BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 13 2013, pages 1693–1695
doi:10.1093/bioinformatics/btt265

Sequence analysis

Advance Access publication May 8, 2013

DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites

Brett Trost[1,*], Ryan Arsenault[2,3], Philip Griebel[2,4], Scott Napper[2,3] and Anthony Kusalik[1]

[1]Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, [2]Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon, SK S7N 5E3, [3]Department of Biochemistry and [4]School of Public Health, University of Saskatchewan, Saskatoon, SK S7N 5E5, Canada

Associate Editor: Janet Kelso

ABSTRACT

Summary: While many experimentally characterized phosphorylation sites exist for certain organisms, such as human, rat and mouse, few sites are known for other organisms, hampering related research efforts. We have developed a software pipeline called DAPPLE that automates the process of using known phosphorylation sites from other organisms to identify putative sites in an organism of interest.

Availability: DAPPLE is available as a web server at http://saphire.usask.ca.

Contact: brett.trost@usask.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 13, 2012; revised on May 2, 2013; accepted on May 3, 2013

1 INTRODUCTION

Protein phosphorylation is the most widespread cellular signaling mechanism in eukaryotes (Johnson and Hunter, 2005). Knowledge of an organism's phosphorylation sites facilitates the study of its cellular signaling pathways, which in turn has many applications in basic and translational research. Although online databases contain many phosphorylation sites for human, rat, and mouse, little data are available for other species. Using the cow as a test species, we previously proposed a protocol for making predictions in species with few known sites (Jalal *et al.*, 2009). Taking advantage of sequence homology between human and bovine proteins, this protocol involved manually using known human phosphorylation sites as BLAST queries to identify bovine sites. If a query and its best match in the bovine proteome had few or no sequence differences, the match was considered a putative bovine site.

While useful, several aspects of this protocol could be improved. First, its manual nature makes it time-consuming, and also limits the amount of known phosphorylation data that can be used. Second, it uses only known phosphorylation sites from human. It is possible, for instance, that a given bovine site might be homologous to a known rat site, but not to any known human site, and by using only known phosphorylation sites from human, this bovine site would be missed. This problem would be even more pronounced for species that are distantly related to human, such as plants. Third, the method used in Jalal *et al.* (2009) to identify non-orthologous proteins (comparing their annotations) has several drawbacks, including its subjective nature, the difficulty of automating these comparisons and the fact that annotations are often inaccurate or incomplete.

DAPPLE is a software pipeline that addresses these concerns, ultimately allowing the user to easily, quickly and accurately identify potential phosphorylation sites in an organism of interest.

2 DESCRIPTION OF DAPPLE

A complete description of the operation of DAPPLE, including a detailed flow chart, is available as Supplementary Material. Below, we briefly describe the input to, and output from, DAPPLE.

DAPPLE's input files are (i) the proteome of the target organism; (ii) a database of known phosphorylation sites; and (iii) the proteomes of the organisms represented in that database. All proteomes must be in FASTA format. Item (iii) is optional, but is necessary for DAPPLE to output information for the 'RBH?' column of the output table (see below). The phosphorylation site database can be obtained from a number of sources; a partial list is included in the DAPPLE documentation. This study uses phosphorylation sites from PhosphoSitePlus (Hornbeck *et al.*, 2012) (www.phosphosite.org/downloads/Phosphorylation_site_dataset.gz). The majority of sites in PhosphoSitePlus are represented by 15-mer peptides, with the phosphorylated residue in the middle. However, some sites are too close to the N- or C-terminus of the full protein to have seven residues on either side, and are thus represented by a shorter peptide. To allow them to attain statistically significant BLAST hits, for these sites DAPPLE uses as a query the first or last 15 residues of the full protein sequence. As such, all queries used in DAPPLE are 15 residues in length. Additionally, entries with identical sequences (from different organisms) are removed.

The remaining phosphorylation site sequences are used as queries to `blastp`, with the target organism's proteome as the database. Unlike in Jalal *et al.* (2009), queries are not limited to those from human. Information about the best match (as explained in the Supplementary Materials, weaker matches may optionally be used) is saved or computed, and ultimately presented in the DAPPLE output table (described below).

Because of the short length of the query sequences, the full protein corresponding to the best match may not be orthologous

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

to the full protein corresponding to the query. In Jalal *et al.* (2009), this problem was addressed by manually comparing the annotations of the two proteins. However, this approach suffers from the drawbacks described previously; thus, DAPPLE uses the well-established reciprocal BLAST hits (RBH) method to ascertain orthology (Overbeek *et al.*, 1999). For a known site X from organism A with match Y in target organism B, let X' be the full protein corresponding to X, and analogously for Y'. DAPPLE declares X' and Y' as orthologues if and only if Y' is the best match when X' is used as a query and the proteome of organism B is used as the database, *and* X' is the best match when Y' is used as a query sequence and the proteome of organism A is used as the database. In this case, 'the best match' is defined as any protein that has the smallest E-value. Soft masking of the query sequences is used when searching full protein sequences as suggested by Moreno-Hagelsieb and Latimer (2008).

