Community-assisted genome annotation: The *Pseudomonas* example

Geoff Winsor, Simon Fraser University
Burnaby (greater Vancouver), Canada
Overview

_Pseudomonas_ Community Annotation Project (PseudoCAP)
  • Past and present

Annotation issues

_Pseudomonas_ Genome Database Version 2
  • New ability to compare annotations within or between species
  • New analyses/updates
PseudoCAP: *Pseudomonas aeruginosa*
Community Annotation Project

**Goals**
- Critical and conservative genome annotation
- Minimize project costs
- Capitalize on large *Pseudomonas aeruginosa* research community

**Solution**
- Community-aided and internet-based approach to continually updated, reviewed genome annotation

PseudoCAP: Past and present

- 61 researchers, 1741 annotations
- Experts provide clarification on gene names
- Tables of data submitted by e-mail.
- Annotations incorporated by 3 annotators
- Increased initial annotation quality/consistency

- 82 researchers, 1231 additional annotations (not including 24174 computationally-derived annotations)
- Submissions made using web-based forms or e-mail
- Annotations subject to review process
- Continually updated genome annotation
PseudoCAP approach to continually updated annotations

1. Modify an existing annotation
2. Correct a sequence
3. Annotate a new gene or sequence (individual annotations or batch submissions)

E-mail or WWW Submission System

PseudoCAP Coordinator

Confidence Level Classification

Review: Biennial Pseudomonas Conference

PseudoCyc NCBI RefSeq

Updated Annotation Database and Updates Log
Annotation submission review process

- Coordinator examines initial submission
  - Looks for reference and contact info for corresponding author
- Responds with requests for any additional information and clarification, if required.
- Entry reviewed by additional reviewer from the research community, if required.
- Coordinator assigns product name confidence classification

---

Weekly Literature Review

PseudoCAP

Computational Analyses

E-mail or WWW Submission System
1. Modify an existing annotation
2. Correct a sequence
3. Annotate a new gene or sequence (individual annotations or batch submissions)

Authors of Literature

PseudoCAP Coordinator

Review: Biennial Pseudomonas Conference

Confidence Level Classification

Review

Updated Annotation Database and Updates Log
Product confidence level classification system

• Reflects the type of evidence upon which the gene/protein name was based.
• Similar confidence classification used for other data
• TIGR has requested our Class 1 annotations

<table>
<thead>
<tr>
<th>Protein Name Confidence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Function experimentally demonstrated in <em>Pseudomonas aeruginosa</em>.</td>
</tr>
<tr>
<td>2</td>
<td>Function of highly similar gene experimentally demonstrated in another organism (and gene context consistent in terms of pathways its involved in, if known).</td>
</tr>
<tr>
<td>3</td>
<td>Function proposed based on presence of conserved amino acid motif, structural feature or limited sequence similarity to an experimentally studied gene.</td>
</tr>
<tr>
<td>4</td>
<td>Homologs of previously reported genes of unknown function, or no similarity to any previously reported sequences.</td>
</tr>
</tbody>
</table>
Literature review process

- Review literature once per week in order to keep number of submissions from growing too large
- Consult with author of paper to make sure they agree with submission
- Not frequently subject to additional review because it has already been peer-reviewed
Updated annotations are recorded in an updates log

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>Genome</th>
<th>Date / Description</th>
<th>Acknowledgements</th>
</tr>
</thead>
</table>
| PA0175   | *Pseudomonas aeruginosa* | **2006-03-01**
| PA0087   | *Pseudomonas aeruginosa* | **2006-03-01**

- Individual gene pages link to all entries in log for that gene and protein
- Boolean search page allows search by locus ID, participant, update description and date updated
- Sort results and download to text file
- Updates log made available for review at *Pseudomonas* conference
Annotation issues
Annotation issues encountered

- **Unpublished data submitted**
  - Review by at least one additional member of the research community
  - If gene and protein names are offered, recorded as alternate names until published or other equivalent consensus is reached.

