Supplementary Data

for the manuscript:

GeneReporter – Sequence based document retrieval and annotation

by Bartsch et al.

Table of Contents

1. Query and result options

1.1. Sequence input options

Type:

GeneReporter searches for amino acid sequence similarities. Consequently, the input is a protein coding DNA or an amino acid sequence. If the DNA sequence is not in frame, the BLAST option NCBI-BLAST;blastx is required.

Format:

Input sequences must have FASTA format.

Number of sequences:

You can post up to 10 sequences in one query. Queries will be conducted successively. Results will be available in the order of sequence input.

1.2. BLAST query options

Algorithm:

You can choose between NCBI-BLAST, WU-BLAST and PSI-BLAST. The algorithms differ in the ranking of similarity scores, sensitivity and program options. PSI-BLAST is the most sensitive algorithm and will detect also distantly related proteins. You can select one, two ore all algorithms in one query.

Program:

With NCBI-BLAST, 2 alternative programs are available: proteinBLAST and blastx. The proteinBLAST option is the default setting to compare a protein sequence with a protein database. Blastx performs a translation of DNA sequence into all 6 possible reading frames. This allows the input of genomic sequences that might have negative orientation or an out-of-frame start.

Database:

This option defines the BLAST database. The default database is SwissProt. Alternatively, you can select UniProt. This expands the search to the complete UniProt knowledgebase, which will probably lead to more BLAST hits but not necessarily to more information, as many UnProtKB/TrEMBL records are predicted and thus not well annotated.

E-value cutoff:

This option adapts the expectation value for the BLAST search. Default value is $1*e^{-1}$.

Bit score cutoff:

This options adapts the bit score cutoff for the BLAST search. Default value is 1 bit.

Max. passes:

This option adapts the number of passes for the PSI BLAST profile search. Default value is 3 passes.

1.3. PubMed query options

To start a PubMed search, a previous BLAST search for the extraction of query words is required. Thus, the selection of at least one BLAST algorithm is mandatory. The default algorithm is NCBI-BLAST.

Start/End year:

Define the period of publication dates you are interested in. Default values are 1990-present.

Number of publications:

Define the number of citations you want to retrieve from one PubMed query. The default value is 20, max. number is 50. To view more than 50 citations you can start your query directly on the PubMed website with one click from the result page (see: View result options).

Organism specific query:

The query option “Organism specific search” extracts the species name from the UniProt record of the homologous sequence and adds it to the subsequent PubMed query. Thus, this search returns citations corresponding to several species.
Alternatively, you can enter a species name in the field “Additional query terms as comma separated lists”. This search will return only documents matching the added species string.

Additional query terms as comma separated lists:

You can specify your PubMed queries by entering additional query terms here. These terms are linked with “AND” or with “NOT” to the automatically derived set of PubMed query words. Multiple comma separated terms are internally combined with “OR”. If you want to add author or journal names, the fields “Journals AND” and “Authors AND” restricts your input to these PubMed fields.

1.4. InterProScan query options

InterProScan integrates applications to search for protein families, domains, regions and sites. The applications use different methodologies and a varying degree of biological information on well characterised proteins to derive protein signatures.

The following applications are available: BlastProDom, FprintScan, Gene3D, HMMPanther, HMMPfam, HMMPIR, HMMSmart, HMMTigr, ProfileScan, PatternScan, SignalPHMM, SuperFamily, TMHMM.

1.5. Phobius query options

Phobius predicts transmembrane topology and signal peptides from the amino acid sequence of a protein. There are no additional query options for Phobius.

1.6. PrediSi query options

PrediSi predicts signal peptides of proteins transported through membranes. The specification of the organism in Gram-positive, Gram-negative Prokaryote or Eukaryote is mandatory.

1.7. Overview page options

The result overview page shows all information that belong to one query sequence in one line. Several buttons link to analysis results and download files.

