A Web-based High-Throughput Tool for Next-Generation Sequence Annotation

Kamal Kumar, Valmik Desai, Li Cheng, Maxim Khitrov, Deepak Grover, Ravi Vijaya Satya, Chenggang Yu, Nela Zavaljevski, and Jaques Reifman
Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Fort Detrick, MD
{kamal, valmik, lcheng, mkhitrov, dgrover, rvijaya, cyu, nelaz, reifman}@bioanalysis.org

Abstract

The availability of a large number of genome sequences, resulting from inexpensive, high-throughput next-generation sequencing platforms, has created the need for an integrated, fully-automated, rapid, and high-throughput annotation capability that is also easy-to-use. Here, we present a web-based software application, Annotation of Genome Sequences (AGeS), which incorporates publicly-available and in-house-developed bioinformatics tools and databases, many of which are parallelized for high-throughput performance. The current version of AGeS provides annotations for bacterial genome sequences, and serves as a readily-accessible resource to Department of Defense (DoD) scientists for storing, annotating, and visualizing genomes of newly-sequenced pathogens of interest.

The AGeS system is composed of two major components. The first component is a web-based application that provides a graphical user interface for managing users' input genomes, submitting annotation jobs, and visualizing results. Sequence contigs are uploaded as a multi-FASTA input file and submitted for annotation, and the resulting annotations are visualized through GBrowse. The input genome sequences and the annotation results are stored in a secure, customized database. The second component is a high-throughput annotation pipeline for finding the genomic regions that code for proteins, RNAs, and other genomic elements through a Do-It-Yourself Annotation framework. The pipeline also functionally annotates the protein-coding regions using an in-house-developed high-throughput pipeline, the Pipeline for Protein Annotation. The annotation pipeline has been deployed on the Mana Linux cluster at the Maui High Performance Computing Center. The two components are connected together using the DoD user interface toolkit application programming interface.

The AGeS system was evaluated for scaling of its parallel execution and annotation performance. AGeS scaled with super-linear speedup for up to 128 processors, after which performance degraded. A 2.2-Mbp bacterial genome sequence can be annotated in ~1 hr using 128 processors. AGeS annotations of draft and complete genomes were compared with the original annotations from three different sources, and were found to be in general agreement with them.

1. Introduction

Access to inexpensive, high-throughput DNA sequencing technology has led to an explosion in the number of sequenced organisms and the volume of sequenced data[1]. To date, due to the so called "next-generation sequencing" technology, the genomes of >1,000 microbial pathogens and their near neighbors are available, and many more are being sequenced. A genome sequence provides valuable information in terms of genomic features, such as genes that code for proteins and RNAs, as well as the positions and numbers of tandem repeats. In addition, we can gain further insights by annotating the functions of the proteins that the genes code for. This valuable information, gleaned from the annotation of a newly sequenced complete genome, can help devise new strategies in diagnostics and forensics. Moreover, these annotations, coupled with comparative genomics, can enable novel approaches to identify vaccine candidates and potentially discover "universal" drug targets. For such downstream applications, the annotation of genomic sequences needs to be integrated, fully-automated, rapid, and high-throughput; and for such annotation capability to be truly effective, it should also be easy-to-use and readily available.

To address this need, we developed the Annotation of Genome Sequences (AGeS) software system, which was designed as a modular and flexible platform to facilitate the annotation, storage, and comparative analysis of sequenced genomes[2]. The AGeS system is composed of a Web-based application and a software pipeline. The Web-based application enables users to upload and store input contig sequences and the resulting annotation data in a central, customized database and users

can visualize the annotations via easy-to-use graphical user interfaces (GUIs). The visualization of annotated sequences is presented using the open-source genome browser GBrowse[3]. The integrated software pipeline analyzes contig sequences, and locates genomic regions that code for proteins, RNAs, and other genomic elements through a Do-It-Yourself Annotation (DIYA) framework[4] and Tandem Repeats Finder (TRF)[5]. The identified protein-coding regions are then functionally annotated using an in-house-developed high-throughput pipeline, the Pipeline for Protein Annotation (PIPA)[6]. All of these capabilities are available for bacterial genomes. Overall, AGeS provides the functionalities to: 1) store input sequences and annotated sequence data, 2) annotate completed and draft bacterial genomes in a fully-integrated and automated manner, 3) use high performance computing (HPC) for high-throughput annotation through efficient parallelization of the various publicly-available and in-house-developed bioinformatics resources, 4) visualize annotations using the familiar GBrowse[3] interface, and 5) download annotated genomes in GenBank[7] format.

