Restauro-G: A Rapid Genome Re-Annotation System for Comparative Genomics

Satoshi Tamaki[1,2], Kazuharu Arakawa[1]*, Nobuaki Kono[1,2], and Masaru Tomita[1,2]

[1] *Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan;* [2] *Department of Environmental Information, Keio University, Fujisawa 252-8520, Japan.*

Annotations of complete genome sequences submitted directly from sequencing projects are diverse in terms of annotation strategies and update frequencies. These inconsistencies make comparative studies difficult. To allow rapid data preparation of a large number of complete genomes, automation and speed are important for genome re-annotation. Here we introduce an open-source rapid genome re-annotation software system, Restauro-G, specialized for bacterial genomes. Restauro-G re-annotates a genome by similarity searches utilizing the BLAST-Like Alignment Tool, referring to protein databases such as UniProt KB, NCBI nr, NCBI COGs, Pfam, and PSORTb. Re-annotation by Restauro-G achieved over 98% accuracy for most bacterial chromosomes in comparison with the original manually curated annotation of EMBL releases. Restauro-G was developed in the generic bioinformatics workbench G-language Genome Analysis Environment and is distributed at http://restauro-g.iab.keio.ac.jp/ under the GNU General Public License.

Key words: bioinformatics, software, annotation, G-language Genome Analysis Environment, complete genomes

Introduction

The advent of genome sequencing technologies has greatly reduced both the time and cost required for identifying complete nucleotide sequences; consequently, the number of complete genomes is growing at an increasingly rapid rate. The Genomes OnLine Database (GOLD) currently lists more than 2,000 published and ongoing genome projects (*1*), and this number is continuously increasing. In addition to sequence information, high-quality genome annotation is indispensable for understanding a genome, its components, and the protein products. Sequences submitted to the International Nucleotide Sequence Database Collaboration (INSDC) through GenBank (*2*), the European Molecular Biology Laboratory (EMBL) (*3*), and the DNA Data Bank of Japan (DDBJ) (*4*) repositories are accompanied with functional descriptions and links to external resources regarding the genetic components, in a way that is useful for bioinformatics and genomics research. However, those submitted complete genome sequences are annotated by each sequencing project using their own methods and criteria (*5*), and annotation updates are also maintained by the submitters (*6*). This leads to a diversity in the annotation completeness, and some genomes at early annotation stages have limited or sometimes no functional information (*7*). Moreover, because gene functional annotation relies heavily on similarity searching techniques with protein sequence databases, automatically annotated entries can become quickly outdated when the reference sequence used for the similarity-based search is updated (*8–10*). This is a central problem of bioinformatics, especially for comparative analyses, which require genome annotation with uniform criteria and computer-friendly semantics.

The Genome Reviews database at the European Bioinformatics Institute (EBI), and complete genome sequences having accession numbers prefixed with "NC_" in RefSeq database (*11*) of the National Center for Biotechnology Information (NCBI), help to solve this problem by automatically re-annotating and regularly updating the annotation of complete genomes and by using manual curation under standardized criteria. For genomes that are not finished or only available in-house, several software systems achiev-

***Corresponding author.**
E-mail: gaou@sfc.keio.ac.jp

ing high efficiency and accuracy have been developed for automatic and semi-automatic annotation, such as GeneQuiz (*12*), MAGPIE (*13*), PEDANT (*14*), Ensembl (*15*), and GenDB (*16*). The majority of these tools, however, are aimed at genome projects; thus they are semi-automatic and premised on having final expert curation, being equipped with rich user interfaces for this purpose. Moreover, these software systems take quite a long time for finishing the full annotation process, on the order of hours to days. While this approach is necessary to ascertain the high quality demanded for genome projects, data preparation for comparative studies has different requirements, such as re-annotation with controlled methods and with sufficient speed. For example, even if an annotation system is quick enough to finish annotating one genome in 2 h, complete re-annotation of the currently available 375 microbes would still require more than 1 month.