View result: This button links to the web service analysis results for the respective sequence.
Excel: This button provides an Excel table that lists all analysis results for one sequence.
Search again: This button links to the query page including the respective sequence to restart the search with other parameter settings.
Zip Excel result: Zip the selected Excel result file(s) before download.
Zip CSV result: Provides the results in CSV format (i.e., one text file for each sequence and service) and zips selected result files before download.

1.8. Result page options

BLAST result:
The BLAST result table provides some columns with hyperlinks. The gene name links to the corresponding PubMed result, whereas other links provide access to the source databases (UniProt, GO and PubMed).

PubMed result:
View references from UniProt: Links to citations from the UniProt records of homologous sequences from the BLAST result.
This query in PubMed: Provides the opportunity to redo the respective search on the PubMed website. Query terms can be added and removed as desired.

Search again with this sequence: links to the query page including the respective sequence to restart the search with other parameter settings.

InterProScan result:

Links provide access to the referring InterPro records and to the analysis results on the web sites of the respective services.

2. Construction of PubMed queries

For the construction of the PubMed queries, the following information is extracted from the BLAST results and the corresponding UniProt entries:

- Gene name
- Synonym names
- Description (i.e., the long gene name):
  Descriptions that include the words “hypothetical”, “putative” and “uncharacterized” are excluded to avoid unspecific matches.
- Species name:
  The species name is only included for organism specific searches. The Species name is added both as systematical name (e.g., “Mus musculus”) and as trivial name “(e.g., “mouse”).

Redundant words are removed from the query.

Additional query words and dates can be added by the user by the definition of query options. To specify the PubMed search, query words are searched within certain fields of the PubMed database:

Type of query word	**PubMed field**
Gene name, synonym, species name	TIAB = Title and abstract
Description	TW = Text word (title, abstract, MeSH terms, …)
User defined query term	TW = Text word (title, abstract, MeSH terms, …)
User defined author	AU = author
User defined journal	TA = journal title abbreviation, full title, or ISSN number

Boolean combination of the query words:

- The default Boolean combination of query words is “OR”.
- For user defined query terms, the combination is “AND” or “NOT” as defined on the query page.
- User defined author and journal names are combined with “AND”.
- Descriptions and species names usually comprise several words. In this case, these words are bordered by brackets, and the words within the brackets are combined with “AND”.
  `Description: “Cobalamin biosynthesis protein BluB”`
  `=> (Cobalamin AND biosynthesis AND protein AND BluB)`
- The species name is combined with “AND”. However, alternative species names (i.e., trivial and systematical names) are combined with “OR”.
  `Species name: “Escherichia coli K12”`
  `=> (Escherichia coli OR E. coli OR K12)`

3. Cutoffs and limits

GeneReporter utilizes web services from the EBI, NCBI and our institute. This ensures that the analysis runs with the original source databases and up-to-date applications, which is of great advantage for the user. However, it includes some limitations in speed and the amount of results in one query.

The following properties are due to the SOAP based nature of GeneReporter:

1. A local queuing system ensures a maximum number of parallel queries.
2. Multiple sequences from a single query are sequentially processed. Equally, the results are sequentially available.
3. The number of input sequences is limited to 10.
4. The BLAST result is limited to the 20 best hits for each similarity search.
5. The PubMed result is limited to the 50 best hits for each query.
6. The Speed of a query depends on the speed of the involved external web services.
7. In case of failure of one web service, the workflow skips this service and returns the result of all other requested services.

