GeneReporter-sequence-based document retrieval and annotation
GeneReporter—sequence-based document retrieval and annotation

Annekathrin Bartsch¹, Boyke Bunk¹, Isam Haddad¹, Johannes Klein¹, Richard Münch¹, Thorsten Johl², Uwe Kärst², Lothar Jänisch², Dieter Jahn¹ and Ida Retter¹.

¹Institute for Microbiology, Technische Universität Braunschweig, Spielmannstr. 7, 38106 Braunschweig and ²Cellular Proteomics Group, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Summary: GeneReporter is a web tool that reports functional information and relevant literature on a protein-coding sequence of interest. Its purpose is to support both manual genome annotation and document retrieval. PubMed references corresponding to a sequence are detected by the extraction of query words from UniProt entries of homologous sequences. Data on protein families, domains, potential cofactors, structure, function, cellular localization, metabolic contribution and corresponding DNA binding sites complement the information on a given gene product of interest.

Availability and implementation: GeneReporter is available at http://www.geneReporter-tu-bs.de. The web site integrates databases and analysis tools as SOAP-based web services from the EBI (European Bioinformatics Institute) and NCBI (National Center for Biotechnology Information).

Contact: i.reter@tu-bs.de; ida.reter@helmholtz-hzi.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on October 3, 2010; revised on December 1, 2010; accepted on January 9, 2011

1 INTRODUCTION

In face of next-generation sequencing and high-throughput analyses, the link between obtained data and existing knowledge is crucial. Automatic annotation pipelines provide useful evidence of potential functions for genes and proteins, but in a last essential step, the scientist must manually evaluate the available information. Usually, the necessary evidence is derived from scientific publications, databases and in silico predictions. Thus, tools that provide a combination of all of these relevant data for a gene or protein of interest are of high practical impact. In this context, GeneReporter offers a customizable workflow for the integrated application of protein sequence analysis and document retrieval.

A large number of diverse text-mining tools exist that provide different strategies and interfaces to satisfy the extensive data-mining demands in biomedical sciences (Krallinger et al., 2010). GeneReporter identifies citations related to a gene or protein sequence of interest. The UniProt annotations of homologous sequences are used to derive keywords such as gene names, synonyms and species. These keywords provide the query terms for a subsequent literature search in PubMed (Sayers et al., 2010). In this way, GeneReporter extends and replaces MineBlast (Dieterich et al., 2005), a similar tool which is discontinued. In comparison with other tools that connect literature to sequence information, like quickLit (Gilchrist et al., 2008) and Metis (Mitchell et al., 2005), GeneReporter is characterized by highly customizable query options, the integration of InterPro and the direct access to the original EBI and NCBI databases.

2 REQUESTING LITERATURE AND SEQUENCE ANALYSIS

The user can enter up to 10 nt or protein sequences to submit a query on the GeneReporter web site. Two different types of analyses are provided: (i) homology-based document retrieval searches information on homologous sequences from the UniProt Knowledgebase (UniProt Consortium, 2010) and citations from PubMed. (ii) Analysis of the protein sequences requests protein annotations from InterPro (Hunter et al., 2009), Phobius (Käll et al., 2007) and PrediSi (Hiller et al., 2004). The complete workflow is depicted in Figure 1. An example for an application is given as Supplementary Material.

Using homology-based document retrieval, the first step is a BLAST search in UniProtKB, where the user can select the desired algorithm, NCBI-BLAST (Altschul et al., 1997) and WU-BLAST (Lopez et al., 2003) result in a different ranking of homology matches, and therefore yield different query word extractions from the respective UniProtKB entries. PSI-BLAST (Altschul et al., 1997) is the most sensitive algorithm and beneficial for sequences that fail to result in significant hits with the other algorithms. Either Swiss-Prot or the complete UniProtKB can be chosen as BLAST
The results are summarized on an overview page. For each query, this page provides a link to a detailed view of the obtained data for the requested service. The result overview page can be bookmarked and results can be retrieved from this URL for at least 24 h. For further analysis, results can be downloaded as Excel or tab-delimited text files.

The detailed view provides one result tab for each requested service. The BLAST result tab shows homologous protein sequences. It is complemented with annotations from the UniProt database, e.g. organism name and GO terms, in order to facilitate their evaluation. The PubMed result tab shows gene-related citations ordered by the respective UniProt queries. Query words that were matched within title and abstract are marked in bold. For each query word combination, the link ‘This query in PubMed’ performs the corresponding query on the PubMed web site. This allows the manual modification and specification of the automatically generated queries with all the sophisticated features of the PubMed search interface. Furthermore, GeneReporter provides citations from UniProt entries of the BLAST hit sequences. In general, these references comprise the key papers on the respective gene or protein. Figure 2 shows the PubMed tab of an example search for a hypothetical protein from Pseudomonas aeruginosa C3719.

The output from InterPro, Phobius and PredSi, which search for putative transmembrane regions and signal peptides, is given in additional tabs. The InterProScan and Phobius output includes graphical visualizations of signature matches and transmembrane regions within the proteins of interest.

The detailed view provides one result tab for each requested service. The BLAST result tab shows homologous protein sequences. It is complemented with annotations from the UniProt database, e.g. organism name and GO terms, in order to facilitate their evaluation. The PubMed result tab shows gene-related citations ordered by the respective UniProt queries. Query words that were matched within title and abstract are marked in bold. For each query word combination, the link ‘This query in PubMed’ performs the corresponding query on the PubMed web site. This allows the manual modification and specification of the automatically generated queries with all the sophisticated features of the PubMed search interface. Furthermore, GeneReporter provides citations from UniProt entries of the BLAST hit sequences. In general, these references comprise the key papers on the respective gene or protein. Figure 2 shows the PubMed tab of an example search for a hypothetical protein from Pseudomonas aeruginosa C3719.

The output from InterPro, Phobius and PredSi, which search for putative transmembrane regions and signal peptides, is given in additional tabs. The InterProScan and Phobius output includes graphical visualizations of signature matches and transmembrane regions within the proteins of interest.

3 RESULTS

The results are summarized on an overview page. For each query, this page provides a link to a detailed view of the obtained data for the requested service. The result overview page can be bookmarked and results can be retrieved from this URL for at least 24 h. For further analysis, results can be downloaded as Excel or tab-delimited text files.

The detailed view provides one result tab for each requested service. The BLAST result tab shows homologous protein sequences. It is complemented with annotations from the UniProt database, e.g. organism name and GO terms, in order to facilitate their evaluation. The PubMed result tab shows gene-related citations ordered by the respective UniProt queries. Query words that were matched within title and abstract are marked in bold. For each query word combination, the link ‘This query in PubMed’ performs the corresponding query on the PubMed web site. This allows the manual modification and specification of the automatically generated queries with all the sophisticated features of the PubMed search interface. Furthermore, GeneReporter provides citations from UniProt entries of the BLAST hit sequences. In general, these references comprise the key papers on the respective gene or protein. Figure 2 shows the PubMed tab of an example search for a hypothetical protein from Pseudomonas aeruginosa C3719.

The output from InterPro, Phobius and PredSi requests is given in additional tabs. The InterProScan and Phobius output includes graphical visualizations of signature matches and transmembrane regions within the proteins of interest.