To avoid erroneous conclusions possibly affected by the accuracy and diversity of genome annotation, computational analyses of genome sequences often require a careful selection of methods and datasets. We previously reported on a benchmarking method, Gene Prediction Accuracy Classification (GPAC) (*17*), to quantify the sensitivity of computational analysis for this purpose. However, this kind of pre-analysis and data selection should ideally be coupled with rapid re-annotation software with a flexible configuration for various annotation strategies. In light of these requirements, here we introduce a rapid open-source automatic genome re-annotation system, Restauro-G, developed by using the generic bioinformatics workbench G-language Genome Analysis Environment (G-language GAE) (*18*, *19*). Restauro-G achieves high accuracy in comparison with manually curated complete genomes in the EMBL repository, and the system generates annotation in computer-friendly semantics with rich links to external resources in a variety of formats.

System and Methods

Strategy overview

Re-annotation by Restauro-G is based on similarity searches of amino acid sequences against public databases of protein sequences. Because gene identification is well established for bacterial genomes and can achieve high sensitivity and specificity (*20*), the system uses the predicted coding regions annotated in the original genome and focuses on the functional annotation of the protein products wherever possible. If only the nucleotide sequence is available, the system can alternatively identify the coding regions by using the GLIMMER software (*21*).

For similarity searches, three databases are used to account for the information reliability and coverage. The manually curated Swiss-Prot database of the UniProt Knowledgebase (UniProt KB) (*22*) is used as the first priority; the computer-annotated TrEMBL is used as the second priority; and finally the NCBI non-redundant (nr) database (*11*) is used for maximum coverage. Reliability levels are assigned to the matches of similarity searches, and information on protein products is obtained from the corresponding entries in UniProt KB. The genomes can be further annotated with three additional types of information: (1) orthologous clusters with amino acid sequence similarity searches against the NCBI Clusters of Orthologous Groups (COGs) database (*23*); (2) protein domains using HMMER (*24*) and HMMPfam (*25*); and (3) protein localization from PSORTb (*26*). Annotated genomes can be generated in numerous formats supported by G-language GAE and Bioperl (*27*), including GenBank, EMBL, and GFF. Because the system was developed in G-language GAE, users can easily modify and adjust the resulting formats according to their needs. Restauro-G adds the new annotation to the genome flatfile without replacing the existing entries.

Implementation

The software performs the following processes for re-annotation. First, upon user selection of the input genome in GenBank or EMBL format, the system performs similarity searches of all genes using the BLAST-Like Alignment Tool (BLAT) (*28*). Credibility of the BLAT search is marked with the five reliability levels outlined in Table 1:

Table 1 The five reliability levels for the BLAT search

Level	E-value	(Match/Subject length) and (Match/Query length)
Level 1	$\leq$ 1E-70	$\geq$ 98%
Level 2	$\leq$ 1E-50	$\geq$ 95%
Level 3	$\leq$ 1E-30	$\geq$ 90%
Level 4	$\leq$ 1E-10	$\geq$ 80%
Level 5	None of the above	

In the table, "Match" denotes "Alignment-Length" minus "Mismatches" from the BLAT output in blast8 format, similar to the levels as defined by the GAMBLER software (*29*). The target databases for the BLAT search are Swiss-Prot, TrEMBL, and NCBI nr, in the order of priority based on the information reliability and coverage.

Based on the level assignment procedures, Restauro-G searches for homologues until the top five hits are recorded. To reduce computational cost, the system searches in the database of lower priority only when it did not find matches in a certain level, and the hits in the same reliability level are sorted by sequence identity. By default, reliability level is defined as follows:

Swiss-Prot Level 1 > Swiss-Prot Level 2 > TrEMBL Level 1 > TrEMBL Level 2 > Swiss-Prot Level 3 > TrEMBL Level 3 > Swiss-Prot Level 4 > TrEMBL Level 4 > NCBI nr Levels 1–4 > Swiss-Prot Level 5 > TrEMBL Level 5 > NCBI nr Level 5.