4. Example of application

The basic idea of GeneReporter is to obtain as much information as possible for homologous proteins to get hints to the function of a predicted but yet uncharacterized protein through a single query. Obtained data must provide the basis for further experimental approaches. The following example was taken from an ongoing current project in our laboratory that aims to characterize the membrane proteome of *Pseudomonas aeruginosa*. Our specific interest here was the effect of aerobic *vs.* anaerobic conditions on the formation of membrane proteins and their regulatory interactions with quorum sensing systems of *P. aeruginosa.* One of the proteins identified as strongly regulated was PA3800. It is yet uncharacterized and its UniProt entry holds no references. The sequence was extracted from UniProt and entered into GeneReporter using the following parameters:

```
>tr|Q9HXJ7|Q9HXJ7_PSEAE Putative uncharacterized protein OS=Pseudomonas
aeruginosa GN=PA3800 PE=4 SV=1
MVQWKHAALLALALAVVGCSSNSKKELPPAELTDFKEEVVLSKQWSRSVGDGQGDLYNLL
EPAVDGSTIYAASAEGRVMAIQRETGDVLWKKDLERPVSGGVGVGYGLVLVGTLRGDVIA
LDEATGKKKWTKRVNSEVLSAPATNGDVVVVQTQDDKLIGLDAASGDQRWIYESTVPVLT
LRGTGAPLIAGNMALAGLASGKVVAVDVQRGLPIWEQRVAIPQGRSELDRVVDIDGGLLL
SGDTLYVVSYQGRAAALDVNSGRLLWQREASSYVGVAEGFGNIYVSQASGSVEGLDSRGA
SSLWNNDALARRQLSAPAVFSSNVVVGDLEGYVHLLSQVDGRFVGRERVDSDGVRVRPLV
VGSWMYVFGNGGKLVAYTIR
```

1. **Parameters for retrieval of literature and further data related to the protein:**
   - **a.** Standard NCBI Blast, with an e-value cutoff of 0.05, which corresponds to twice the standard deviation, and the default 1 bit score cutoff.
   - **b.** The PubMed search was limited to the ten years prior to the publication of the *Pseudomonas aeruginosa* PAO1 genome sequence [1] and to ten years after (i.e. 1990-2010). Organism specific search was unchecked to allow for the retrieval of documents on similar proteins from other organisms, We selected to see up to 50 publications.
   - **c.** InterProScan: We selected Gene3D for three dimensional structures, HMMPanther for predicted function, HMMSmart to predict domains, HMMTigr to double check homologous protein sequences, and finally TMHMM to predict transmembrane helices in the protein.
   - **d.** We also activated Phobius to double check transmembrane elements with TMHMM.

2. **Result: http://www.genereporter.tu-bs.de/example/1289837034.php**

3. BLAST result:

```
>SW:YFGL_ECOLI P77774 Lipoprotein yfgL OS=Escherichia coli (strain K12)
GN=yfgL
           PE=1 SV=1
          Length = 392

 Score =  213 bits (542), Expect = 2e-54
 Identities = 121/382 (31%), Positives = 201/382 (52%), Gaps = 13/382 (3%)

Query: 9   LLALALAVVGCSSNSKKE--LPPAELTDFKEEVVLSKQWSRSVGDGQGDLYNLLEPAVDG 66
           LL++ L + GCS  + +E  +  + L   + +   +  WS SVG G G+ Y+ L PA+
Sbjct: 11  LLSVTL-LSGCSLFNSEEDVVKMSPLPTVENQFTPTTAWSTSVGSGIGNFYSNLHPALAD 69

Query: 67  STIYAASAEGRVMAIQRETGDVLWKKDLERP----------VSGGVGVGYGLVLVGTLRG 116
           + +YAA   G V A+  + G  +W   L              +SGGV V  G V +G+ +
Sbjct: 70  NVVYAADRAGLVKALNADDGKEIWSVSLAEKDGWFSKEPALLSGGVTVSGGHVYIGSEKA 129

Query: 117 DVIALDEATGKKKWTKRVNSEVLSAPATNGDVVVVQTQDDKLIGLDAASGDQRWIYESTV 176
            V AL+ + G   W  +V  E LS P  +  +V++ T + +L  L+ A G  +W     +
Sbjct: 130 QVYALNTSDGTVAWQTKVAGEALSRPVVSDGLVLIHTSNGQLQALNEADGAVKWTVNLDM 189

Query: 177 PVLTLRGTGAPLIAGNMALAGLASGKVVAVDVQRGLPIWEQRVAIPQGRSELDRVVDIDG 236
           P L+LRG  AP  A   A+ G  +G+V AV +++G  IW+QR++   G +E+DR+ D+D
Sbjct: 190 PSLSLRGESAPTTAFGAAVVGGDNGRVSAVLMEQGQMIWQQRISQATGSTEIDRLSDVDT 249

Query: 237 GLLLSGDTLYVVSYQGRAAALDVNSGRLLWQREASSYVGVAEGFGNIYVSQASGSVEGLD 296
             ++    ++ ++Y G   ALD+ SG+++W+RE  S          IY+   +  V  L
Sbjct: 250 TPVVVNGVVFALAYNGNLTALDLRSGQIMWKRELGSVNDFIVDGNRIYLVDQNDRVMALT 309

Query: 297 SRGASSLWNNDALARRQLSAPAVFSSNVVVGDLEGYVHLLSQVDGRFVGRERVDSDGVRV 356
           G  +LW    L  R L++P +++ N+VVGD EGY+H ++  DGRFV +++VDS G +
Sbjct: 310 IDGGVTLWTQSDLLHRLLTSPVLYNGNLVVGDSEGYLHWINVEDGRFVAQQKVDSSGFQT 369

Query: 357 RPLVVGSWMYVFGNGGKLVAYT 378
            P+     + +    G + + T
Sbjct: 370 EPVAADGKLLIQAKDGTVYSIT 391
```

