BMC Bioinformatics

Meeting abstract

Comparison of annotation terms between automated and curated *E. coli* K12 databases

ReddySailaja Marpuri* and Claire A Rinehart

Address: Department of Biology, Western Kentucky University, Bowling Green, KY 42101, USA

Email: ReddySailaja Marpuri* - reddysailaja.marpuri418@wku.edu

* Corresponding author

from UT-ORNL-KBRIN Bioinformatics Summit 2009
Pikeville, TN, USA. 20–22 March 2009

Published: 25 June 2009

BMC Bioinformatics 2009, **10**(Suppl 7):A10 doi:10.1186/1471-2105-10-S7-A10

This abstract is available from: http://www.biomedcentral.com/1471-2105/10/S7/A10

Background

Genome sequencing and annotation may provide ways to understand genomes. Annotation of genome results in identification of genes in terms of precise start and end sites and description of cellular components, molecular functions and biological process. Increase in the wealth of the genomic data has led to the necessity of identification of information encoded within the genome which in turn resulted in the development of automated annotation techniques that assigns functions to newly sequenced genes based on similarity to previously annotated genes. This approach has a few problems, for example if there was a mistake or error in previously annotated genomes it will result in whole family of misannotated genes. Annotation usually fails to meet the "golden standard" of the curated databases as the level of details in automated annotation systems is reduced, classifying proteins into more broader categories. To overcome this problem;

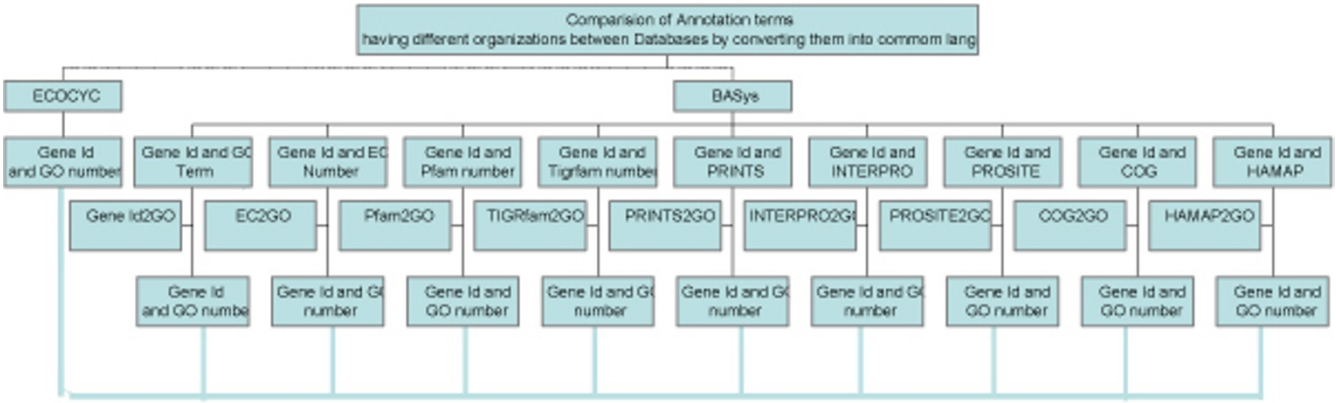

Figure 1

Flow chart for processing of annotation terms. EcoCyc was used as the standard for the comparison of annotation terms in the form of Gene Id's and GO numbers. In order to convert BASys terms from different datasets (row: 3, lanes: 2–9) into gene ID's and GO numbers (row 5: lanes 2–9); conversion files (row: 4; lanes 2–9) from gene ontology site were used. Each of the BASys Gene Id's and numbers (row 5: lanes 2–9); were compared to EcoCyc Gene Id's and GO number (row 3: lane-1).

Table 1: Summary of matches and mismatches between databases

Databases	True Positives	False Positives	True Negatives	False Negatives	Correlation Coefficient
TIGRfam	373	1832	2340	3277	- 0.37
HAMAP	192	1629	2592	3458	- 0.39
PRINTS	88	1723	2524	3562	- 0.45
EC: Numbers	288	3369	2382	3362	- 0.51
PROSITE	161	2226	2076	3489	- 0.51
PFAM	305	2729	1329	3345	- 0.60
INTERPRO	502	3758	1095	3148	- 0.63
GO:TERMS	349	7356	1344	3301	- 0.71
COG	71	2095	972	3579	- 0.71
Composite	838	15219	2839	2812	- 0.52

EcoCyc data compared to BASys Database sources was listed in column: 1. Column: 2 to 4 Show the number of matching annotation terms between the databases that were true positives, false positives, true negatives and false negatives. Column: 6 show the correlation coefficient for each of the databases. The composite is the sum of column data.

ontology terms were used in automated databases as a means of understanding and recognizing types of proteins to the level of curated databases.

In this project we tried to compare the results of predictive automated bacterial annotation programs to a curated annotation databases such as EcoCyc. EcoCyc is a conservative multidimensional annotation system that is validated by over 15,000 publications. Automated annotation systems, such as BASys can be used as first pass annotation tools that try to add as many annotations as possible by drawing upon over 30 sources. Gene Ontology is described by a defined library of terms related to the biological process, cellular components and molecular functions of a gene in an organism. Because of the limited and common terms in the ontology annotations, we compared ontology's between the BASys and EcoCyc databases. Additional, non-ontology terms and metadata were generated in BASys. Methods were developed to compare these additional terms to the EcoCyc database and it was found that approximately 17% of the BASys predicted ontology's matched the EcoCyc database.

Materials and methods

Gene Ontology database [5] was used to convert each of the annotation terms into corresponding GO numbers as shown in Figure 1 using annotation term-2-GO files. BASys and EcoCyc were the databases used for comparison (3 and 4).

Each of the annotation terms from the respective databases were converted into common GO numbers by using the respective conversion files from the Gene Ontology site http://www.geneontology.org/.

Results and conclusion

Our results showed that of the appoximately 4200 genes in *E. coli,* 1594 of them have been validated by EcoCyc based on Ontology numbers. EcoCyc is a conservative annotation system that requires strict validation before entering annotation terms into its database. On the other hand BASys was found to be more liberal in assigning annotations to 2511 genes based on ontologies, because it was designed to annotate each gene as fully as possible. Total GO numbers based shown in Table 1 was found to be 21,708. Table 1 shows that about 17% (4% true positives and 13% true negatives) of BASys ontology assignments were validated with EcoCyc. About 70% of them were false positives and 13% of them were false negatives. The high false positive rate might be due to incomplete literature validation of EcoCyc database.

Acknowledgements

Bioinformatics and Information Science Center, Western Kentucky University.

References

1. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, Peralta-Gil M, *et al.*: **Multidimensional annotation of the *Escherichia coli* K-12 genome.** *Nucleic Acids Res* 2007, **35(22):**7577-7590.
2. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS: **BASys: a web server for automated bacterial genome annotation.** *Nucleic Acids Res* 2005:W455-W459.
3. **ECOCYC** [http://www.ecocyc.org/]
4. **BASys Bacterial Annotation System** [http://wishart.biology.ualberta.ca/basys/cgi/gallery.pl]
5. **Gene Ontology** [http://www.geneontology.org/]