Several software systems have recently been developed for high-quality, automated annotation of bacterial genomes. These include BASys[8], RAST[9], and Microbial Genomes Database Web resources[10] as well as annotation services provided by some of the large genomic annotation centers, such as the Annotation Engine at the J. Craig Venter Institute (http://www.jcvi.org/cms/research/projects/annotation-service/), the Genoscope's annotation service MicroScope[11], and the Microbial Annotation Pipeline at Integrated Microbial Genomes[12]. However, these systems or services do not provide integrated, fully-automated, rapid, high-throughput, and readily-available capability, and some of the important features, such as mapping to standard Gene Ontology (GO) annotation[13], are also missing. Although most annotation systems contain components that are based on publicly-available bioinformatics programs and databases, integration of these components into pipelines is not a trivial task for researchers without significant bioinformatics and computer science expertise. While recently published DIYA[4] and the Genome Reverse Compiler[14] provide integrated software packages for genome annotation, they do not enable the full use of parallel computing and lack fully-integrated and automated visualization of annotations.

2. Methods and Implementation

The AGeS system is composed of two main components: a Web-based application that provides user-friendly GUIs accessible via a standard Web browser; and a high-throughput software pipeline for the annotation of input genome sequences. Figure 1 shows the overall system architecture of the AGeS system. The AGeS Web application has been designed to control all aspects of the annotation process, i.e., input sequence management for uploading and manipulating genomic sequences, submitting annotation jobs to the AGeS annotation pipeline at an HPC cluster, storing input sequences as well as annotation results into a central relational database management system (RDBMS), and visualizing the annotations in the integrated GBrowse genome browser. For uploading the genome sequences, along with the required genus, species, and strain information, users have the option to upload the data pertinent to the minimum information about a genomic sequence (MIGS)[15]. Internally, the AGeS Web application uses a workflow manager module to guide the entire lifecycle of the annotation process, starting from the upload of an input sequence and ending with the visualization of the annotated sequences.

2.1 Web Application

The AGeS system is accessible at https://applications.bioanalysis.org/ages/, and is available to the Department of Defense (DoD) Supercomputing Resource Centers (DSRCs) users for genome sequence annotation using a standard Web browser. The AGeS Web application has been designed as a modular application for the easy integration of future sequence analysis modules, as they become available, and uses a workflow manager to invoke its modules. Resource-intensive annotation tools are run on the Mana Linux cluster at the Maui HPC Center (MHPCC), which is accessed by the Web application using the DoD User Interface Toolkit (UIT) application programming interface (API) (https://www.uit.hpc.mil/). UIT is a Web service-based API that provides secure access to DoD HPC resources. AGeS users are authenticated through the UIT API using their Kerberos credentials. The AGeS Web application provides GUIs for managing sequences, submitting annotation jobs to the HPC cluster, and visualizing and downloading the annotation results. Figure 2 shows a screenshot of the AGeS Web application, showing the sequence management GUI. When an annotation job is completed on the HPC end, the results are automatically transferred back to the Web server and stored into the central database for visualization and download. Upon completion of an annotation job, an e-mail is also sent automatically to the user.

