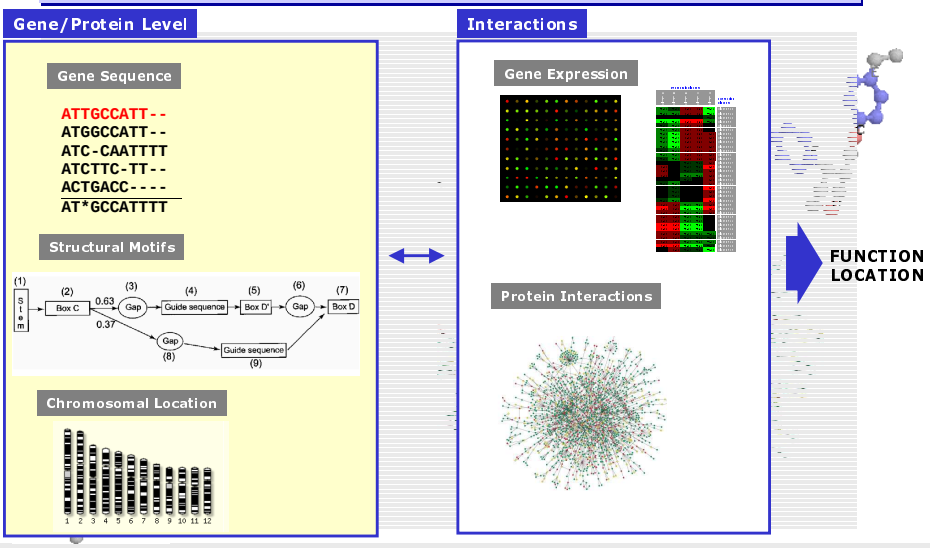


- Feature selection used in almost 70% of the entries for Task 1
- Ensemble classifiers based on more than one algorithm used extensively
- Decision trees among the most commonly used, with Naive Bayes and k-NN
- Cross-validation to deal with small dataset size

Task 1 Highlights

- Test set was challenging second round of compounds made by chemists -- change in distribution.
- Far more features than data points; can't run most commercial systems even with 1G RAM.
- Varying degrees of correlation among features.
- Better than 60% weighted accuracy is impressive.
- Pure binary prediction task, yet the winner is a Bayes net learning system (after feature selection).

Tasks 2 & 3: Relational Prediction



Task 2 Highlights

- Average of about 3 functions per protein.
- Multi-relational, as are many real-world databases.
- Yet top-scoring approaches were not pure relational learners.
- But top-scoring approaches did account for multi-relational structure of the data.
  - Krogel: novel form of feature construction to capture relational information in a feature vector.
  - Sese, Hayashi, and Morishita: instance-based learning, but using the interactions relation as part of the distance function.

Task 3 Highlights

- Similar to task 3, but only one localization per protein.
- Similar lessons.
- High overlap in top scorers for both tasks.
- Question: did anyone "bootstrap" by using their predictions for function to help predict localization, or vice-versa?

KDD-2001 Cup Winners

- Task 1: Jie Cheng, CIBC
- Task 2: Mark-A. Krogel, Magdeburg Univ.
- Task 3: Hisashi Hayashi, Jun Sese, and Shinichi Morishita, Univ. of Tokyo

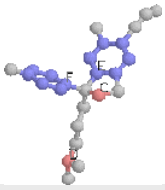
Task 1 Winner

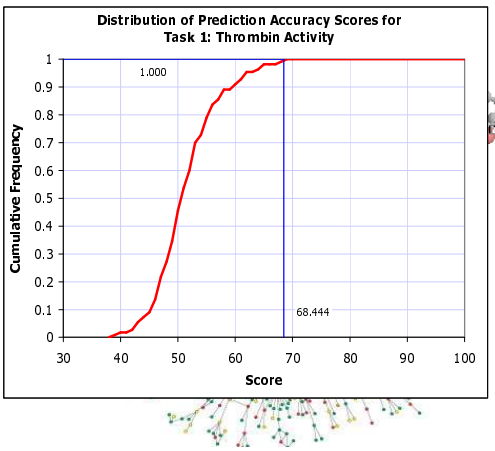
KDD Cup 2001 Results
Task 1: Thrombin

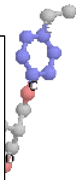
Name: Jie Cheng
Rank: 1
Weighted Accuracy: 68.4435
Accuracy: 71.1356

