Manual curation is not sufficient for annotation of genomic databases

William A. Baumgartner Jr.
K. Bretonnel Cohen
Lynne M. Fox
George Acquaah-Mensah
Lawrence Hunter

Center for Computational Pharmacology
University of Colorado School of Medicine
william.baumgartner@uchsc.edu

Originally presented at ISMB 2007, Vienna, Austria
Found/Fixed Graph

Good software development process

Non-terminating software development process

Center for Computational Pharmacology

william.baumgartner@uchsc.edu
Translating to the Biomedical Domain

Software

Found Bug  Fixed Bug

Genomic Resource

Missing Annotation  Added Annotation*

*Simplifying Assumption: “fixed” = >0 annotations
Modeling Annotation of Swiss-Prot and Entrez Gene

- Drosophila proteins missing a GO annotation vs. Drosophila proteins gaining at least one GO annotation
- Mouse proteins missing a GO annotation vs. Mouse proteins gaining at least one GO annotation
- Proteins missing a FUNCTION comment vs. Proteins gaining a FUNCTION comment
- Human Entrez Gene records missing a GeneRIF vs. Human Entrez Gene records gaining at least one GeneRIF
Predicting completion of GO annotation for mouse proteins in Swiss-Prot
Predicting completion of annotation

How long will it take?

<table>
<thead>
<tr>
<th>Data type</th>
<th>linear</th>
<th>$R^2$</th>
<th>exponential</th>
<th>$R^2$</th>
<th>logarithmic</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swiss-Prot Drosophila GO annotations</td>
<td>1.16</td>
<td>0.9570</td>
<td>0.55</td>
<td>0.9506</td>
<td>1.38</td>
<td>0.9572</td>
</tr>
<tr>
<td>Swiss-Prot Mouse GO annotations</td>
<td>3.06</td>
<td>0.8778</td>
<td>0.90</td>
<td>0.8436</td>
<td>3.75</td>
<td>0.8845</td>
</tr>
<tr>
<td>Swiss-Prot all species GO annotations</td>
<td>10.5</td>
<td>0.5746</td>
<td>3.05</td>
<td>0.7852</td>
<td>16.68</td>
<td>0.5530</td>
</tr>
<tr>
<td>Swiss-Prot all species function annotations</td>
<td>99.0</td>
<td>0.9807</td>
<td>9.12</td>
<td>0.8870</td>
<td>1.07 x 10^9</td>
<td>0.8207</td>
</tr>
<tr>
<td>Entrez Gene Human GeneRIFs</td>
<td>13.0</td>
<td>0.9788</td>
<td>0.003</td>
<td>0.7132</td>
<td>24.83</td>
<td>0.9784</td>
</tr>
<tr>
<td>Entrez Gene Mouse GeneRIFs</td>
<td>38.3</td>
<td>0.9777</td>
<td>0.40</td>
<td>0.7227</td>
<td>629,396</td>
<td>0.9221</td>
</tr>
</tbody>
</table>
Acknowledgements

• Thanks to the anonymous ISMB reviewers for their helpful comments.

• Thanks to my colleagues in the Hunter Lab, especially Dr. Sonia Leach, Dr. Michael Bada, and Dr. Hannah Tipney.

• This work funded by NIH grant R01-LM008111 to Lawrence Hunter.