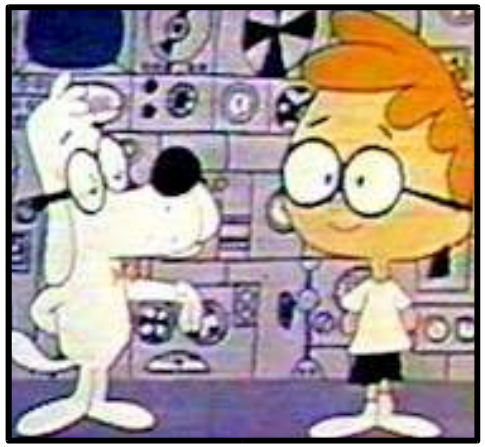
An open-source meta-analysis tool for protein localization prediction

or…



The conscientious developer: How to make a predictive tool and make friends at the same time

Jennifer Gardy
The Brinkman Laboratory
Simon Fraser University
British Columbia, Canada
jlgardy@sfu.ca

Slides @ http://www.sfu.ca/~jlgardy

PSORTb in One Slide

- **Predicts protein localization in bacteria**

Meta-analysis { *Homology to protein of known localization, frequent subsequence-based SVMs, signal peptides, secondary structure, patterns & motifs*

- **Web-based/standalone versions (GNU GPL)**
- **Flexible output, no parsers necessary**
- **Precomputed genome results available**
- **96% overall precision (most precise tool)**
- **High predictive coverage for proteomes**
  - » 57% Gram-negative, 75% Gram-positive
- **Latest release: June 2004**

www.psort.org/psortb

Poster I-80

Original version – PMID 12824378
New version – submitted

BOSC 2004 Glasgow, UK

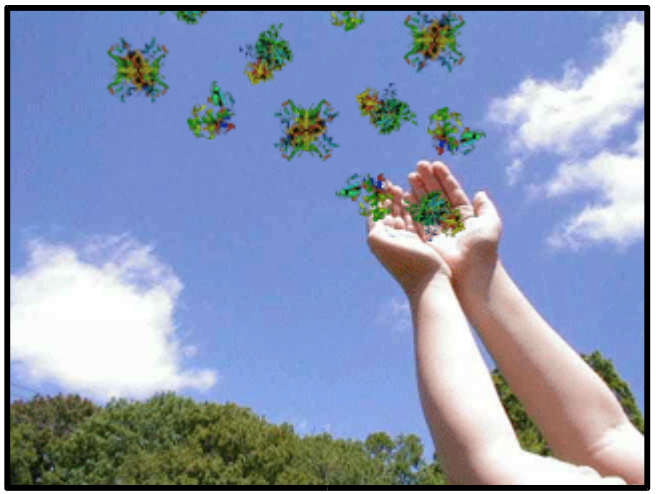
psortb

29 July, 2004

Lessons Learned

Despite was Mom said, sometimes asking nicely doesn't even help. The best tool for a job may be under restrictive licenses.

If you love something, set it free. Namely, your dataset.



Don't be shy. Publish confusion matrices to allow objective comparisons.

Pred. → / Actual ↓	X	Y
X	432	18
Y	4	639

Together We're Better

"…improved prediction performance by a consensus prediction method"

"…overall best predictions are obtained by combining predictions from these methods"

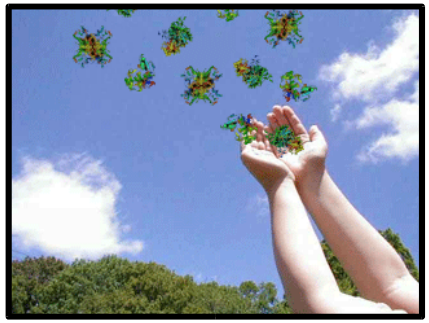
"..consensus predictions improve accuracy"

"…prediction accuracy has also been improved by combining more than one algorithm"

- **Consensus/meta-analytical methods = best**
- **What do to when the best tools available aren't open-source?**
  - *Single analysis method*
  - *"Next-best" tool*
  - *Grow-your-own*

Free as a ~~Bird~~ Protein

- **Freely-available datasets:**
  - *Promote advances in tool development*
  - *Allow comparisons*
  - *Promote your work*



- **PSORTdb used to develop & test new tools**
  - *ProteomeAnalyst, CELLO*

Taking the *Confusion* Out of *Confusion Matrix*

- **Grid comparing actual and predicted classes**

Pred. → / Actual ↓	X	Y
X	432	18
Y	4	639

- *Can't hide behind metrics*
- *Shows performance for each task, not just overall*
- *Gives users the power to evaluate tool objectively*

The Last Word

- **Closed source predictive tools can't be incorporated into open source meta-analyses**
  - Keep preaching the joys of open source. Convert!
  - Open source = keeps development moving forward
- **A dataset is a tool too, make it free**
  - Advances discovery, promotes your hard work
- **"Open source results" – confusion matrices reduce confusion**
  - No matrices = "Do they have something to hide?"
  - Should be required in a predictive method publication

Slides @ http://www.sfu.ca/~jlgardy PSORTb: http://www.psort.org/psortb