		Predicted	
		Positive	Negative
Actual	Positive	95	55
	Negative	128	356

True Positive Rate: 63.3%
True Negative Rate: 73.6%







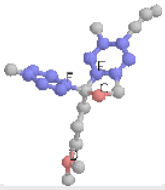
Task 2 Winner

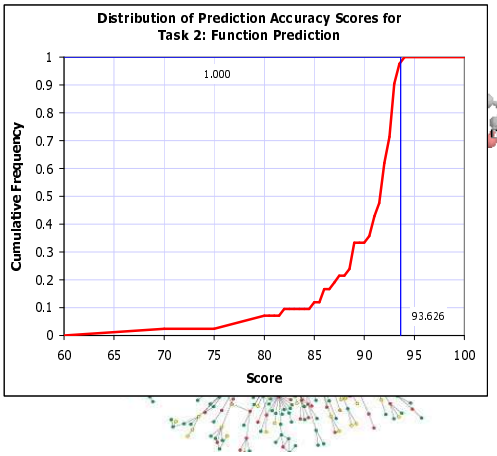
KDD Cup 2001 Results
Task 2: Function

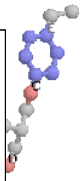
Name: Mark-A. Krogel
Rank: 1
Accuracy: 93.6258

		Predicted	
		Positive	Negative
Actual	Positive	690	282
	Negative	58	4304

True Positive Rate: 71.0%
True Negative Rate: 98.7%



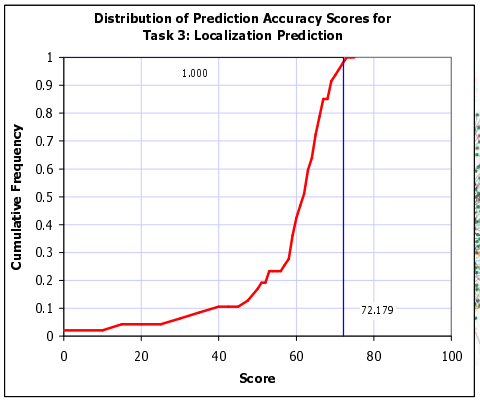




Task 3 Winner

KDD Cup 2001 Results
Task 3: Localization

Name: Hisashi Hayashi, Jun Sese, and Shinichi Morishita
Rank: 1
Accuracy: 72.1785

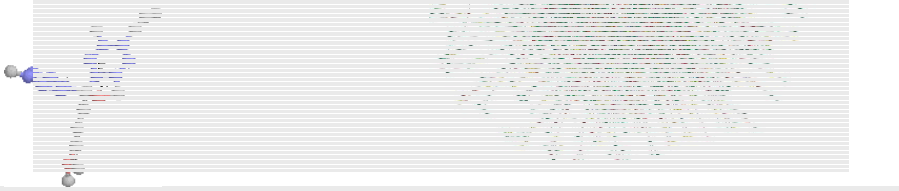


KDD-2001 Honorable Mentions

Task 1: Silander, Univ. of Helsinki

Task 2: Lambert, Golden Helix;
Sese & Hayashi & Morishita;
Vogel & Srinivasan, A.I. Insight

Task 3: Schonlau & DuMouchel & Volinsky & Cortes, RAND and AT&T Labs;
Frasca & Zheng & Parekh & Kohavi, Blue Martini

KDD-2001 Cup Winners

- **Task 1:** Jie Cheng, CIBC
- **Task 2:** Mark-A. Krogel, Magdeburg Univ.
- **Task 3:** Hisashi Hayashi, Jun Sese, and Shinichi Morishita, Univ. of Tokyo