- **Addition of new genes** (e.g. recently added rRNA, tRNA genes)
  - Decimal numbering system

  Example, PA4802.1 would be used to indicate the gene PA number for a new gene identified between PA4802 and PA4803. PA4802.01 would be used to identify a new gene between PA4802 and PA4802.1.
Annotation issues encountered

- More than one gene with same name OR one gene with multiple proposed primary names
  - Conflicts should be resolved within reasonable time frame to avoid confusion in the literature
  - All researchers involved in the conflict should be involved in the resolution
  - In absence of researcher consensus, the more predominant name used in the literature and by research groups is favored
  - In the absence of literature and researcher consensus, the first published name in the literature will be given priority
  - Alternate names are still recorded under the “alt gene” name field
  - Biennial *Pseudomonas* conference review
Pseudomonas Genome Database Version 2

New ability to compare annotations within or between species
Search annotations associated with any *Pseudomonas* genome made publicly available

- Text-based annotation search page
  - Boolean search engine
- Sequence-based search page
  - BLAST search of nucleotide or protein sequence database

View results as list of proteins
  (BLAST output including pairwise alignments is still provided with BLAST Search Results)

- Download text file
- Add proteins in list to clipboard utility
- Compare annotations from within or between genomes
- Sort
View and compare annotations within a *Pseudomonas* species or between species.

- Get to this point using text- or sequence-based search
- Sort results, download clipboard annotations to text file, flip genome orientation
View and compare annotations within a *Pseudomonas* species or between species.

- Perform pre-formatted BLAST searches, multiple sequence alignment (ClustalW)
- Go to individual gene pages for links to:
  - GBrowse, PseudoCyc, KEGG, TIGR, updates log
New analyses and updates
New and updated analyses from Brinkman Lab

**PSORTb**...very accurate protein location prediction

- Multi-component (or module) approach to localization prediction

**Modules included with PSORTb:**

- **Signal peptides:** Non-cytoplasmic
  - HMM

- **Amino acid composition/patterns:** All localizations
  - Support Vector Machine’s trained with frequent subsequences

- **Transmembrane helices:** Cytoplasmic membrane
  - HMMTOP

- **PROSITE motifs with 100% precision:** All localizations

- **Outer membrane motifs:** Outer membrane
  - Identified by association-rule mining

- **Homology to proteins of experimentally known localization:** All localizations
  - “SCL-BLAST” against database of proteins of known localization

Integration with a Bayesian Network

- **96% precision**
- **82% recall**
New and updated analyses from Brinkman Lab

PSORTb…very accurate protein location prediction

- Most precise subcellular localization prediction method available
- First computer-based method that exceeded the accuracy of high-throughput laboratory methods (and more than 500x faster)
- New version in development (improved recall)

http://www.psort.org/psortb

Gardy et al. (2005) Bioinformatics 21(5):617-623

New and updated analyses from Brinkman Lab

Improved (more precise) ortholog predictions using Ortholuge

- http://www.pathogenomics.ca/ortholuge
- Fulton et al. (2006) BMC Bioinformatics 7:270
- High throughput method for evaluating ortholog predictions
- Examines phylogenetic distance ratios between two comparison species and an outgroup species
- Identifies predicted orthologs undergoing unusual rates of divergence
- Assigns ortholog predictions as ‘probable orthologs’, ‘uncertain’ or ‘probable paralogs’
- Notable number of orthologs predicted by reciprocal best-BLAST-hit analysis are likely false positives (i.e. are paralogs)
  - Bacterial genome dataset (~5%), eukaryotic genome dataset (~10%)
New and updated analyses

PseudoCyc

- Pathway/Genome database for *Pseudomonas aeruginosa* PAO1 developed by SRI International
- Now maintained by PseudoCAP
- Using Pathway Tools version 9.5

Updates
- Gene, product names updated to include latest annotation submissions
- Genomic DNA sequence and base pair coordinates updated to accommodate nucleotide insertion at position 2669175
- Primary keys in database updated
Acknowledgments

• Fiona SL Brinkman
• Robert EW Hancock
• Raymond Lo, Brinkman Lab
• SRI International – Pathway Tools Support
• Cystic Fibrosis Foundation Therapeutics Inc.
• PseudoCAP participants

• v2.pseudomonas.com
• www.pseudomonas.com
• All web site source code freely available under GNU public license