4. Evaluation of the BLAST result:

The *Escherichia coli* Lipoprotein YfgL revealed a similarity score of 213, E=2e-54. All other sequences scored no better than 65.5 and were excluded from further analyses.

5. Phobius result:

Phobius suggests a long non-cytoplasmic domain, and up to two transmembrane domains that are part of it.

6. InterPro result:

InterPro links to the 3D structure of YfgL, and HMMSmart predicts several transmembrane beta propeller repeats, which partly overlap with Phobius. Finally, HMMTigr comes to the same conclusion, that YfgL is the closest related protein.

7. PubMed result:

A survey of the nine retrieved references revealed that *yfgL* encodes a periplasmatic lipoprotein, which is part of a greater complex that is essential for outer membrane protein biogenesis and organisation [e.g., 2, 3].

8. Subsequent analyses:

As PA3800 is not annotated as a lipoprotein in UniProt, this was checked and consecutively affirmed by Lipo (http://www.bioinfo.no/tools/lipo), a specific tool for this purpose. Beta-propeller repeats occur in proteins that bind pyrrolo-quinoline quinone, a redox cofactor that in prokaryotes is associated with dehydrogenases, though probably not in this context of outer membrane assembly in Gram-negative bacteria.

Based on the obtained data from literature and from the several predictive tools retrieved by a single query we now assume that PA3800 in *P. aeruginosa* is a periplasmatic membrane-associated lipoprotein. In must participate in outer membrane (OM) assembly as part of a complex required for embedding proteins in the OM. This interpretation nicely fits with the observed regulation of *PA3800* gene expression when comparing aerobic *vs.* anaerobic cultivation conditions. This shift involves a major reorganisation of the OM. The combination of retrieved literature data and bioinformatic analyses results now provides a solid basis to devise experiments to test this hypothesis.

[1] Stover C.K., Pham X.-Q.T., Erwin A.L., Mizoguchi S.D., Warrener P., Hickey M.J., Brinkman F.S.L., Hufnagle W.O., Kowalik D.J., Lagrou M., Garber R.L., Goltry L., Tolentino E., Westbrock-Wadman S., Yuan Y., Brody L.L., Coulter S.N., Folger K.R., Olson M.V. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406,959-964(2000).

[2] Vuong P, Bennion D, Mantei J, Frost D, Misra R. Analysis of YfgL and YaeT interactions through bioinformatics, mutagenesis, and biochemistry. J. Bacteriol. **190**,1507-17 (2008).

[3] Charlson ES, Werner JN, Misra R. Differential effects of yfgL mutation on *Escherichia coli* outer membrane proteins and lipopolysaccharide. J Bacteriol. **188**, 7186-94 (2006).