GEPETTO (GEne PrioriTization Tool) is an open-source framework, distributed under the LGPL license. The source code is available at sourceforge.net/projects/gepetto/files/.

GEPETTO update: An Open Source Framework for Gene Prioritization

Hoan Nguyen
nguyen@igbmc.fr
Integrated structural Biology department, (IGBMC), Illkirch, France
Integrative Genomics and Bioinformatic Laboratory-LBGI, Strasbourg, France

Recently, the use of high-throughput biotechnologies has emphasized the need for new prioritization tools to identify the most promising genes/proteins among a list of candidates resulting from high-throughput experiments [1]. Large sets of genes must be evaluated, in order to score and rank them according to their similarity to known genes and their potential viability as candidates for important applications, such as diagnostic/prognostic markers, drug targets, etc. The biomedical community urgently needs a customizable and extensible framework for gene selection that can handle large-scale biological information from public, as well as private data resources.

GEPETTO (GEne PrioriTization Tool) is an original open-source framework, distributed under the LGPL license, for gene selection and prioritization on a desktop computer that ensures confidentiality of personal data. It takes advantage of the data integration capabilities from public database, combined with in-house developed gene prioritization methods. It currently incorporates six prioritization modules, based on gene sequence, protein-protein interactions, gene expression, disease-causing probabilities, protein evolution and genomic context). Each module integrates specialized evaluation or ranking approaches including K-means clustering, Pearson Correlation and Fisher's omnibus analysis and network-based approaches with neighbourhood evaluation of candidate or disease genes. The final overall prioritized candidate list is determined using several methods: order statistics [2], Robust Rank Aggregation[3] and GPSy's optimal weight [4].

GEPETTO is written in Java/Python and supported by an advanced modular architecture, which means that it can easily be modified and extended by the user, in order to include alternative scoring methods and new public/private data sources. Recently, we used the jBPM (JBoss Business Process Management) workflow engine to define and execute the prioritization process. The GEPETTO software and applications are available at sourceforge.net/projects/gepetto/files/ or decrypthon.igbmc.fr/sm2ph/cgi-bin/gepetto.

1) Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012 Jul 3;13(8):523-36.
2) Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012 Feb 15;28(4):573-80. doi: 10.1093/bioinformatics/btr709. Epub 2012 Jan 12.
3) Britto R, Sallou O, Collin O, Michaux G, Primig M, Chalmel F. GPSy: a cross-species gene prioritization system for conserved biological processes--application in male gamete development. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W458-65. doi: 10.1093/nar/gks380. Epub 2012 May 8.
4) Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y.A guide to web tools to prioritize candidate genes. Brief Bioinform. 2011 Jan;12(1):22-32. doi: 10.1093/bib/bbq007. Epub 2010 Mar 21.