DAPPLE outputs a table in which each row represents the result of a BLAST search using, as a query, one of the known sites in the phosphorylation site database. The table is in a tab-delimited plain text format that can easily be manipulated or imported into a spreadsheet program. This table contains many columns designed to help the user decide on the accuracy and usefulness of a given match; the following list describes most of these (for the full list, see the Supplementary Materials).

- Query accession, query description, query organism, query sequence, query site—the accession number, description, organism, amino acid sequence and phosphorylated residue (e.g. Y482) of X', respectively.
- Hit site, hit accession, hit description, hit sequence—the same as above, except for Y' rather than X'.
- Sequence differences—the number of differences between all of X (not just the portion that matched in the BLAST local alignment) and Y.
- Hit protein E-value—the E-value of the match between X' and Y' when X' is used as the query and B is used as the database.
- RBH?—'yes' or 'no', depending on whether X' and Y' are RBH.

3 RESULTS

To test DAPPLE, phosphorylation sites in the cow (*Bos taurus*) were identified, as was done by Jalal *et al.* (2009). The files described below were used as input to DAPPLE. The PhosphoSitePlus database was downloaded, and contained 214 185 unique phosphorylation sites. The proteomes corresponding to the target organism (cow) and the organisms represented in the PhosphoSitePlus database were downloaded from UniProtKB.

Table 1 compares the results given by Jalal *et al.* with those produced by DAPPLE. Note that both the methodology and input data used are not identical, so DAPPLE's output is not expected to be exactly the same. Nevertheless, the percentages of known phosphorylation sites that had a given number of sequence differences with their best bovine BLAST match were similar between the two approaches. For DAPPLE, the percentage of peptides under the 'no homology' category differed depending on the criterion for declaring two proteins as orthologues (see Table 1 caption), with the RBH method being less sensitive but more specific than the E-value method. Note that the sites reported by DAPPLE are only predictions; further, the functional significance of a homologous site may differ in the target organism, especially when the target is a distantly related species.

Table 1. Comparison of the results of Jalal *et al.* (2009) with those of DAPPLE

Sequence differences	% (Jalal *et al.*)	% (RBH)	% (E-value)
0	50	27.6	32.9
1	13	14.3	17.2
2	7	9.0	11.0
3	4	6.2	7.7
4	1.5	4.3	5.5
5	0.4	3.0	3.9
6	0.6	1.9	2.6
7+	0	1.4	2.0
No homology	22	32.2	17.1

Note: The first column indicates the number of sequence differences between a known site from PhosphoSitePlus and its best bovine match. The second column indicates the percentage of known sites with the indicated number of sequence differences in Jalal *et al.* (2009). The 'no homology' row indicates known sites for which there was either no match in the bovine proteome, or the annotation of the match differed from that of the query. The third column represents output from DAPPLE, with the 'no homology' row indicating that either the phosphorylation site had no match in the bovine proteome, or that 'RBH?' = 'no' (see Section 2). The fourth column is similar to the third, except instead of a site falling under the 'No homology' row if 'RBH?' = 'no', it does so if the hit protein E-value (see Section 2) is $>10^{-5}$. The E-value method represents a less stringent method of ascertaining homology (though not necessarily orthology).

Both the gain in efficiency using DAPPLE, and the value of using RBH as opposed to comparing annotations, are illustrated with examples in the Supplementary Materials.

4 CONCLUSION

DAPPLE improves on an already-successful method for predicting phosphorylation sites for non-typical model species. Our lab has used its output to help design peptide arrays containing targets of protein kinases (Houseman *et al.*, 2002) for studying honeybee, pig and chicken (our manuscripts in preparation), and it should be applicable to many other organisms, as well as other research problems related to protein phosphorylation. Finally, DAPPLE is not limited to phosphorylation; it could easily be applied to other post-translational modifications or to any problem that involves finding homologous motifs.

ACKNOWLEDGEMENT

The authors thank Stephen Johnson for helping test the software.

Funding: Natural Sciences and Engineering Research Council of Canada (NSERC).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Conflict of Interest: none declared.

REFERENCES

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Houseman,B.T. *et al.* (2002) Peptide chips for the quantitative evaluation of protein kinase activity. *Nat. Biotechnol.*, **20**, 270–274.

Jalal,S. *et al.* (2009) Genome to kinome: species-specific peptide arrays for kinome analysis. *Sci. Signal*, **2**, pl1.

Johnson,S.A. and Hunter,T. (2005) Kinomics: methods for deciphering the kinome. *Nat. Methods*, **2**, 17–25.

Moreno-Hagelsieb,G. and Latimer,K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, **24**, 319–324.

Overbeek,R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016