BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 10 2005, pages 2520–2521
doi:10.1093/bioinformatics/bti380

Genome analysis

MIPS bacterial genomes functional annotation benchmark dataset

Igor V. Tetko*, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Gisela Fobo, Andreas Ruepp, Alexey V. Antonov, Dimitrij Surmeli and Hans-Wernen Mewes

Institute for Bioinformatics (MIPS), GSF National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany

Received on January 28, 2005; accepted on March 6, 2005
Advance Access publication March 15, 2005

ABSTRACT

Motivation: Any development of new methods for automatic functional annotation of proteins according to their sequences requires high-quality data (as benchmark) as well as tedious preparatory work to generate sequence parameters required as input data for the machine learning methods. Different program settings and incompatible protocols make a comparison of the analyzed methods difficult.

Results: The MIPS Bacterial Functional Annotation Benchmark dataset (MIPS-BFAB) is a new, high-quality resource comprising four bacterial genomes manually annotated according to the MIPS functional catalogue (FunCat). These resources include precalculated sequence parameters, such as sequence similarity scores, InterPro domain composition and other parameters that could be used to develop and benchmark methods for functional annotation of bacterial protein sequences. These data are provided in XML format and can be used by scientists who are not necessarily experts in genome annotation.

Availability: BFAB is available at http://mips.gsf.de/proj/bfab

Contact: i.tetko@gsf.de

Numerous genome-sequencing projects have caused a rapid growth of protein databases. In contrast to the pregenomic era, when the selection of sequences was highly biased toward known and characterized genes, the systematic exploration of genomes now allows to assign more and precise functional properties in the majority of cases. However, a manual annotation of sequences is laborious and expensive. Thus the development of methods for reliable functional annotation of bacterial genomes is of great importance for the bioinformatic community.

The annotation of protein sequences is a classification task. The problem is to build a classifier that can predict the function of new proteins based on the annotation of previously annotated sequences. One can always separate three stages: (1) preparation of the data for analysis using molecular indices, (2) development of a classifier and (3) validation (testing) of the developed model. All three steps are important when considering and comparing the performance of different annotation methods. The prediction performance of the annotation depends on a combination of the first two steps. However, a proper comparison of different models also depends on the validation protocol used. The use of different protocols or non-comparable statistical parameters makes a straightforward comparison of different procedures impossible.

A lot of functional annotation studies in the bioinformatics field were performed for one organism only using cross-validation approaches. However, the annotation within one species may not provide a proper test of the developed scheme. The performance of the annotation procedures in such analysis is usually limited to the classification of duplicated paralogous genes, which have similar or even identical function. At the same time, such methods could not be appropriate to annotate conserved genes that are not abundant in the organisms, even if they were conserved across different organisms. Thus, such analysis may provide a biased performance for the estimation of its prediction ability in cross-genome annotation settings.

At the same time, the evaluation of annotation across genomes may also meet some difficulties. Despite the existence of data for a number of genomes, there is a possibility that annotations performed by different teams of scientists could be inconsistent. For example, annotations of *Drosophila melanogaster* were performed independently by two groups, both using Gene Ontology (GO) (Mi *et al*., 2003). The result for the ontology 'biological process' was that only 1156 proteins were annotated consistently by both groups, but the GO assignments for 4137 proteins were different. If such annotations would be performed for similar but still different genomes, the observed difference could be interpreted as differences in the protein functions of both genomes. Thus, the inconsistent annotation may complicate development and testing of methods for automatic assignment of protein function.

Recently, we re-annotated four bacterial genomes, *Bacillus subtilis*, *Helicobacter pylori*, *Listeria innocua* and *Listeria monocytogenes* (in total, 11 502 sequences), previously annotated at different times according to the MIPS FunCat (Ruepp *et al*., 2004). The FunCat is a well established annotation scheme for the functional description of proteins from prokaryotes, unicellular eukaryotes, plants and animals, which has been intensively used in bioinformatics and machine learning studies (see Cai and Doig, 2004; Clare and King, 2003; Mateos *et al*., 2002). A detailed comparison of the new version of FunCat 2.0, GO and other schemes was recently published elsewhere (Ruepp *et al*., 2004). The joint efforts enabled us to achieve a consistent manual annotation of

*To whom correspondence should be addressed.

the genomes as it is exemplified in the sample benchmark study provided at our site that was performed using the sequence similarity scores only. Thus, the annotation of these genomes proved to be consistent and is therefore perfectly suitable for benchmarking any functional annotation approach. A use of more sophisticated methods of data analysis and/or different input data may provide even better results.

In order to facilitate the development of new annotation schemata for scientists from other fields of science, particularly the machine learning specialists, we precalculated most widely used and important parameters of the gene sequences. The parameters calculated include pairwise sequence similarity scores, namely BLAST, PSI-BLAST (Altschul *et al*., 1997) and FASTA (Pearson, 1996) as well as InterPro (Mulder *et al*., 2005) domain composition, similarities to COGs and SCOP domains, sequence localization (Nakai and Kanehisa, 1991) and amino acid composition were derived from the Pedant and SIMAP databases (Mewes *et al*., 2004).

All these data, as well as functional annotations of sequences are available for free download at http://mips.gsf.de/proj/bfab. The use of the same input datasets allows a direct comparison of different classification approaches. On the other hand, the use of different input parameters within a method allows to estimate the influence of molecule representation on the quality of the annotation and to propose new sets of indices that can be used in the annotation process. The input parameters are stored in XML. All XML files have the same structure. The text files can be derived from the XML file using a Perl script. This facilitates an easy conversion of files to different input files required by the users.

We invite all users to report their annotation performance for these data with leave-one-out genome schema, i.e. to predict the test target genome using the annotation information from the other genomes. Notice, that *L.innocua* and *L.monocytogenes* are very similar and they should not be used to predict one another. The schema provides a realistic scenario for the annotation of complete prokaryotic genomes.

We encourage authors to submit their preprints and upon publication, links to the published articles with analysis of MIPS-BFAB dataset to be included on our site for comparison purposes. The annotation data provide many different ways to estimate the performance of methods. For example, the performance can be estimated in terms of specificity, sensitivity, coverage, Receiver Operator Curve on different levels of annotation, e.g. the most general or the most fine level. Each measure can be more suitable for one or another purpose. That is why we invite users to submit their prediction results as a standard XML file with all results and also provide software tools to evaluate the performance of methods according to their favorite measure. This makes it possible for new users to compare all results in terms of their preferred performance measure(s).

ACKNOWLEDGEMENTS

This work was supported by grants 031U212C BFAM (BMFB) to H.W.M., BFAM and TE/308/1-1 (DFG) to I.V.T. and H.W.M. and TE 380/1-1 grant. We thank Dmitrij Frishman for his helpful suggestions.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Cai,Y.D. and Doig,A.J. (2004) Prediction of *Saccharomyces cerevisiae* protein functional class from functional domain composition. *Bioinformatics*, **20**, 1292–1300.

Clare,A. and King,R.D. (2003) Predicting gene function in *Saccharomyces cerevisiae*. *Bioinformatics*, **19**, II42–II49.

Mateos,A. *et al*. (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res.*, **12**, 1703–1715.

Mewes,H.W. *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.

Mi,H. *et al*. (2003) Assessment of genome-wide protein function classification for *Drosophila melanogaster*. *Genome Res.*, **13**, 2118–2128.

Mulder,N.J. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**, 95–110.

Pearson,W.R. (1996) Effective protein sequence comparison. *Meth. Enzymol.*, **266**, 227–258.

Ruepp,A. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.