The AGeS Web application was developed using standards-based technologies, which include Java (http://www.oracle.com/technetwork/java/), J2EE (http://www.oracle.com/technetwork/java/javaee/overview/), JavaServer Faces (JSF) (http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html), asynchronous JavaScript and XML (AJAX)[16],

ICEfaces (http://www.icefaces.org/), jBPM (http://www.jboss.org/jbpm/), and Apache ActiveMQ (http://activemq.apache.org/). The Web application mainly consists of server-side Java codes that use JSF- and AJAX-based APIs from ICEfaces. ICEfaces provides a rich set of user interface components, such as menus, buttons, etc., and generates updated views of Webpages without reloading the entire page. The workflow manager module has been implemented, within the Web application, using the jBPM workflow engine API for controlling the execution of various modules. The Web application uses an Apache ActiveMQ server for asynchronous message passing between the modules and the workflow engine. A PostgreSQL (http://www.postgresql.org/) RDBMS server is used to store users' input genome sequences, annotation results, and other job-related data. The Web application is deployed on an Apache Tomcat (http://tomcat.apache.org/) server, using a secure hypertext transfer protocol over a secure socket layer connection for encrypting all of the data flowing to and from the user's Web browser.

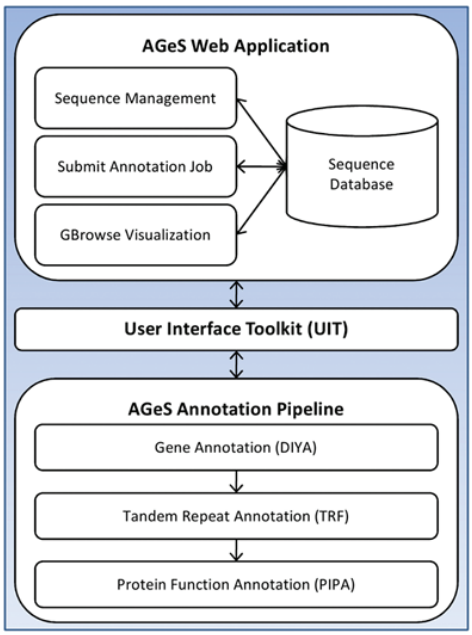

Figure 1. Overall system architecture for the Annotation of Genome Sequences (AGeS) system

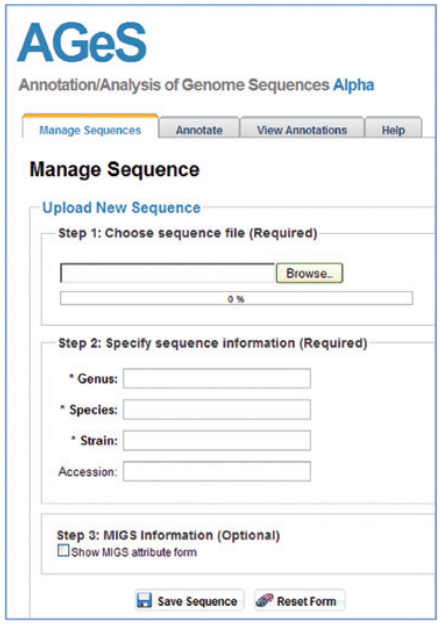

Figure 2. The AGeS graphical user interface used for sequence data management

2.2 Annotation Pipeline

As shown in Figure 1, the AGeS annotation pipeline is composed of three modules for gene, tandem repeats, and protein function annotations. The annotation pipeline takes assembled contiguous sequences, or contigs, as input in multi-FASTA format files generated by high-throughput, next-generation sequencing technologies (http://www.454.com/, http://www.illumina.com/, and http://www.appliedbiosystems.com/). First, a customized DIYA[4] framework is used to locate protein-coding genes using Glimmer[17] and RNA genes using RNAmmer[18] and tRNAscan-SE[19]. Within the DIYA framework, the system uses BLAST[20] searches to extract coding regions from the Glimmer predictions, and to infer gene products by transferring annotation from the best BLAST match. Next, the system finds tandem repeats in the pseudo-assembled sequence using TRF[5]. Outputs from the different DIYA component programs and TRF are post-processed and parsed to generate a file in GenBank format.

After annotation of the genomic regions is complete, the identified protein-coding regions are annotated using the high-throughput protein function annotation methods implemented in PIPA[6]. One of the most useful features of PIPA is that it exploits and consistently consolidates protein function information from disparate sources, including the in-house-developed CatFam enzyme profile database[21]. As an added benefit, the consolidated function predictions are given in GO terms, which is the de facto standard for protein annotation. The protein annotation results from PIPA are included in the GenBank file from the previous step, and are transferred back to the AGeS Web application for storage into the central database.