In this way, users can select the genes used for the comparative study according to the annotation credibility, removing those erroneous annotations that are likely for a certain percentage of genes annotated with this kind of automatic method. After the BLAT search, Restauro-G refers to the UniProt KB for annotation. Annotations done by the system depend on the target database (Table 2). In all cases, ID, gene name, gene description, similarity E-value, and reliability levels are annotated. For the similarity search with UniProt KB/Swiss-Prot and UniProt KB/TrEMBL, database cross-reference, comments, and feature table are also annotated. In the orthologous search, COG family is retrieved from NCBI COGs. In addition, domain information is annotated from HMMPfam, and protein location information is annotated from PSORTb. All hits to the employed databases are recorded with database IDs and information in text for users, as well as with semantic ontological identifiers including the Gene Ontology (GO) terms (*30*).

Performance optimization

BLAT is chosen for similarity searches due to its rapidity, and "minScore" option is set to 100 for speed but with sufficient accuracy based on the levels as defined above. In our server (Dual Pentium 4 Xeon 2.8 GHz, 4 GB RAM), BLAT was more than 130 times faster than BLAST (*31*), with equivalent accuracy in terms of the above reliability levels. Hierarchy of the database is defined not only to establish high accuracy, but also to reduce computational resources, because the size of the database increases when the rank order of the database decreases to account for higher coverage. To cope with the massive databases that the system has to handle, databases are parsed and stripped down to only contain the information necessary for Restauro-G annotation, and the data are further converted to be stored in virtual memory and heavily indexed to achieve high performance. All of the database access is performed in memory, both physically and virtually; therefore, the performance is greater even than relational databases.

Results and Validations

Restauro-G is implemented with G-language GAE, packaged for UNIX platforms, and distributed at http://restauro-g.iab.keio.ac.jp/ under the GNU General Public License. The web site contains documentations about the software as well as the 375 re-annotated bacterial genomes.

Re-annotated genomes were validated for accuracy by comparing the annotated external database reference to UniProt KB entries with those in the original bacterial genomes released in the EMBL repository. Data used for validation were from release 8.5 of UniProt KB, 50.5 of Swiss-Prot, 33.5 of TrEMBL, and 2006-04-04 of NCBI nr. Here we show the results of five example genomes, *Bacillus subtilis* (*32*), *Escherichia coli* K12 (*33*), *Mycoplasma genitalium* (*34*),

Table 2 Types of annotations and information included in Restauro-G annotation

Type	Database	Annotataion
Similarity	UniProt KB/Swiss-Prot UniProt KB/TrEMBL	ID/gene name/description/database cross-reference/E-value/level/comments/feature table
	NCBI nr	ID/gene name/description/E-value/level
Orthologous	NCBI COGs	ID/gene name/description/COG family/E-value/level
Domain	HMMPfam	ID/gene name/description/domain information/E-value
Protein location	PSORTb	Protein location information

Mycobacterium tuberculosis (*35*), and *Pyrococcus furiosus* (*36*) (Table 3). Annotations were also compared for corresponding genomes in release 61 of the EBI Genome Reviews.

The overall results are displayed in Table 3. Both for the comparison with EMBL and EBI Genome Reviews, all genomes achieved over 98% accuracy, among which *M. genitalium* scored the highest with a perfect match. The annotation time ranged from 5 to 45 min. For each of the genomes, missed entries are shown with respective reasons in Tables S1–S4. Genes in EMBL without external reference to UniProt KB are listed in Tables S5–S7.

