

**Updates from the
Protein Localization
Prediction Front:
PSORTb, PSORTdb and
Perspectives on
Predictive Methods**



Jennifer Gardy
The Brinkman Lab
9th September 2004



A Subcellular Localization Prediction Tool for Bacteria

- Initial release July '03 (PMID: 12824378)
- Web-based, standalone versions (GNU GPL)
- v.2.0 = Gram-negative & Gram-positive
- 2-3X more confident predictions than v.1
 - Frequent subsequence-based SVMs
 - She et al. *Proc. 9th ACM SIGKDD*, pp. 436-445. ACM Press, NY. (2003) <http://www.acm.org/pubs>
- 96% precision in 5X cross-validation
- 98% precision in comparative evaluation:
 - Proteome Analyst: 90%, CELLO: 72%

<http://www.psort.org/psortb>



http://db.psort.org: New Resource

- Database of prokaryotic localization information

- Experimental

- Literature
- 2165 proteins

- Computational

- PSORTb v.2.0
- 411029 protein
- 140 organisms

Display Options

Sort Columns: Protein Name, Organism, -- Sort Field 3 --

Columns NOT in Display: SwissProt Accession ID, Alternate Protein Name, Alternate Gene Name, Taxonomy ID, Class

Columns to Display: Protein Name, Gene Name, Organism, Phylum, Gram Stain

Results Per Page: 3

Download Options: TAB DELIMITED, FASTA

798 Results - 1 2 3 4 5 - Records 1-3 (Page 1 of 266 pages)

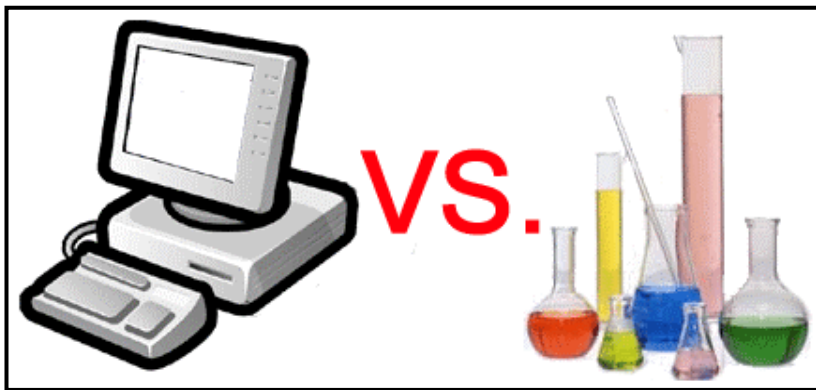
GI	Protein Name	Gene Name	Organism	Phylum	Gram Stain	Experimental SCL (terse)	Experimental SCL (verbose)	GO Accession ID
129099	15 kDa cysteine-rich outer membrane protein, serovar C	CRPA	Chlamydia trachomatis	Chlamydiae	negative	OuterMembrane	Outer membrane integral membrane protein	0009279
6831523	15 kDa cysteine-rich outer membrane protein, serovar D	CRPA	Chlamydia trachomatis	Chlamydiae	negative	OuterMembrane	Outer membrane integral membrane protein	0009279
129100	15 kDa cysteine-rich outer membrane protein, serovar E	CRPA	Chlamydia trachomatis	Chlamydiae	negative	OuterMembrane	Outer membrane integral membrane protein	0009279

798 Results - 1 2 3 4 5 - Records 1-3 (Page 1 of 266 pages)

- Search, browse and BLAST entry points
- Customizable results display, downloadable

Computational vs. Laboratory Data

- Incorrect results possible with both
- *"Data from large-scale analyses is often less reliable than data from small-scale analyses"*
 - N techniques, 1 entity = consensus
 - 1 technique, N entities = FPs, unusual cases
- *P. aeruginosa* exported proteins:



- 4 probable cytoplasmic proteins predicted as exported
- PhoA fusions indicated 9 probable cytoplasmic proteins as exported

Mission Impossible: Developing a Quality Open Source Consensus Predictive Method

- Consensus methods generally outperform single predictive methods
- What do you do when the best tool for the job isn't open source?
 - Abandon the idea of open source?
 - Go with the next or next-next best tool?
 - Make your own?
- *"Choosing an open-source license for your software promotes development of new and better tools and moves research forward"*



Acknowledgements

- **PSORTb/db**: Fiona Brinkman, Sébastien Rey, Matt Laird, Mike Acab, The Brinkman Lab, Rong She, Martin Ester, SFU Data Mining Group, Chris Walsh, Gabor Tusnady, Christophe Lambert.
- **Pseudomonas/PhoA**: Shawn Lewenza, Bob Hancock



For more information:

- www.psort.org – PSORTb & PSORTdb
 - psort-update@sfu.ca (“subscribe”) – maillist
- www.pathogenomics.sfu.ca/brinkman – Lab
- www.sfu.ca/~jlgardy – talk slides