3. Results

AGeS provides the capability to annotate whole bacterial genomes, including both genomic features and protein functions. The annotation pipeline that has been deployed on the Mana Linux cluster at the MHPCC scales well and is suited for whole genome sequence annotation. In this section, we present the results of the parallel processing performance testing of AGeS as well as of the software validation experiments.

3.1 Parallel Performance

To assess the scalability of the parallelization of the annotation modules of the AGeS pipeline, we computed the speedup curve for the annotation of a typical bacterial genome (Figure 3). Speedup is defined as the ratio of the time taken by a program to run on *N* processors to the time taken to run the same program on a single processor, with an ideal speedup being linear, meaning that the speedup is directly proportional to the number of processors. AGeS achieves super-linear speedup for up to 128 processors, after which its performance declines. The super-linear speedup is attributed to faster processing achieved by fully using the processors' local memory, and the speedup decline beyond 128 processors is attributed to communication overhead. A 2.2-Mbp bacterial genome sequence (e.g., *Staphylococcus hominis* SK119, which is an opportunistic pathogen in patients with a compromised immune system) can be annotated in ~1 hr using 128 processors.

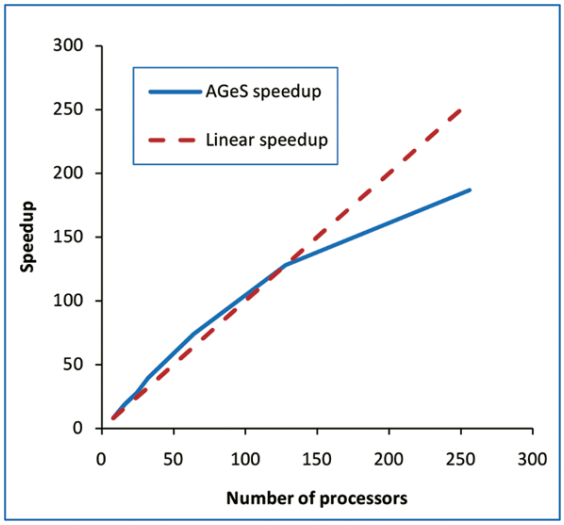

Figure 3. AGeS performance speedup as a function of the number of processors

3.2 Software Validation

We validated AGeS by comparing its annotations of bacterial genomes with annotations from three other sources. We evaluated two draft genomes, *Staphylococcus hominis* SK119 and *Staphylococcus aureus* subsp. *aureus* TCH60, and one completed genome, *Yersinia pestis* CO92. The *S. hominis* draft genome, sequenced by J. Craig Venter Institute (http://www.jcvi.org/cms/research/groups/microbial-environmental-genomics/), consists of 37 contigs, and the *S. aureus* draft genome, sequenced by the Human Genome Sequencing Center at Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/), consists of 65 contigs. Both of these draft genomes were sequenced using 454 pyrosequencing technology (http://www.454.com/). The complete *Y. pestis* genome was sequenced by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/resources/downloads/bacteria/yersinia.html) using Sanger sequencing technology.

The annotations for these three genomes were retrieved from the corresponding sequencing centers, and their sequences were re-annotated using the AGeS system. Figure 4A shows a subset of the compared genomic features[2]. The total number of annotated genes for each of these genomes was compared with the original annotations provided by the corresponding centers. Each of the two compared annotation sources predicted similar numbers of genes. For *S. hominis* (Sh), we found that 1,753 (~78%) genes were identical across both predictions. Most of the remaining genes overlapped at the start or end positions, with only 0.2% of the predictions unique to AGeS (data not shown). For the *S. aureus* (Sa) genome, 2,037 (~77%) genes were identical, with only 1% of the predictions unique to AGeS (data not shown). For the *Y. pestis* (Yp) genome, 2,637 (>60%) genes were identical across the 2 annotations, and another ~30% had identical start or end positions (data not shown). Annotation comparisons indicated larger differences for the *Y. pestis* completed genome than for the two draft genomes. These differences could be attributed to the more extensive studies performed in this well-studied genome. A similar level of agreement was observed for other genomic features, such as CDSs, rRNAs, and tRNAs.