The majority of the missed entries were caused by the change in sequence information or the entry ID. Of the 57 missed entries of *E. coli*, 50 were due to changes in the start codon position, 1 for methionine excision, 4 for deletion of the entry in UniProt KB, 1 for the priority of the database, and 1 was not annotated by Restauro-G. Similarly, 51 of the 59 misannotated genes in *M. tuberculosis*, 6 of 6 in *B. subtilis*, and 5 of 15 in *P. furiosus* were due to updates in UniProt KB. Genes in EMBL without external reference to UniProt KB were mostly fragmented coding regions, pseudogenes, or prophages. Although the number was minimal, several entries in the genomes of *M. tuberculosis* and *P. furiosus*, which are not as well studied as *E. coli* and *B. subtilis* and thus have most of their entries in TrEMBL instead of Swiss-Prot, were missed due to the priority of the databases.

Discussion

Rapid and automated re-annotation is essential to cope with the large amount of genomic information for comparative bioinformatics studies. Re-annotation by Restauro-G is sufficiently accurate, achieving over 98% accuracy compared with EMBL and Genome Reviews. The system is also rapid, finishing one microbial genome within 5 to 45 min, depending on the size of the genome and the number of corresponding entries in Swiss-Prot. Most of the missed entries are due to the update of UniProt KB. Because the EMBL releases used in the validation were annotated with UniProt KB release 7, whereas release 8.5 was used in this work, the use of the latest predicted coding regions for EMBL should allow reclamation of most of the missed entries. In addition, because UniProt has a very rapid biweekly release cycle, automatically annotated entries in EMBL can become outdated quickly. A major fraction of the "missed" entries in the validation described in this work may be regarded as the identification of outdated entries in EMBL, rather than being annotated mistakenly. Therefore, Restauro-G may be used for the estimation of updated entries, taking advantage of its rapidity. Similar results were also obtained for the comparison with Genome Reviews. There were also genes that did not have a correct match due to the system architecture in the level clustering system, because organisms that are not well studied tend to have their genes included in TrEMBL instead of Swiss-Prot. Adjusting the database priority beforehand should correct this problem. Moreover, because these problems only accounted for about 0.2%–0.5% of the genes in a whole genome, the system should be sufficient for use in comparative studies. The number of genome sequences is continuing to grow at a rapid rate, and the emerging field of microbial comparative genomics through metagenomics, a shotgun sequencing of entire genetic material from environmental samples (*37*), greatly increases the total number of genes that need to be functionally annotated. The number of available genome sequences will likely continue to grow at an exponential rate in the foreseeable future, and

Table 3 Validation of Restauro-G annotation accuracy

Genome*	No. of coding sequences	Annotation in EMBL	Restauro-G prediction	Matches with EMBL	Matches with Genome Reviews (%)	Time (s)
B. subtilis	4,106	4,106	4,105	4,100 (99.85%)	99.70%	1,110
E. coli	4,331	4,259	4,301	4,202 (98.66%)	99.46%	510
M. genitalium	476	476	476	476 (100.00%)	100.00%	212
M. tuberculosis	4,189	4,186	4,188	4,127 (98.59%)	98.50%	2,430
P. furiosus	2,065	2,057	2,062	2,043 (99.31%)	99.27%	1,195

*Genome versions: *B. subtilis*—EMBL: AL009126 07-JUL-2003 (rel. 76, ver. 3); *E. coli*—EMBL: U00096 13-AUG-2006 (rel. 88, ver. 6); *M. genitalium*—EMBL: L43967 14-JAN-2006 (rel. 86, ver. 2); *M. tuberculosis*—EMBL: AE000516 14-APR-2005 (rel. 83, ver. 2); *P. furiosus*—EMBL: AE009950 22-JAN-2004 (rel. 78, ver. 2).

therefore rapid automated annotation methods will be indispensable for data preparation prior to comparative analyses. Coupled with GPAC of G-language GAE, Restauro-G will be a useful tool for this purpose, achieving high accuracy while reducing the time required from the order of hours to minutes.

Acknowledgements

We thank the members of MGSP at the Institute for Advanced Biosciences, Keio University, for critical suggestions. This work was supported by the Japan Society for the Promotion of Science (JSPS).