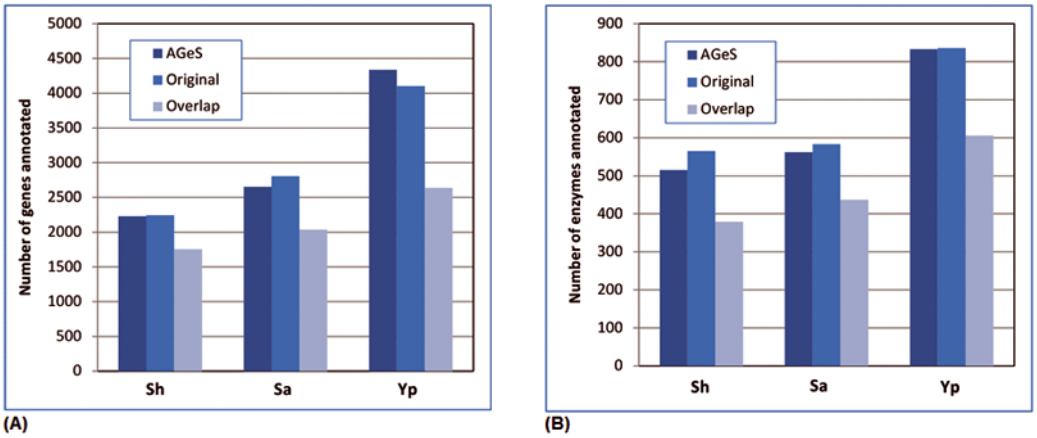

Figure 4. Comparison of gene annotations and enzyme function predictions between AGeS and the other three annotation systems for the three analyzed genomes, *Staphylococcus hominis* SK119 (Sh), *Staphylococcus aureus* subsp. *aureus* TCH60 (Sa), and *Yersinia pestis* CO92 (Yp). *A*: the number of genes predicted by the original annotation centers and AGeS, with the overlap corresponding to identical predictions. *B*: the number of enzymes predicted by the original annotation centers and AGeS, with the overlap corresponding to identical predictions.

We also compared the annotations of the enzyme functions predicted by the CatFam enzyme profile database with those provided by the other three annotation centers. Figure 4B shows the similar numbers of annotated enzymes for each of the three compared genomes[2]. For example, for the *S. hominis* (Sh) draft genome, CatFam assigned Enzyme Commission (EC) numbers for 515 genes, whereas the J. Craig Venter Institute assigned EC numbers to 565 genes, with 379 enzymes having identical EC number annotations. In general, our results indicate that the AGeS annotations are in agreement with the other evaluated methods both on the genomic and proteomic annotation levels.

4. Conclusion

The Web-based AGeS system described in this paper is a computationally-efficient and scalable system for high-throughput genome annotation of newly sequenced pathogens of military relevance and their near neighbors. The AGeS annotation pipeline is fully-parallelized and is currently operational at the Mana Linux cluster at the MHPCC, where we performed scalability tests and found that a 2.2-Mbp bacterial genome sequence can be annotated in ~1 hr using

128 processors. Validation results indicated that the AGeS system's annotations are in general agreement with the other evaluated methods, both on the genomic and proteomic annotation levels. Due to significant cost reductions afforded by the recently developed next-generation genome sequencing technologies, we expect that software applications such as AGeS will become vital for microbial comparative genomics studies.

Acknowledgements

This work was partially sponsored by the US DoD High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes Initiative.

Disclaimer

The opinions and assertions contained herein are the private views of the authors, and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense.

References

1. Hall, N., "Advanced sequencing technologies and their wider impact in microbiology", *The Journal of Experimental Biology*, 210(9), pp. 1518–1525, 2007.

2. Kumar, K., V. Desai, L. Cheng, M. Khitrov, D. Grover, R.V. Satya, C. Yu, N. Zavaljevski, and J. Reifman, "AGeS, a software system for microbial genome sequence annotation", *PLoS ONE,* 6(3), e17469, 2011.