Authors' contributions

ST developed and validated the software, and drafted the manuscript. KA conceived the software, designed the fundamental architecture, and edited the manuscript. NK developed the web database and edited the manuscript. MT supervised the research and revised the final manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Liolios, K., *et al.* 2006. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* 34: D332-334.
2. Benson, D.A., *et al.* 2007. GenBank. *Nucleic Acids Res.* 35: D21-25.
3. Kulikova, T., *et al.* 2007. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.* 35: D16-20.
4. Sugawara, H., *et al.* 2007. DDBJ working on evaluation and classification of bacterial genes in INSDC. *Nucleic Acids Res.* 35: D13-15.
5. Iliopoulos, I., *et al.* 2003. Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* 19: 717-726.
6. Sterk, P., *et al.* 2006. Genome Reviews: standardizing content and representation of information about complete genomes. *Omics* 10: 114-118.
7. Ouzounis, C.A. and Karp, P.D. 2002. The past, present and future of genome-wide re-annotation. *Genome Biol.* 3: COMMENT2001.
8. Serres, M.H., *et al.* 2001. A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* 2: RESEARCH0035.
9. Iliopoulos, I., *et al.* 2001. Genome sequences and great expectations. *Genome Biol.* 2: INTERACTIONS0001.
10. Camus, J.C., *et al.* 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 148: 2967-2973.
11. Pruitt, K.D., *et al.* 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35: D61-65.
12. Andrade, M.A., *et al.* 1999. Automated genome sequence analysis and annotation. *Bioinformatics* 15: 391-412.
13. Gaasterland, T. and Sensen, C.W. 1996. MAGPIE: automated genome interpretation. *Trends Genet.* 12: 76-78.
14. Riley, M.L., *et al.* 2007. PEDANT genome database: 10 years online. *Nucleic Acids Res.* 35: D354-357.
15. Hubbard, T.J., *et al.* 2007. Ensembl 2007. *Nucleic Acids Res.* 35: D610-617.
16. Meyer, F., *et al.* 2003. GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* 31: 2187-2195.
17. Arakawa, K., *et al.* 2006. GPAC: benchmarking the sensitivity of genome informatics analysis to genome annotation completeness. *In Silico Biol.* 6: 49-60.
18. Arakawa, K. and Tomita, M. 2006. G-language System as a platform for large-scale analysis of high-throughput omics data. *J. Pesticide Sci.* 31: 282-288.
19. Arakawa, K., *et al.* 2003. G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* 19: 305-306.
20. Audic, S. and Claverie, J.M. 1998. Self-identification of protein-coding regions in microbial genomes. *Proc. Natl. Acad. Sci. USA* 95: 10026-10031.
21. Delcher, A.L., *et al.* 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27: 4636-4641.
22. The UniProt Consortium. 2007. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 35: D193-197.
23. Tatusov, R.L., *et al.* 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.
24. Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755-763.
25. Bateman, A., *et al.* 2004. The Pfam protein families database. *Nucleic Acids Res.* 32: D138-141.
26. Gardy, J.L., *et al.* 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 21: 617-623.

27. Stajich, J.E., *et al.* 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12: 1611-1618.
28. Kent, W.J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12: 656-664.
29. Sakiyama, T., *et al.* 2000. An automated system for genome analysis to support microbial whole-genome shotgun sequencing. *Biosci. Biotechnol. Biochem.* 64: 670-673.
30. Ashburner, M., *et al.* 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25: 25-29.
31. Altschul, S.F., *et al.* 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
32. Kunst, F., *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249-256.
33. Blattner, F.R., *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453-1474.
34. Fraser, C.M., *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.
35. Fleischmann, R.D., *et al.* 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* 184: 5479-5490.
36. Maeder, D.L., *et al.* 1999. Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. *Genetics* 152: 1299-1305.
37. Venter, J.C., *et al.* 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66-74.

Supporting Online Material

http://restauro-g.iab.keio.ac.jp/
Tables S1–S7