3. Donlin, M.J., "Using the Generic Genome Browser (GBrowse)", *Current Protocols in Bioinformatics*, Chapter 9, pp. 9.9.1–25, 2009.

4. Stewart, A.C., B. Osborne, and T.D. Read, "DIYA, a bacterial annotation pipeline for any genomics lab", *Bioinformatics*, 25(7), pp. 962–963, 2009.

5. Benson, G., "Tandem repeats finder, a program to analyze DNA sequences", *Nucleic Acids Research*, 27(2), pp. 573–580, 1999.

6. Yu, C., N. Zavaljevski, V. Desai, S. Johnson, F.J. Stevens, and J. Reifman, "The development of PIPA, an integrated and automated pipeline for genome–wide protein function annotation", *BMC Bioinformatics*, 9, 52, 2008.

7. Benson, D.A., I. Karsch–Mizrachi, D.J. Lipman, J. Ostell, and E. W. Sayers, "GenBank", *Nucleic Acids Research*, 38(suppl 1), pp. D46–D51, 2010.

8. Van Domselaar, G.H., P. Stothard, S. Shrivastava, J.A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner, and D.S. Wishart, "BASys, a web server for automated bacterial genome annotation", *Nucleic Acids Research*, 33(suppl 2), pp. W455–W459, 2005.

9. Aziz, R.K., D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, F. Meyer, G.J. Olsen, R. Olson, A.L. Osterman, R.A. Overbeek, L.K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G.D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko, "The RAST Server, rapid annotations using subsystems technology", *BMC Genomics*, 9, 75, 2008.

10. Uchiyama, I., T. Higuchi, and M. Kawai, "MBGD update 2010, toward a comprehensive resource for exploring microbial genome diversity", *Nucleic Acids Research*, 38(suppl 1), pp. D361–D365, 2010.

11. Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, and C. Médigue, "MicroScope, a platform for microbial genome annotation and comparative genomics", *Database*, 2009, 2009.

12. Markowitz, V.M., I.–M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, I. Anderson, A. Lykidis, K. Mavromatis, N.N. Ivanova, and N.C. Kyrpides, "The integrated microbial genomes system, an expanding comparative analysis resource", *Nucleic Acids Research*, 38(suppl 1), pp. D382–D390, 2010.

13. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel–Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene ontology, tool for the unification of biology", *Nature Genetics*, 25(1), pp. 25–29, 2000.

14. Warren, A. S. and J. C. Setubal, "The Genome Reverse Compiler, an explorative annotation tool", *BMC Bioinformatics*, 10, 35, 2009.

15. Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M.J. Allen, S.V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. DePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glockner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz–Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens–Mack, S.E. Lewis, K. Li, A.L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S. A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil, G. Wilson, and A. Wipat, "The minimum information about a genome sequence (MIGS) specification", *Nature Biotechnology*, 26(5), pp. 541–547, 2008.

16. Paulson, L.D., “Building Rich Web Applications with Ajax”, *Computer*, 38(10), pp. 14–17, 2005.

17. Delcher, A.L., K.A. Bratke, E.C. Powers, and S.L. Salzberg, “Identifying bacterial genes and endosymbiont DNA with Glimmer”, *Bioinformatics*, 23(6), pp. 673–679, 2007.

18. Lagesen, K., P. Hallin, E.A. Rødland, H.–H. Stærfeldt, T. Rognes, and D.W. Ussery, “RNAmmer, consistent and rapid annotation of ribosomal RNA genes”, *Nucleic Acids Research*, 35(9), pp. 3100–3108, 2007.

19. Lowe, T.M. and S.R. Eddy, “tRNAscan–SE, a program for improved detection of transfer RNA genes in genomic sequence”, *Nucleic Acids Research*, 25(5), pp. 0955–0964, 1997.

20. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, “Basic local alignment search tool”, *Journal of Molecular Biology*, 215(3), pp. 403–410, 1990.

21. Yu, C., N. Zavaljevski, V. Desai, and J. Reifman, “Genome–wide enzyme annotation with precision control, catalytic families (CatFam) databases”, *Proteins*, 74(2), pp. 449–460, 2009.